EDUCATIONAL EVALUATION, ASSESSMENT, AND MONITORING

A SYSTEMIC APPROACH

JAAP SCHEERENS, CEES GLAS AND
SALLY M. THOMAS

CONTEXTS OF LEARNING

SWETS & ZEITLINGER
PUBLISHERS

EDUCATIONAL EVALUATION, ASSESSMENT, AND MONITORING

A Systemic Approach

JAAP SCHEERENS
Department of Educational Organization and Management, University of Twente, Enschede, The Netherlands

CEES GLAS
Department of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

SALLY M. THOMAS
Graduate School of Education, University of Bristol, United Kingdom

SWETS & ZEITLINGER
PUBLISHERS

LISSE ABINGDON EXTON (PA) TOKYO

EDUCATIONAL EVALUATION, ASSESSMENT, AND MONITORING

CONTEXTS OF LEARNING
Classrooms, Schools and Society

Managing Editors:

Bert Creemers, *Faculty of Psychology, Education and Sociology, University of Groningen, The Netherlands.*

David Reynolds, *School of Education, University of Exeter, Exeter, UK.*

Janet Chrispeels, *Graduate School of Education, University of California, Santa Barbara, USA.*

EDUCATIONAL EVALUATION, ASSESSMENT, AND MONITORING

A Systemic Approach

JAAP SCHEERENS

Department of Educational Organization and Management, University of Twente, Enschede, The Netherlands

CEES GLAS

Department of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands

SALLY M.THOMAS

Graduate School of Education, University of Bristol, United Kingdom

SWETS & ZEITLINGER PUBLISHERS

LISSE ABINGDON EXTON (PA) TOKYO

Published by: Swets & Zeitlinger Publishers
http://www.szp.swets.nl/
This edition published in the Taylor & Francis e-Library, 2005.
" To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to http://www.ebookstore.tandf.co.uk/."

ISBN 0-203-97105-1 Master e-book ISBN

ISBN 90 265 1959 1 (HB) (Print Edition)
ISSN 1384-1181 (Print Edition)

Contents

Preface

In applications of educational evaluation “ongoing” monitoring approaches and various applications of assessing and measuring student achievement, appear to be more common and more frequent than program evaluations. Yet, the bulk of textbooks on educational evaluation address program evaluation. Without in any way denying the importance of the basic logic of program evaluation, including possible analytic pre-stages in conceptualizing the program to be evaluated, methodological contributions to attributing effects to “the program”, and conceptualization of the use of results, this book has a different orientation.

Based on an encompassing framework
It concentrates on the application of educational evaluation, assessment and monitoring activities that are embedded in organizational, managerial and instructional processes. The structure of the book is built around a three-dimensional model on the basis of which different types of educational “M&E”, as it is sometimes abbreviated, are distinguished. The three dimensions being:the basic function of the evaluation, the level of application in the education system and the data strategy.

We have called this approach “systemic” since it depends upon applied evaluation forms, embedded in a multi-level representation of an educational system.

Systemic approach
More specifically the term “systemic” is given the following interpretations:

- a *systems perspective* in the sense that M&E is used in the context of institutionalized application of M&E in education systems, not restricted to program evaluation;
- M&E is seen as functional to the day-to-day running and improvement of education systems; the theoretical principle that lies behind this view is *the cybernetic principle* from systems theory, which describes learning and control as contingent on evaluation and feedback;
- strategic use of M&E is seen as *dependent on the decision-making structure of multi-level education systems* and the dispersion of authority across levels;
- *comprehensiveness* in the sense that all forms of educational testing, monitoring and evaluation are seen as components that have a place to provide feedback with different orientations at different levels of education systems; strategic M&E is seen as an economic selection of components exploiting synergy between specific forms;
- an *input-throughput/process-output model* of education systems is used as a framework to indicate educational content and generate key object areas of education M&E.

Analyzing educational M&E “in context”

There are two ways in which monitoring and evaluation activities can be described as embedded in an organizational context. From a more theoretical stance the role of evaluation and monitoring can be analyzed, depending on specific ideal-type models of governance and management, such as synoptic rational planning, the functioning of market mechanisms in education, the "cybernetic principle" and related ideas such as "retro-active planning" and "schools as learning organizations". These conceptual frameworks are analyzed to get a better grip on the various functional roles evaluation and monitoring could play as part of governance and management. From a more pragmatic action oriented point of view the institutional, political and organizational context in which M&E takes place can be analyzed in terms of constraints. Understanding these contextual constraints is a first step in trying to overcome them and create better pre-conditions for evaluations to unfold in a technically appropriate way. Implementing M&E provisions at national level, particularly within the context of developing countries, is to be seen as an innovation process, that often requires organization development. This may involve recruiting and coordinating the necessary technical human resources potential, and facilitating the use of the information to decision-makers and other stakeholders.

Three core methodological orientations in educational M&E

In the introductory part (Part I) an overview of monitoring, assessment and evaluation approaches is given, the idea of a systemic approach is sketched and an overview of "basic concepts" is given. Part II is dedicated to organizational conditions and constraints as explained in the above. The book further concentrates on three major data strategies and approaches: assessment and measuring student achievement, monitoring school effectiveness on the basis of indicators and inspection and school self-evaluation, respectively in Parts III, IV and V.

In Part III on assessment and educational measurement a basic introduction to classical test theory and approaches depending on particular item response models is provided, including a varied set of applications in education.

Part IV is concentrated around the concept of educational effectiveness, and monitoring effectiveness on the basis of various kinds of indicator systems.

Part V finally, looks at different approaches to school self-evaluation, discusses various methodological and practical problem areas and provides an illustrative case-study of school self-evaluation projects in the UK. This part also includes a chapter on modern approaches to school-inspection.

Given these "ingredients" of an applied, systemic, monitoring emphasis, the attention for organizational context and constraints, the link with substantive knowledge on educational effectiveness, and state of the art methodology this book is believed to be relevant for several audiences. As a text-book for graduate students in education, as a guide-book for practicing evaluation researchers, and last but not least for education innovators and developers who work in the field of designing and organizing the "evaluation" function in developing countries. The authors have specific responsibility for certain parts of the book. Cees Glas has written Part III of the book and Sally Thomas has contributed the chapter on valueadded analysis in Part IV, and the chapters on school inspection (co-authored by Wen Jung Peng) and the Lancashire school self-evaluation

project (co-authored by Rebecca Smees) in Part V. Jaap Scheerens is responsible for the basic idea and design of this book and has written all the other chapters

The authors are particularly grateful to Carola Groeneweg who has edited the manuscript.

PART 1
Basic Concepts

1
Monitoring and Evaluation (M&E) in Education: Concepts, Functions and Context

1.1 Introduction

In this chapter basic concepts like evaluation, monitoring and assessment are defined. The chapter provides the outline of a framework to distinguish fifteen types of educational monitoring and evaluation. The framework depends on three basic dimensions: functions, data-strategy and (level of aggregation of) evaluation object.

All forms of *evaluation* consist of systematic information gathering and making some kind of judgment on the basis of this information. A further expectation is that this "valued information" is used for decisions on the day-to-day running of education systems or for more involving decisions on the revision and change of the system. The term "monitoring" is to be seen as a further qualification of evaluation, stressing the association with ongoing information gathering as a basis for management decisions, a reliance on administrative data and a stronger preoccupation with description than with "valuing".

In the description the term "education system", as the object of M&E, can be given different interpretations. It could be the national education system, a specific educational program, a school, or a classroom. The object of educational M&E can be defined at different levels of aggregation. Sometimes different terms are used when the object of evaluation differs. The term monitoring is often associated with the education system at macro level. Evaluation can be used for all objects but is most frequently associated with programs, as in program evaluation. When teachers are the object of evaluation the term "appraisal" is preferred in some national contexts (the UK in this case). And, finally, when the achievements of individual students are evaluated the term "assessment" is frequently used.

"Making empirically based checks on quality" can be seen as the overall purpose of educational M&E. Core functions are:

a. certification and accreditation; i.e. checking whether object characteristics conform to formally established norms and standards;
b. accountability; whereby object quality is made available for inspection to other units or the society at large;
c. (organizational) learning; whereby quality assessment is used a basis for improvement at the same object level.

M&E directed at these three core functions differs from a to c (in ascending order) in the degree of formality of criteria and standards, the external vs. internal nature of the procedures and a summative vs. a formative orientation.

Educational M&E makes use of different data sources. A pragmatic distinction is between data based on educational achievement measurement, data that is available from administrative records (including education statistics) and data that becomes available from expert review and educational research type of methods.

In subsequent sections specific types of educational M&E will be distinguished and elaborated by crossing *object, function* and *data source* as three basic dimensions.

1.2 Why do we Need Monitoring and Evaluation in Education?

The main motives for creating or improving provisions for monitoring and evaluation in education are three major concerns: to formally regulate desired levels of quality of educational outcomes and provisions; to hold educational service providers accountable and to support ongoing improvement in education. Decentralization policies in many countries are discussed as a stimulating contextual condition for systemic M&E.

• *to formally regulate desired levels of quality of educational outcomes and provisions*

Monitoring the quality of educational systems is not the first purpose of *examinations* that comes to mind. Examinations, for example at the end of lower secondary education, are there to certify individual students and to regulate what society can expect from those students (purposes of selection and stratification). Still examination can also be seen as a basis for determining the quality of educational systems and sub-systems, i.e. schools. Pass-rates on examinations are frequently used as performance indicators in judging the quality of educational programs and of schools.

When the unit of analysis to be formally evaluated is not the individual student but the school as an organization the term *accreditation* rather than *certification* is most commonly used. Quality control systems like the well-known ISO norms can be applied to schools to check whether central work and managerial processes are in place and the organization is customer oriented.

Finally, when explicit criteria and norms are used to compare educational achievement of national educational systems the term *benchmarking* is used. International assessment studies are needed to obtain basic and comparable data. In a global economy international benchmarking of educational quality is increasingly relevant for countries.

• *to hold education systems accountable for their functioning and performance and support direct democracy in education*

Accountability in education means that schools should provide information on their performance and functioning to outside parties. In this way schools and educational provisions are open to public review. Outside agencies, which have vested interest in the quality of education, may use this information for sanctioning (provide rewards or punishments). Such sanctions may be of an administrative nature, when originating from national, regional or local governing bodies, or take the shape of certain reactions from the consumers of education. Parents, for example, may try to persuade schools to alter

their practices, or, in situations of free school choice, may take their children to another school.

Several global developments have stimulated demands for accountability in education, these are:

- the growing realization of the increasing importance of education, when economies develop into "knowledge societies";
- the high costs of education, which in many countries are the highest post in government expenditure, paired with economic decline in the eighties this realization led to an increased concern with the *efficiency* of education provisions;
- an increased sense of openness and making public sector provisions in general accountable for the quality of their services (in the Netherlands for example the education inspectorate was forced by law to make public detailed reports on school reviews conducted by inspectors).

The substantive interest in accountability is usually in checking the *quality* or the general "well-functioning" of educational provisions. Quality is a rather general term. In actual practice concerns may relate to a good choice of educational objectives *(relevance)* or to the question whether the educational objectives are actually attained *(effectiveness)*. There may also be an emphasis on the fair and equal distributions of educational resources *(equity)* or a specific concern with an economic use of these resources *(efficiency)*. Recognition that schools are to be accountable to other stakeholders than just administrators or governmental units also points at a basic requirement for democracy. Particularly when this concerns the immediate consumers and the clients of educational provisions, information from M&E can be seen as a basis for more direct democracy in education. In its turn more influence from the immediate clients and stakeholders is also seen as a stimulant of effectiveness and efficiency.

- *as a mechanism to stimulate improvement in education*

Next to formal regulation of performance norms and stimulating accountability and democracy a third major function of M&E. in education is stimulating learning and self-improvement of educational units. When evaluative information is fed back to the units concerned, this can be an important basis for corrective action and improvement. The evaluation-feedback-action sequence is a central mechanism for all kinds of learning processes, including so called "organizational learning". The idea of learning from evaluation is central in the concept *of formative* evaluation, which is usually included in schemes for design and development in education.

- *decentralization policies in education in many countries as a stimulating condition (either by decentralizing M&E as well, or centralizing M&E as a counterbalance of more autonomy at lower levels in other domains)*

During the last two decades shifts in the patterns of centralization and decentralization have taken place in many countries, both in Western as in developing countries (OECD, 1998). Patterns of (de)centralization are best seen in terms of *functional decentralization*. This concept recognizes the fact that countries may decentralize educational systems in one domain, for example financial management, while simultaneously centralizing in other domains, like for example the curriculum.

This type of restructuring has stimulated the application of education M&E in two ways:

- more centralized control and stimulation of M&E as a counterbalance to providing more leeway and freedom with respect to school management and pedagogy (this pattern is most clearly visible in the UK);
- stimulation of school-based evaluation as part of decentralization "quality care" to the school level; to some extend this trend is discernable in Italy; in other cases despite decentralizing quality care to schools, M&E strategies are still mixed in the sense that more centralized forms are strengthened simultaneously (the Netherlands is a case in point).

What all three functions of M&E that were discussed in this section have in common is the purpose to stimulate quality. The first one, accreditation/certification depends on formally and officially established criteria and norms. The second one, accountability, may benefit from these formal criteria and norms, but is essentially relational in that lower level units in the system account for their performance to either official or unofficial (clients) stakeholders. The third (organizational learning) has a focus on within-unit improvement. Although accountability is ultimately related to improvement as well, the feedback-loop is shorter when M&E is applied internally.

In subsequent chapters specific applications of M&E will be discussed, which are differentially oriented towards one of these three basic functions (accreditation, accountability and self-improvement). The differences and correspondences between M&E types, primarily serving a particular function, will become clear when doing so. An important perspective, which will be given specific attention, is the option to exploit synergy between different basic forms and make efficient combinations.

1.3 A Conceptual Framework to Distinguish Technical Options in Educational M&E

Functions, data sources and objects are used as the basic dimensions to categorize M&E types in education. In this way 15 different types are distinguished.

Considering terminology *assessment, appraisal, evaluation* and *monitoring* are almost synonyms when one looks them up in the dictionary. They all have elements of valuing and judgments, being authorized to do so and of attributing numerical estimates. Monitoring stands out for its connotation of "detection" and association with controlling the running of a system over time and "keeping order". (One definition the Concise English Dictionary gives of "monitor" is "a lizard supposed to give warning of approach of crocodiles"). In the usage of these terms in education, the most frequently chosen *objects* that are judged, appraised, evaluated and monitored seem to be most decisive in the choice of terms:

Assessment, when students are the object;

Appraisal, when teachers are the object;

Evaluation, when an educational program is the object;

Monitoring, when the day-to-day running of educational systems and organizations is at stake.

It should be noted, however, that the use of these terms differs between countries. The above definitions more or less confirm to the way they are used in the United Kingdom. In the USA, the term “testing”, or “educational testing” is more commonly used for the assessment of “traditional” subject matter mastery, whereas “assessment” has the connotation of “alternative assessment”, in the sense of measuring more general skills and attitudes.

The conceptual framework to categorize types of educational evaluation, assessment and monitoring consists of *three basic data sources, three core functions and five different evaluation objects, each of which is defined at a particular level of aggregation.*

Table 1.1 Overview of M&E types; MIS means Management Information System.

Data Source	**Test and assessment data**			**Adminis trative data; statistics**			**Systematic inquiry and reivew**		
Function Object	**Accoun tability**	**Impro vement**	**Accr ditation**	**Accou ntability**	**Impro vement**	**Accredi tation**	**Accoun tability**	**Impro vement**	**Accre di tation**
System	National /Inter national Assessment			MIS	MIS		Inter national Review panels	Intern ational Review panels	
Program	Formative and summative evaluation of outcomes and processes using various data sources								
School	School Performance Report	Test -based school self-eval uation	School accreditati on/audits	School MIS	School MIS		Insp ection	Insp ection School Self Eval uation	Quality audits
Teacher	Assessment of comp etencies		Teacher certification	School MIS	School MIS		Insp ection	Inspection	
Student		Student monitoring system	Exams		School MIS			Mon itoring of behaviour by teachers	

The three basic data sources are:

- student achievement and assessment data
- administrative data and descriptive statistics
- data from expert reviews and systematic inquiry (surveys, observations and ratings)

The three functional areas are as described in an earlier section of this chapter:

- accreditation and certification

• accountability
• diagnosis/organizational learning

The five evaluation objects are:

• the education system at national level
• an educational program
• the school
• the teacher
• the individual student

By crossing these three dimensions (see Table 1.1) the main forms of educational M&E can be characterized.

The following test and assessment based types are distinguished:

1. national assessment programs
2. international assessment programs
3. school performance reporting
4. student monitoring systems
5. assessment-based school self-evaluation
6. examinations

Next, there are two basic kinds of monitoring systems that depend on statistics and administrative data:

7. system level Management Information Systems
8. school Management Information Systems

The following forms depend on data from expert review and systematic inquiry:

9. international review panels
10. school inspection/supervision
11. school self-evaluation, including teacher appraisal
12. school audits
13. monitoring and evaluation as part of teaching

Finally, there are two forms that will be discussed globally and not differentiated according to data source:

14. program evaluation
15. various forms of teacher evaluation

In a subsequent chapter (Chapter 3) these M&E types are described in more detail, in this section a further clarification on the three basic dimensions is given.

Basic data sources

Educational measurement, or rather the measurement of educational achievement, is one of the two basic forms of educational evaluation, the other one being program evaluation. At this stage it is sufficient to say that the technology and formal conceptual background of educational measurement are highly developed. Important issues are:

- the degree to which tests are curriculum-tied or aimed at general skills and "cross-curricular competencies";
- item formats, closed vs open;
- the idea of authentic assessment (measuring skills in real-life settings or simulations thereof);
- norm referenced vs criterion referenced testing, an issue that is related to standard setting; here the central issues are whether tests are well fit to discriminate and select (norm referenced testing), or should give a clear indication about which educational content is mastered when a particular score is obtained (criterion and standard based testing);
- the psychometric model to which a test confirms; here an important development is item response theory (IRT), which allows for tests results to be better interpretable (for further details see Part 3 of this book).

Rather than addressing technical issues in assessment the present focus is more on presenting the different options of applying it for different functions and as a component in broader M&E strategies, like the inclusion of student assessments in system level indicators or management information systems.

Educational statistics provide numerical data on the inputs (costs and resources, human resources), flows (participation rates, position of students with a particular level of education on the labor market) and outcomes (graduation rates, proportion of students that enroll in a higher education level) of education. Of course data from educational testing and assessment can also be expressed in summary statistics. Sometimes basic data on the larger societal and economic context in which education system operate are included as well. The term indicator is often used with the same meaning as an "education statistic" It is used to express the view that a particular statistic stands for a key aspect of education. Also, use of the term indicator is the more likely when reference is made to a composite of several basic statistics or variables (like for example the pupil/teacher ratio). Finally, when a statistic has an explicit evaluative rather than a merely descriptive interpretation, the term *indicator* is likely to be used as well.

When sets of indicators are selected on the basis of an implicit or explicit model of the functioning of an educational system one usually refers to them as *indicator systems*. The degree of connectivity or integration between indicators may vary. In many applications each indicator stands more or less on its own. Only when different types of indicators are linked by information that is collected on the same or explicitly related units can interrelationships between indicators be examined.

When sets of education statistics or indicators are collected at system level they are sometimes referred to as forming a Management Information System (MIS). Similarly systems that are based on administrative data at school-level are referred to as School Management Information Systems. A school MIS may contain, for example, data on the number of students with a particular socioeconomic background, absenteeism and school careers of students.

Data based on expert review and systematic inquiry are seen as the third basic data source for educational M&E. Methods obtain their legitimacy either by the professional status and expertise of the reviewer or through meeting scientific criteria concerning objectivity, reliability and validity of procedures and datacollection instruments.

Inspection, peer-review, audits, methods of school-based review and education research methods, are all united under this broad category.

Functional areas

Summarizing the earlier description the following three functional areas are distinguished:

Accreditation and certification are meant to ascertain that organizations or individuals have reached legally and formally established norms. The term accreditation is used when this happens for organizations, for example when the ISO norms are applied. Certification is used when students get a diploma, obtained on the basis of an examination.

Accountability refers to holding public institutions and services responsible for the quality of their performance. It has the following ingredients: disclosure of the product or service being provided; product or performance testing; and redress for poor performance, in other words: sanctions.

Diagnosis, improvement and organizational learning. When this function is aimed for M&E is meant to provide information to facilitate learning and modification as part of a developmental process or as part of the day-to day running of a system. M&E plays a more formative role. For example when a diagnostic achievement test is used the primary aim is not to decide whether or not the performance of a student is good enough to grant him or her a diploma. In stead the aim is to find out where he or she has particular weaknesses, so that these can be addressed in remediation. The same applies to school diagnosis where the strengths and weaknesses in the functioning of a school organization are used to find out in which area improvement should take place.

Evaluation objects

The five types of objects of educational M&E that are distinguished refer to levels of aggregation in educational systems, with the system defined at the national system containing the others: programs, schools, teachers and students. The limitation to five is somewhat arbitrary, since other administrative levels, like the state in federal systems, the province and the municipality can also be distinguished. The five object levels that are discerned here are considered sufficient to clarify basic M&E types, however. Of these evaluation objects, only “programs”, and program evaluation need some further clarification at this stage.

Program evaluation is aimed at determining whether or not a program or project has been successful in attaining its goals. Here the term program can either be taken as a specific, well-defined part of the normal day-to-day functioning of an education system, or as a new, innovative program.

Program-evaluation cannot be easily classified according to functional area. It could be seen as serving a more distinct and specific function, namely to determine the success of programs. This function is not expressed well by either accountability or organizational learning. It has elements of both functions. When a program evaluation is designed to have both *formative* and *summative* elements, the former is close to the improvement perspective and the latter close to the accountability perspective. Formative

evaluation is defined as evaluation that takes place during pilot-stages or the implementation phase of a program. It is aimed at providing feedback that is relevant to support and improve the process of implementation. Summative evaluation makes up the balance in checking whether a program has reached its objectives.

Program evaluation cannot be classified so well with respect to basic data sources either. In most cases various data sources (particularly data based on research-like systematic inquiry) are used, according to a specific design that allows for attributing program effects to program characteristics.

In Chapter 3 each of the 15 M&E types that were listed in this section will be further described with respect to: general characterization, major audience and types of use of the information, technical issues, technical and organizational capacity required and controversial issues.

1.4 Pre-Conditions in Educational M&E

The General Service Director General of the Ministry of Education in a NorthAfrican country sets out to improve "school evaluation", as part of a more general aim to improve educational evaluation and assessment in the country. The idea is that not only will schools be externally evaluated, but also internally, in the sense of *school self-evaluation.* The source of this idea is an education improvement project for the elementary and secondary education sectors, supported by an international organization. Within the governance structure the idea of improving educational evaluation is supported by two members of the Cabinet of the Minister. When consultants are called in to develop concrete proposals for school evaluation it becomes clear that several units of the Ministry and several semi-independent institutes have some kind of involvement. One of the most striking experiences of the consultants is that the Director of the Project Implementation Unit of the educational improvement project is unavailable to discuss the context of their study. The consultants report to the Director General of General Services, but the Director General of the Planning department appears to be most directly involved. Other units that have some kind of involvement are an institute for educational research and evaluation, an informatics unit in the Ministry and the Inspectorate. Among the various parties that are involved the two members of the Minister's Cabinet appear to be the most engaged; others display a more passive attitude, although all agree to the general sensibility of improving school evaluation. The impression that the overall purpose to do so has not been thoroughly thought through is enforced when the purpose of doing school self-evaluation is considered. The education system is centralized to a degree that there is hardly any autonomy for school directors so that it is questionable whether there is a real context for internal school self-evaluation as a tool for *school improvement.*

Educational monitoring and evaluation is definitely part of the rhetoric of systematic educational innovation. It is part and parcel of any type of rational planning scheme used in the preparation of reform and improvement programs. The logic is clear and it makes perfect sense. Yet, in actual practice it is very often the last item on the policy agenda, and doing something about it more like a necessary symbolic ritual than something "for real". Even though the logic is unbeatable and there are usually some stakeholders genuinely interested, it still is a "hard sell" when it comes to developing and

implementing sustainable monitoring and evaluation provisions. How could one possibly hope to turn this overriding attitude around? Here are a few possible answers to this:

- use the momentum of M&E being part of current reform models and planning schemes;
- consider the design and implementation of M&E as an innovation program in its own right, justified by the global call for *quality* in education;
- stress the innovative and "learning" potential of institutionalized M&E as a lever for educational improvement; in other words show that educational M&E can be useful.

What the example illustrates is that, when it comes down to taking concrete steps in establishing or improving educational M&E, one cannot take "the political will" to do so for granted. There is more to this issue than M&E possibly having low priority among other items on the educational reform agenda. Once it gets off the ground M&E provides information, maybe even strategic information, and information means power and shifting the balance of power when it becomes available more readily to some stakeholder as compared to others. Moreover, M&E leads to "valuation" and judgement which are likely to evoke resistance among those being judged, particularly when this takes place in a context where there is already some antagonism among the parties concerned.

The success of improving M&E as a support function of the day to day running and strategic planning of educational systems, as a matter of course, depends on the degree to which the country already has a history in the employment of this function. Ideally there should be some societal patterns in which such a function has a place. This is a matter of structural and formal arrangements, like an examination system, but also of something that could be indicated as an "evaluation culture". In the case of the North-African country there were several relevant elements. The country had at the time of study four different school inspectorates. A national program of assessing student achievement was being prepared. Special policy to support schools in so called "priority zones" (characterized by a high proportion of disadvantaged students) was accompanied by some systematic monitoring activities.

Apart from political aspects and institutional pre-conditions improving M&E also depends to a large extend on *organizational pre-conditions*. As in the case of the example of the North-African country there are usually several organizational units active in this field. In that country it was not so clear where to locate the basis for the furthering of school-evaluation: at the Planning Unit of the Ministry, with the Research and Evaluation Institute, in the Informatics Unit or with one or more than one of the School Inspectorates. When it was proposed that several of these units join forces in a more comprehensive approach to school evaluation the issue of coordination between these independent organizational units arose.

What the example indicates is that even before anything is said about the growing range of technical possibilities in educational M&E political, institutional and organizational aspects of the local context need to be taken into consideration. Successful use and implementation of these technical options depend on creating supportive conditions in these areas.

1.5 Conclusion: Why Speak of "Systemic Educational Evaluation"?

In a News Release on achievement differences between states in The USA, based on analysis of National Assessment of Education Progress (NAEP) tests, the Rand Corporation states the following conclusion about the remarkable achievement gains in two states, North Carolina and Texas:

> *"The most plausible explanation for the remarkable rate of math gains by North Carolina and Texas is the integrated set of policies involving standards, assessment and accountability that both states implemented in the late 1980s and early 1990s."* (Rand News Release, July 25, 2000).

Among the not always unambiguous findings of empirical school effectiveness research *frequent monitoring and evaluation of students' progress* stands out as a factor that is consistently mentioned in research reviews as a correlate of educational achievement. It appears to have a significant effect in meta-analysis and has a clear theoretical interpretation (Scheerens & Bosker, 1997).

Clearly, educational monitoring, evaluation and assessment should not just be seen as discrete events of appraisal and reflection (important as these are in their own right) but also as key-mechanisms that drive regulation and improved functioning of education systems. Of course, this is far from being a new insight. Movements in the evaluation field marked by terms like *"decision-oriented evaluation"*, *"utilization focused evaluation"* and *"stake-holder based evaluation"* have tried to make the same point. In the current context this functional view of M&E is seen as particularly compelling. As the range of technical options is expanding "improving the evaluation function of education systems" is seen as an educational reform in its own right. At the same time, it is increasingly being realized that political, organizational and technical pre-conditions need to be systematically considered when implementing M&E. Seeing educational monitoring and evaluation as a more permanent function of information provision, appraisal and feed-back to relevant units also marks a departure from "stand alone" program evaluations, as the prototype form of educational evaluation. The term "systemic M&E" is coined to underline this view.

The term "systemic" refers to the educational system as a whole, not confined to a particular part (cf. Concise Oxford Dictionary). More particularly the term *"systemic M&E"* is used to express the following points:

- *a systems perspective* in the sense that M&E is used in the context of institutionalized application of M&E in education systems, and is not restricted to program evaluation:
- M&E is seen as functional to the day-to-day running and improvement of education systems; the theoretical principle that lies behind this view is *the cybernetic principle* from systems theory, which describes learning and control as contingent on evaluation and feedback;
- strategic use of M&E is seen as *dependent on the decision-making structure of multi-level education systems* and the dispersion of authority across levels;
- *comprehensiveness* in the sense that all forms of educational testing, monitoring and evaluation are seen as components that have a place to provide feedback with different

orientations at different levels of education systems; strategic M&E is seen as an economic selection of components, while exploiting synergy between specific forms;
- an *input-throughput/process-output model* of education systems is used as a framework to indicate educational content and generate key object areas of education M&E.

Throughout this book this view of educational evaluation will be worked out in more detail, with respect to technical, organizational and substantive aspects.

2
Basics of Educational Evaluation

2.1 Introduction

In this chapter some essential aspects of general evaluation methodology will be discussed. The chapter provides an at a glance view on aspects that will be treated in more detail in subsequent chapters. In this way the core “logic” of systematic evaluation is introduced. In addition, important distinctions in evaluation theory will be discussed concerning ideal-type stages in evaluation projects, methodological implications of accountability and improvement perspectives, and formative and summative roles of evaluation. The section on “rationality assumptions” links this chapter to the next one on relevant aspects of the organizational and political context of evaluations.

2.2 Basics of Evaluation Methodology

2.2.1 Evaluation objects, criteria and standards

The sentence: “who is evaluating what and for what purposes” provides a general ordering framework to further qualify evaluations as defined in the first chapter. In this section the *“what”* of the framing sentence will be dealt with. The evaluation object or the “evaluandum” is the entity that is, ultimately, to be judged on the basis of systematic information gathering. Of course the delineation of the evaluation object defines the borders of the system (i.e. the set of elements and relationships between elements) that is to be subjected to evaluation. Examples of evaluation objects within the educational domain are: the national educational system as a whole or a specific sector within that system, for example vocational schools, the collection of schools that take part in a particular program that is to be evaluated, individual schools, teachers or pupils.

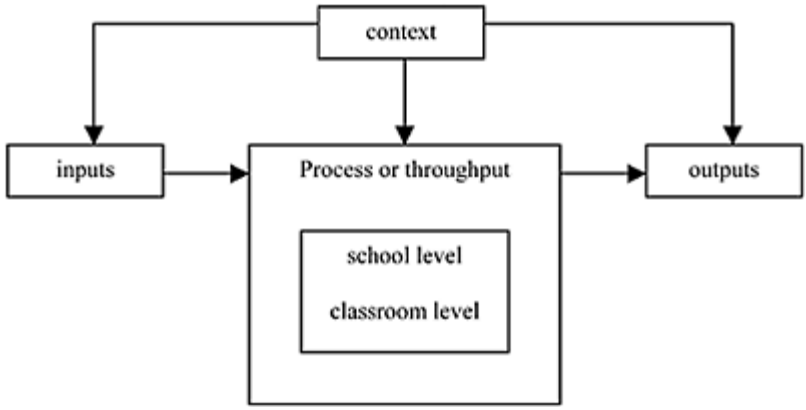


Figure 2.1 A basic systems model.

It is usually helpful to define evaluation objects according to an input, process, output and context framework (see Fig. 2.1). This basic framework will be used throughout this book as a model to categorize different kind of educational content. The model can be defined at different levels of aggregation. The central box can be viewed as the national education system, a particular program an individual school, classroom and even student. It is important to note that the model is dynamic in the sense that it views education as a production function: educational inputs are transformed to educational outputs. Typical inputs are material and financial resources, process characteristics are organizational and instructional structures and process, outputs of schooling are, for example, scores on achievement tests and relevant context aspects are, for example, attainment targets set by a higher administrative level.

Each of these abstract entities, inputs, processes, outputs and contextual conditions may be evaluated on its own. Accordingly a distinction can be made between input, process, output and context evaluation; of these four, process and output evaluation are most commonly encountered. In the case of input evaluation the actual financial resources of a national system, program or school may be described and judged according to the level that is thought to be necessarily in order to keep the system running. Process indicators may be assessed by comparing them to generally accepted ideas on educational good practice, while outputs can be judged according to pre-fixed attainment levels, margins of variability that are deemed acceptable and by comparing them to other relevant situations.

Indicators of the context of, for example, a school may be judged with respect to their being considered as favorable or unfavorable to a proper functioning of the school.

The terms evaluation criterion and evaluation standard are often confused. The criterion is the dimension on which the evaluative interpretations are ultimately made. For example, a math test can be used as the criterion in an educational evaluation. The standard refers to two things: the criterion (in the sense just defined in the above) *and* a norm on the basis of which it can be decided whether a “success” or a “failure” has been achieved. Cutting scores defined on a particular achievement test provide an example of a standard. In this case, the example of a cutting score, the standard is absolute. An

alternative is the use of comparative standard; for example the statistical significance of difference in mean scores of a treated and a control group.

2.2.2 Measurement of criteria and antecedent conditions

Measuring outcomes

In many educational evaluations attainment, in the sense of scores on achievement tests in particular subject-matter areas, is the central criterion. In such situations achievement tests are used as the operationalization of educational goals or objectives. When such objectives are tied to subject-matter areas the process of operationalization and test construction is quite straightforward. The basic steps of this process are as follows:

- precise statement of the general educational goal, e.g. pupils should master all aspects of numeracy at the level of the final grade of primary school;
- delineation of components, e.g. fractions, mental arithmetic, decimal system, etc.;
- further specification of subject-matter elements and required skills for each component, e.g. multiplication, addition and division of fractions;
- ordering of subject-matter elements and types of problems/questions/items according to difficulty-level;
- formulation of specific questions, items, problems;
- scaling of items and questions, i.e. determining on the basis of try-outs and by means of specific psychometric analyses whether sets of items are homogeneous and can be placed on a one-dimensional continuum or scale;
- further analyses concerning the reliability and validity of the tests (in other words, the provisionally scaled sets of items).

In so-called classical test theory the concepts of reliability and validity are based on the idea of a “true score” that tests are trying to measure. Since tests are considered to be less than perfect it is recognized that an actual test will contain some error, defined as the discrepancy between the theoretically assumed “true” score and the actual test score. When the sources of error are considered as random fluctuations that depend on all kinds of “disturbing” conditions, such as, a child having a cold during a particular testing session, disturbances because of construction activities in the school during the administering of a test, the imperfection stemming from these sources is seen as lack of *reliability*. When the sources of error are systematic, however, this is considered as a problem of *validity*.

In order to overcome problems of reliability test items should be sufficiently precise and clear in order to make them less vulnerable to influences from random disturbances. Reliability is usually checked, in a try-out situation during test construction, by administering the test twice to the same group, by having two homogeneous groups in ability level do the same test, and by splitting tests in two equal halves, and them computing the correspondence (correlation).

Checking the validity means making sure that the test is measuring what it is supposed to measure. Reliability can be seen as a pre-condition for validity, i.e. in order to be valid a test should also be reliable. On the other hand reliability is no guarantee that the test will also be valid. The *content validity* of an achievement test is checked by analyzing

whether the set of subject-matter components (i.e. the test items) adequately represent all subject-matter elements that together constitute the subject-matter domain in question. The *construct validity* of the test is checked by making sure that a test is measuring the construct it is supposed to measure, irrespective of the test format. One way to do this is to compare various testing formats of an achievement test with similar formats that measure other traits, like, for example, general intelligence. In this way "multi-trait", multi-methods (formats) are constructed and analyzed. The prediction is that correlations between different formats of measuring trait A will be higher than correlations between trait A and B when the same format or method is used.

The *predictive validity* is concerned with the correspondence of test outcomes with other criteria that are measured at a future point in time. For example, achievement test scores at the end of primary school are expected to correlate with performance of the same student in secondary school.

In more recent developments on educational measurement and psychometric theory, so-called "item response" theory, stronger assumptions according to the invariance of test outcomes across difficulty levels and stub-populations taking the test are used. When sets of items conform to the assumptions of particular item response models this makes the interpretation of the scores and the comparison of scores across difficulty levels easier.

It should be noted that educational measurement and underlying psychometric theory forms a discipline in itself that has reached high levels of formal and mathematical sophistication. In this chapter only a sketchy introduction is aimed for; Part 3 of this book on student achievement measurement provides a more thorough introduction.

Before leaving the issue of measuring educational attainment by means of achievement tests altogether two further issues need to be briefly referred to: constructivist views on learning and instruction and absolute achievement standards. Recently there is a certain tendency to emphasize mental skills over and above subject-matter mastery in the statement of educational objectives. Partly this tendency is inspired by "constractivist" views on learning and instruction, where knowledge of subject-matter is seen as a "means" to cognitive skill development, rather than as and end in itself. In the field of test development and educational measurement this has led to achievement test with two characteristics:

- testing of specific skills (among them so-called meta-cognitive skills) applied to all types of subject-matter categories;
- embedding test items and problems in real-life situations (this practice is also referred to as authentic testing).

Finally, when achievement tests are used as the criterion in educational evaluations, the additional question of stating *standards* or *"norms"* can also be approached on the basis of empirical methods and specific analysis of data acquired by these methods. Typically such methods depend on interviewing panels of experts (including teachers) about the appropriateness of norms and difficulty levels (see e.g. Van der Linden, Meijer & Vos, 1997).

Apart from achievement tests educational outcomes can be measured on the basis of performance indicators like: the proportion of students of a cohort that succeeds on a final examination in the minimum number of years, the proportion of drop-outs, the percentage of students that had to repeat a grade and proportions of students that reach a particular

position in further education or a particular position on the labor market. In a growing number of countries and/or states within federal countries, public reporting of performance indicators is being practiced as a basis for school evaluation by higher administrative levels and the “consumers” of education.

Measuring inputs, processes and contextual conditions

Inputs to schooling can be categorized as malleable versus “given”, as “human resources” versus material and financial resources and distinguished according to aggregation level: student level inputs, classroom level inputs, school level inputs and context level inputs.

In the table on the next page examples of inputs at four levels of aggregation are presented.

Malleable input conditions are those that are under the direct control of a certain actor, in this case, for example, the school. The school can influence the allocation of resources to sub-units of the organizations, and, to a more limited extent, also the quantity of resources, e.g. by means of active policies to obtain funding and contributions from parents. At the classroom level, through active recruitment policies, the school can influence the teachers’ experience, while teacher commitment can be boosted by means of dealing a productive working climate and by providing incentives.

Many other inputs, particularly those at context and student level are to be seen as “given”, and are largely beyond the control of the school. With respect to pupils schools might exercise selection procedures, although in many national contexts these will be considered as unacceptable.

Context

- achievement stimulants from higher administrative levels
- development of educational consumerism
- the school system at a particular level being categorical or “comprehensive”
- urban/rural

School level

- material resources
- financial resources
- pupil-teacher ratio
- parental pressure/support
- school size
- student-body composition

Classroom level

- teacher experience
- teacher commitment
- class size

Student level

- gender
- age

- nationality
- language spoken at home
- SES
- educational resources at home
- family support
- IQ/aptitude
- previous achievement

"Given" student characteristics such as SES-background and scholastic aptitude have a large impact on student achievement. The impact of human resources inputs and material inputs, like teacher experience on educational attainment is very modest in Western countries, where between school variations on these inputs are relatively low (cf. Hanushek, 1979; Hedges et al., 1994; Scheerens & Bosker, 1997).

Inputs to schooling are a rather heterogeneous category of variables (indicators which require different data-collection methods and types of scaling). Financial resources are based on accounting methods and development of indicators where monetary inputs are usually divided by the number of students. The most important summary indicator is the costs or the available financial resources per student. In principle such an indicator can be computed for each school. Pupil-teacher ratios are computed in a similar way; and it is also possible to present ratios on the school managerial overhead and professional support per student.

Other variables like external conditions and variables like teacher commitment, and home-background characteristics of students will usually be measured by means of questionnaires. The data that are obtained from questionnaires usually yield discrete categories rather than continuous scales, although SES is sometimes measured on a continuous scale. Student background characteristics like previous attainment and scholastic aptitudes are usually measured by means of tests (see the previous sub-section).

Preferably data on classroom level processes should be gathered by means of direct observation techniques, where an observer (who, incidentally, might well be a colleague) is present during lessons. Other methods that have also been used are self-observations, where teachers use structured grids to observe (some) of their pupils, and lags, where teachers are asked to report retrospectively on certain aspects of their lessons.

Case-study methodology like interviews, incidental observations and documentary analysis may be used for measuring school level process indicators on factors like leadership and school climate. Another method that is frequently used, particularly in school effectiveness research, consists of administering structured questionnaires.

Usually such questionnaires ask for self-observations, for example the head teachers are asked questions about their own leadership behavior, teachers about their teaching strategies and students about, for example, their work attitudes. In recent studies (e.g. Bosker & Hendriks, 1997; Hill & Rowe, 1996) multiple actors are asked to report on the same variables, so, for example, school leadership is not only measured on the basis of self-reports but also on observations of teachers and students, while teacher behavior is also measured on the basis of pupil observations and judgements.

2.2.3 Controlling for background variables (value added)

Usually, in educational evaluation, the aim will be to know the effect of the conditions that are malleable by means of policy measures, organizational arrangements or teaching strategies. This is the case when the evaluandum is a program, but also when the overall functioning of a school, as compared to other schools—as in school effectiveness research—is the object of evaluation.

As we saw in earlier sections, outcomes are also determined by "given" characteristics that exist independent of active manipulation. The most important category of these are "given" characteristics of pupils, such as SES and scholastic aptitude. In order to separate the impact of the "program" or "malleable conditions" from the "given" characteristics two things are required:

- first, the relevant background conditions need to be measured;
- second, appropriate analysis techniques will have to be employed to separate both types of impact, and to arrive at "net" estimates of the program effect.

In educational evaluation the most straightforward way to obtain "net" effect estimates is to make adjustments for previous attainment. Conceptually this adjustment may be seen as a measure of learning gain, as when the difference between the score on an entrance test and a test at the end of a period of schooling is computed. Technically, usually other adjustment techniques are used, however. These techniques are based on making a prediction of expected outcomes on the basis of entrance scores on achievement tests or other relevant background variables. The most strictest adjustment uses both background characteristics and pre-test information (cf. Scheerens & Bosker, 1997, p. 54).

Since these procedures try to arrive at estimates of the effects of schooling over and above what might have been expected on the basis of initial abilities of pupils, they are frequently referred to as assessing the "added value" of schooling. More will be said about the use of value added results within the context of school evaluation in subsequent chapters (particularly in Chapter 12).

2.2.4 Design: answering the attribution question

Education, and the functioning of educational organizations, can be abstractly described as a whole of *means* and *ends*. Educational objectives, attainment standards, acquired position after a period of schooling all belong to the "ends" category. Organizational arrangements, provisions of equipment, time investments of educational staff and teaching and learning strategies are all "means" to reach these ends.

According to this distinction educational evaluation can be considered as *meansto-end analysis*, it is not only the question to what extent objectives are attained and standards are reached, but also whether degrees of effect-attainment can be attributed to "means", particularly the "malleable conditions" which were referred to in earlier sections. Means-to-end analysis is logically similar to causal (cause and effect) analysis, and therefore evaluations that go beyond the mere assessment of the attainment of standards, has the nature of causal analysis.

In research methodology causality is addressed by means of the design of experiments. When conceiving of an experiment in education pupils would have to be randomly assigned to two, so-called "treatment groups", an "experimental" group where the pupils

would be exposed to a particular treatment, such as, for example, the use of computer-assisted instruction during math. lessons, and a "control" group where this specific treatment would be absent and math. lessons would be taught in the traditional way.

The random assignment is a crucial aspect of experimentation, namely as a mechanism that should rule out systematic initial differences between the experimental and the control group. In this way mean-differences in attainment, measured at the end of the experimental period, between the two groups, could then be unequivocally attributed to the experimental program. Even though in "true" experiments outcome scores could well be adjusted for initial achievement, this is only to be seen as enhancement of the precision of the causal attribution and not as a substitute of the randomization mechanism.

In many cases in educational evaluation random assignment to treated and untreated groups will be unfeasible, however. In that case "experimentation" is approached by means of "approximations", known as "quasi-experiments" or "ex post facto" research. The term "quasi-experiments" is used for all situations where there is one (or more) treated group, but random assignment to the treatment conditions has not been possible. A frequently occurring example in educational evaluation is when existing, intact parallel classes are compared. Preferably the effect measure, or "post-test" should be adjusted on the basis of pretest information. Particularly when intact groups are used this type of adjustment is especially relevant. To some extent statistical adjustment techniques are used as a substitute for random selection, although biases in the comparison cannot be ruled out, unless the selection process through which the groups got their composition is completely known.

In the case of quasi-experiments there are all kinds of threats to the unequivocality with which causal relationships can be established. In well-known textbooks like Campbell and Stanley (1963) and Cook and Campbell (1979), these threats to the "internal" and "external validity" of quasi-experimentation are dealt with.

"Internal validity" refers to the approximate validity with which we infer that a relationship between two variables is causal or that the absence of a relationship implies the absence of cause. External validity refers to the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measures of the cause and effect and across different types of persons, settings, and times" (Cook & Campbell, 1979, p. 37).

Threats to internal validity in quasi-experiments occur if the conditions of the experiment, such as test administration and the "watertight" separation of treatment groups interact with the treatment, or with characteristics of pupils in the treated or untreated groups. The specialized literature is referred to for further information on such kind of reactive arrangements.

External validity is threatened by selection biases (uncontrolled initial differences between the treatment groups that interact with treatment conditions) and artificial aspects of the experimental situation. This latter feature would apply in the case of laboratory-type of (quasi)-experiments.

In instances when the distinction between treatment groups is made after the fact and not initially, as when a quasi-experiment is planned, research designs are considered as "ex post facto". Technically speaking ex post facto designs may be identical to planned quasi-experiments, depending on the availability of pre-test measures. In the growing

number of applications of longitudinal achievement monitoring, or periodic assessments, such pre-test information is likely to be available and ex post facto comparison of intact groups would be largely similar to conducting a quasi-experiment.

Conditions that might further complicate causal inference would be the absence of pre-test information and the absence of any type of discrete treatment conditions. But even for such situations statistical modeling techniques exist which try to base causal inferences on correlational structures. LISREL-analysis is the best-known of these type of techniques.

It should be noted that in all cases of educational evaluation in which one would wish to go further than merely monitor results and/or describe processes, and try to discover causes for, for example, the lagging behind of a particular classroom in a particular subject, one would come across these problems of causal attribution in non-experimental field research. In such situations the complexities are likely to be so great that schools will need to seek external expert advice.

2.3 Important Distinctions in Evaluation Theory

2.3.1 Ideal-type stages in evaluation

The following stages will be distinguished:

a. delineation of the evaluation objectives and relevant audiences;
b. evaluability assessment;
c. gathering descriptive information;
d. valuing;
e. use.

re a) *delineation of the evaluation objectives and identification of relevant audiences*

Evaluation studies usually have a contractor, or, in any case, an initiator who might not always be the same as the persons or institute that carries out the evaluation. The persons or agencies that initiate the evaluation can usually be addressed as a source concerning the evaluation objectives.

At the stage where an evaluation plan is formed the initiator, together with the evaluator, should work on the statement of the evaluation objectives and the specific role and context of the evaluation. As specific roles (formative and summative roles) and contexts (improvement-oriented and accountability-oriented) will be discussed in subsequent sections, evaluation objectives will be taken in this section as the specification of the evaluation objects, standards and criteria.

Evaluation objects, criteria and standards have already been defined. In this section the process of obtaining clarity on these elements is the focus of attention. In many instances in evaluation practice the objects, criteria and standards of the evaluation remain relatively vague during the initial stage. Usually the focus is put directly on methods for data gathering, while the more normative evaluative framework (standards and norms) remains implicit.

But even the evaluation object, obvious as it may seem, may give rise to debate and practical problems, if it has not been delineated clearly from the outset of the evaluation. For example, it currently takes some convincing of teachers to make it clear that they will not be appraised personally if they provide information that is used within the framework of a program evaluation or a whole-school evaluation. So it is definitely to be preferred to clarify the precise object of the evaluation at an early stage of the (planning of) an evaluation.

When program goals are unclear and fuzzy, which is often the case in larger scale policy evaluations, it is not easy to agree on evaluation criteria, let alone standards. Yet, it is important for evaluation to try and commit initiators/contractors and other stakeholders to clearly delineated criteria and standards. If this does not happen there is a certain risk that the evaluative results will ultimately be disregarded because it could be maintained that they did not address the "real issues". In many instances, where evaluators are in a position where they can by no means force contractors to be explicit in this way, there may be no other way than initially proposing criteria and standards themselves, and then try to attain commitment from stakeholders later on.

re b) *evaluability assessment*

"Evaluability assessment is a diagnostic and prescriptive tool for improving programs and making evaluations more useful. It is a systematic process for describing the structure of a program (i.e., the objectives, logic, activities, and indicators of successful performance); and for analyzing the plausibility and feasibility for achieving objectives, their suitability for in-depth evaluation, and their acceptability to program managers, policy makers and program operators. This is accomplished by:

- clarifying the program in text form from the points of view of key actors in and around the program;
- exploring program reality to clarify the plausibility of program objectives and the feasibility of program performance; and
- identifying opportunities to improve program performance" (Smith, 1989, p. 1).

In this situation evaluability assessment is described as an analytic activity that focuses at the structure and feasibility of the program that is to be evaluated. As such it can be seen as a potent means to provide early warnings on a phenomenon that is known as "non event evaluation", where in fact a program is unlikely to be implemented or have its desired effects. At the same time this initial analysis provides information that is useful to the design of the evaluation study, or—in extreme cases—on the point of whether evaluation will be at all feasible.

The results of evaluability assessments may be fed back to program officers or other stakeholders responsible for the "evaluandum" and, for example, contribute to a sharper focus in both program design and evaluations.

Evaluability assessment may also be more directly focussed at the feasibility of carrying out an evaluation; relevant desiderata are:

- whether the goals of the program to be evaluated are clear and unequivocal, generally supported or contested;
- whether the program is clearly described and implementation is feasible;

- to what extent program goals are measurable and evaluations have sufficient leeway and time to design the evaluation in such a way that the attribution question can be answered;
- to what extent program officers and practitioners have a favorable attitude towards the evaluation, and, e.g. are ready to cooperate when data gathering is to take place;
- whether the evaluator has sufficient independence and professional credibility to survive in a—sometimes—turbulent, heavily political setting;
- whether conditions for the communication and "use" of results are favorable or unfavorable.

re c) *gathering descriptive information*

The key variables on which data should be gathered are to be chosen on the basis of the determination of evaluation criteria and standards (ends) and the structure of the program (means)—see re a) and re b) in the above.

Generally, next to educational measurement, standard methods of social scientific inquiry like testing, survey techniques, observation methods and case study methods, are to be used as methods for information gathering in evaluations. The reader is referred to standard texts and handbooks for a further description of such data-collection methods (e.g. Seltiz, Wrightsman & Cook, 1976). Clarifying the purposes of data-collection methods and specific measures to ensure standardization across units to practitioners is an important organizational aspect of this stage of an evaluation. In many cases specific efforts are required to obtain cooperation and prevent "non-response".

re d) *valuing*

In case evaluation standards and norms have been clearly specified in advance, the evaluative interpretation of the data that has been gathered is straightforward: it will immediately be clear whether or not attainment is above or below the standard.

It should be noted in passing that, in the case of absolute evaluation standards, it is debatable whether control groups are required. A control group, however, would provide more certainty that a "success" was indeed attributable to the program and not to other circumstances.

In the cases where one is dependent on comparisons of treated and untreated, or even just differently treated groups, evaluative interpretation is more complex. When it has not been possible to state evaluation standards in advance, evaluation results may get "thrown in a political arena", where evaluators may not be able to do more than provide some sort of structure or set out "rules of the game". For example by organizing advocacy types of discussions among stakeholders. It is not unlikely that in such situations evaluations become prone to political preferences and biases in interpreting results.

re e) *evaluation use*

As stated at the beginning of this chapter evaluations are undertaken to be used, for purposes of program redesign, changes in school policies, or (more generally) to assist administrative and political decision-making.

There exists a classical ideal-type linear model where evaluation results clearly delineate decision alternatives which are then, in the next step, used by rational decision-makers. Studies which researched phenomena of evaluation use painted a rather different picture, however (e.g. Weiss, 1980). There appeared to be many instances where decision-makers did exactly the opposite from what the evaluation recommended (e.g. terminate a successful program). Also there were many examples of a total disregard of evaluation findings or of a selective, biased use of the results. And situations where it was completely unclear when and by whom decisions were supposed to be taken were no exceptions.

The general background for these types of "mismatches" between evaluation and decision-making is the fact that the linear model of evaluation use, presupposes a rational world of decision-making, which frequently does not appear to exist—at least not in the simple and straightforward way as was presupposed. These contextual conditions will be discussed further in a subsequent chapter.

As a reaction to the remarkable results of these studies on evaluation use, several authors published handbooks on how to enhance evaluation use (e.g. Alkin, 1979; Patton, 1978). Huberman (1987) provides an elaborate model of evaluation use where important aspects are:

- credibility of the evaluators;
- special measures to enhance communication of results to stakeholders;
- relevant characteristics of the agencies that are recipient of the evaluative information, such as: experience with evaluation, attitude towards the evaluation, costs and benefits of the evaluation;
- communication channels and other liaisons between evaluators and users.

An important practical aspect in the use of school(self)-evaluation is the availability of guidelines for "therapies" or remedies that are feasible, depending on the diagnosis that a school evaluation provides.

2.3.2 Formative and summative roles

The terms formative and summative evaluation were introduced by Scriven (1967). Formative evaluation has the function of ongoing assessment during a development process. Summative evaluation has the function of the overall, final, assessment of a program.

When, for example, a new textbook is being developed, it could be formatively assessed at various stages. Firstly, the overall design or outline could be presented to subject matter and pedagogical experts. Next, parts of the book could be tried out in practice on a small scale. Finally, a first edition could also be assessed with an eye to its implementation. In such a situation teachers using the new textbook could be observed during lessons. The results of such a formative evaluation could then be used to modify or elaborate suggested for a proper use of the method for a second edition. Thus the

concept of formative evaluation is closely related to stages of design and development processes (cf. Maslowski & Visscher, 1997).

There is a fine line of distinction between rather implicit, and rather unformalized try-out and feedback stages in normal practice and development of new approaches on the one hand, and formative evaluation on the other. Only if such processes have a degree of systematic information gathering and impartiality in drawing evaluative conclusions they would meet our initial definition of evaluation. Scriven (1991) further reflects on these qualifications and states that formative evaluation should still have a critical potential and even lead to terminating developmental processes in case of "dismissing for incompetence". In principle, therefore, formative evaluation ought to be as methodologically rigorous as summative evaluation. Without denying this principle, practical conditions may often lead to procedures that have a degree of "lightness", "informality" and simplicity which foregoes explicit testing of reliability and validity. Such practical conditions are related to the time scale that is required to provide "early" or "ongoing" feedback and to constraints with respect to costs. Also formative evaluation procedures cannot have a degree of obtrusiveness that would upset the developing process in a structural way.

In the case of summative evaluation the time scale may provide more room for rigorous methodology, like for example, process-output assessments. The type of use of summative evaluation is usually contrasted to the context of use of formative evaluation by stating that summative evaluation is used for overall and "final" decision-making, about the continuation of a program versus guiding development processes in the case of formative evaluation. As the literature on evaluation use indicates, this distinction should not be seen as very sharp, however, since policydecision-making seldomly has the nature of "go/no go" decisions. So, the results of summative evaluation can also lead to a gradual shaping of policy-making and program development.

The distinction made in the next section, between improvement and accountability perspectives again sets these two types of use further apart, as we shall see.

2.3.3 Accountability and improvement perspectives reconsidered

The distinction between formative and summative roles of evaluation is closely related to two major perspectives from which evaluations are conducted, the first perspective is generally known as the "accountability perspective" and the second as the "improvement perspective". Both orientations were introduced in Chapter 1 as 2 out of 3 basic functions of educational evaluation.

In general terms, accountability refers to holding public institutions and services responsible for the quality and output of their performance. Glass (1972) states that accountability involves several loosely connected strands: "disclosure concerning the product or service being provided; product or performance testing; and redress for poor performance (Glass, 1972). The third element implies that accountability is not just a matter of providing and judging information but at least also "foreshadows" actions by competent authorities in the sense of sanctions or rewards.

When evaluation is situated in a context of organizational learning and improvement this last element ("redress for poor performance") is replaced by the assumption of a less

"controlling" type of use of evaluative information, where adjustment has the nature of mutual adaptation and evaluation is formative rather than summative.

Part of the complexities of the fitting of evaluation within the context of administrative or political decision-making have to do with real or alleged conflicts of interest between decision-makers, practitioners and evaluators. The sometimes even antagonistic nature of the relationships between these actors is much more associated with evaluation from an accountability perspective than is the case for improvement-oriented evaluation. In case of an accountability orientation, evaluation is more judgmental and controlling and closely tied to either vertical relationships within an administrative hierarchy or to demands from important external constituencies on which the existence of the organizations may depend. From the outside, particularly in the service sector, accountability is likely to be seen as a rightful requirement of tax-payers and other stakeholders to check the merits of their "investments". From the inside, such control and assessment is oftentimes perceived as threatening or even unjust. The "cognitive" complexity of decision-making and primary processes in the tertiary sector [e.g. education], together with potentially threatening nature of accountability-oriented evaluation may give rise to a distrust of evaluation procedures. Cronbach and his colleagues (1980) go even as far as calling "accountability being dangerously close to totalitarisation". Scheerens (1983) provides a case study of the evaluation of a pilot project in the sector of adult education in the Netherlands. Although in this case evaluators tried to combine "summative" and "formative" evaluation functions, teachers, however, reacted as if the project was only carried out for accountability purposes. They felt so threatened by the partly internal and partly external evaluation activities that a majority refused to cooperate in data collection activities which, of course, had a detrimental effect on the implementation of the evaluation program.

Even if this is considered too strong a qualification, evaluation practices carried out from an external accountability perspective may be hampered or frustrated by behavior of actors involved who feel threatened, when they consider the stakes to be high. It will be interesting to check and see to what extent external school evaluations, indicated from an accountability perspective suffer from such attitudes, and to what extent there are differences between countries which could be grounded in different educational cultures, and different traditions in acceptability of external control. The antagonistic nature of accountability-oriented evaluations in the education sector is supported by organizational theoretical constructs such as the "professional bureaucracy" (Mintzberg, 1979), of which resistance to rationalization and external review is one of the defining characteristics. Educational professionals, i.e. teachers may contest the expertise and methodology of evaluation researchers, as when they consider themselves as the best medium and service of knowledge on what goes on in classrooms.

Evaluation from an improvement perspective has an altogether different orientation. Here, learning, feedback, a formative role of evaluation, intrinsic interest in processes and a methodology that is controllable by teachers are the central characteristics. External school evaluation is more likely to be accountability-oriented and internal evaluation is more likely to be improvementoriented, although exceptions may occur, as when a school deploys an external consultant to review, for instance, its managerial structure.

While "control" is the key-term in accountability, "learning" is the key term in internal improvement-oriented evaluation. A question of internal "self evaluation", however is

whether, it can be sufficiently objective and impartial to provide a firm basis for such learning processes.

Examples of accountability-oriented school evaluation are the public reporting of average pupil achievement tests in public media and the composition of so-called “league” tables. The use of such approaches is much contested. The most severe criticism is that such procedures could stimulate selection processes which are detrimental to the principle of equality in education. Thus, schools might seek to enhance their average output, not by improving the school organization and teaching, but by attempting to obtain a “favorable” input of students. Particularly when school-scores have not been adjusted for intake or prior achievement, there is a danger that such practices might occur. But in addition, there might be other selection processes, having the same type of effect. Stronger schools with “better” intake could become more attractive to better teachers. And parents from more privileged socioeconomic backgrounds would be more likely to profit from the information that is made available by the publishing of school-scores.

Types of improvement-oriented school evaluation are school-based review procedures which are part of school improvement projects and school selfevaluations which monitor the “normal” functioning of a school.

3
Schematic Description of 15 Types of Educational Evaluation

3.1 Introduction

In this chapter each of the 15 specific types of M&E is briefly described for its general characteristics. In subsequent chapters each of the main data strategies (student assessment, monitoring on the basis of indicators and review and research methods) will be further characterized with respect to the use and follow-up of the M&E activities by different audiences and for the technical and organizational capacity that is required for a proper application.

Table 1.1 in Chapter 1 refers to 15 types of educational assessment, monitoring and evaluation:

1. national assessment programs
2. international assessment programs
3. school performance reporting
4. student monitoring systems
5. assessment-based school self-evaluation
6. examinations
7. system level Management Information Systems
8. school Management Information Systems
9. international review panels
10. school inspection/supervision
11. school self-evaluation, including teacher appraisal
12. school audits
13. monitoring and evaluation as part of teaching
14. program evaluation
15. school effectiveness and educational productivity studies

Each of the M&E types that are mentioned in Table 1.1 is briefly described on the following aspects:

General description
Main audiences and types of use of the information
Technical issues
Technical and organizational capacity required
Controversial issues

3.2 Forms That are Based on Student Achievement Measurement

3.2.1 National assessment programs

General description

Assessment programs consist of educational achievement tests that are meant to monitor acceptable levels of performance in the basic school subjects in a country. Likely age-levels at which the tests are taken are 11/12, (end of primary school), sometimes also 14/15 (end of lower secondary school). Assessment tests in a particular subject need not be administered each year; for example, when there are 6 subjects in the assessment program, each subject may be tested every 6 years. Application of multiple matrix sampling, however, makes a more frequent testing (shorter time interval) for each subject matter area feasible. Typically national assessment programs will target samples of students. Depending on whether conclusions about schools as a particular organizational level would also be aimed for, sampling design would need to accommodate this by ensuring a sufficient number of students per school.

Main audiences and type of use of the information

Decision-makers at the central level, i.e. the Ministry of Education, parliament, organizations representing stakeholders in education like school governors, teachers, parent association, employers are also relevant.

The information from assessment programs can lead to adaptations in the curriculum in the sense of goals (standards) or means (curriculum contents) and all conditions that have an impact on the performance in a particular subject (e.g. teacher training in the particular subject matter area, the textbook-industry, use of computers).

Technical issues

Norm- versus criterion referenced testing. Procedures for standard-setting. Psychometric properties of the tests, in particular the content validity of the tests (do test adequately represent the universe of subject-matter elements in the specific curriculum domain). Sampling issues; not all students need to do all tests, application of multiple matrix sampling (a technique where students do sub-sets of items of comparable content and difficulty level).

Technical and organizational capacity required

Skill-areas that should be covered are: subject-matter expertise, skills in curriculum analysis, skills in writing test-items. Expertise in psychometrics, methods of standard settings, sampling expertise. Communicative and PR skills in disseminating information to decision-makers and the education field.

Concerning the organizational infrastructure the degree to which specialists in subject-matter and subject-related didactics are organized in special interest groups is relevant for mobilization of this expertise. The same applies to curriculumdevelopment institutions.

Depending on the size of the country a specialized institute like ETS in the USA or CITO in the Netherlands could be considered, at least a specialized unit as part of the "techno-structure" of a Ministry of Education would be required. In case of a smaller assessment unit organizational links with curriculum and subject-matter specialist units is very important. Technical support concerning logistics of distribution and retrieval of test-material from schools, dataanalysis and reporting should also have a place in either a specialized institute or a network with sufficient cohesion. Boards of officials and experts should be created to authorize newly developed tests.

Controversial points

Controversy about national assessment programs can arise with respect to the scope of what is being measured. The often-heard argument is that important goals of education cannot be measured. Also the issue, referred to in the above, of curriculum-tied, as compared to "cross-curricular competencies" can be a controversial issue. In developing countries expectations about low performance as compared to industrialized countries might be a difficult point.

3.2.2 International assessment programs

General description

Over recent years there has been an increased interest from governments and international organizations in international assessments. Examples are:

- the Third International Mathematics and Science Study-Repeat of the IEA (TIMSS-R);
- the Civic Education Study (CivED) of IEA;
- the OECD Program for International Student Assessment (PISA)
- the IEA Progress in Reading Literacy Study (PIRLS) and
- the Adult Literacy and Lifeskills (ALL)Study (formerly ILSS).

There are two major advantages for taking part in these international assessment studies. The first is practical: if a country does not already have a national assessment program, important developmental costs can be foregone by making use of the internationally available instruments. This can be the case, even if instruments are modified or extended according to the specific national circumstances. The second potential advantage is the opportunity to compare national performance levels to international standards. This application of comparative "benchmarking" could be seen as an important feature of accomplishing globalization of educational provisions. Of course this possible advantage of international standardization can also be seen as an undesired uniformity. Perhaps a compromise could be found in defining a set of core competencies, which would be expected to meet international performance standards next to a set of more country-specific, or region specific standards.

Main audiences and types of use of the information

These are more or less the same as in national assessment programs.

Technical issues

Making tests internationally comparable is the biggest challenge for international assessment programs. The range of difficulty levels on the scales should be sufficiently broad to cover potentially large differences in achievement levels between countries. IRT-modeling is important for this. Remaining comparability problems can be tackled by means of national options and "add on-s" and by measuring test-curriculum overlap or "opportunity to learn".

Technical and organizational capacity required

Much of the technical capacity for international assessment programs will be located with the international study-coordinating organization, which may be a consortium of top-level institutes at the global level.

Usually at national level a small team with the required research-technical skills and logistic facilities is sufficient to carry out the work at national level.

Controversial points

The main controversy had already been referred to in the above: can specific national priorities sufficiently be represented in international test programs.

3.2.3 School performance reporting

General description

School performance reporting (SPR) is a prototype of accountability oriented assessment. It uses statistical information and/or achievement tests to generate output indicators per school. These are then made public, as, for example, in the form of "league tables" (rankings of schools) that are published in the newspapers.

The achievement test data used for SPR could have various sources:

- national assessment tests;
- tests used in student monitoring programs (an M&E type that will be described further on);
- examinations.

Examples of statistical performance indicators are the success rate (e.g. finalizing the period of schooling without delay), average absenteeism, drop-out and classrepetition rates).

An important issue is whether or not output indicators should be adjusted for previous achievement or other relevant student background characteristics (the issue of "value-

added" output indicators). Another question is whether or not school process or input indicators should be included in the school reports.

Main audiences and types of use of the information

The results of SPR is meant to be used by administrative levels above the school, like municipalities, regional and central government and/or by the consumers of education. In countries with freedom of school choice, parents could make use of this information to select a school for their children.

Decisions about school funding could be made dependent on the results of SPR. Next, different "markets" might use the information for either selecting a particular school or not: markets of parents choosing schools, markets of teachers choosing a school and schools actively marketing themselves with respect to these audiences.

As a "side-effect" schools might also use the information from SPR to diagnose their own performance and use it to target improvement oriented measures. In fact, empirical results indicate that this latter use may be even more important than the accountability oriented uses (cf. Bosker & Scheerens, 1999).

Technical issues

Computing value-added performance indicators is a technical problem both in the sense of statistical analysis as in terms of communication. Although the value-added option may be considered as the fairer one to judge schools, its meaning may be difficult to communicate to broad audiences. Besides, "raw" outcome scores are also informative.

Technical and organizational capacity required

This is highly dependent on the provisions for the basic assessment, monitoring and evaluation types that SPR is likely to depend on. If these are in place a relatively small research team, which contains a unit of data-analysts, would be sufficient.

Controversial points

SPR is quite controversial as it can be seen as stimulating selection mechanisms that are not easily reconcilable with the ideal of equity in education. When the stakes for schools are made high, undesired strategic behavior to artificially create higher scores are likely to occur.

3.2.4 Student monitoring systems

General description

Student monitoring systems operate at the micro level (class level) of educational systems.

Basically student monitoring systems are sets of educational achievement tests that are used for purposes of formative didactic evaluation. An important function is to identify those pupils who fall behind, and also to indicate in which subject matter areas or skills they experience difficulties.

Items should preferably be scaled according to a particular IRT model. Student monitoring systems should be longitudinal and allow for the "following" of students throughout a particular educational program. For example in the Dutch *Leerlingvolgsysteem* for primary schools two tests per grade level are administered. The scope of a student monitoring system depends on the number of school subjects that are covered.

Main audiences and types of use of the information

Student monitoring systems are used in the interaction between teachers and students. Apart from the achievement tests *remedial material* should be seen as the major component of a pupil monitoring system. One type of remedial follow-up material consists of guidelines for further diagnosis of deficiencies. Exercises to remedy deficiencies form another. Such exercises take the form of performance tasks to stimulate learning.

Technical issues

Test construction is an important technical issue. Because of the intended longitudinal use of the instruments "vertical equating" is an essential asset. This requires scales that confirm to the assumptions of IRT models. An important precondition for the curriculum validity of the tests is that there is at least consensus about the educational objectives at the end of the program. If prescribed curricula do not exist they need to be "reconstructed" in the form of a sequence of subject-matter areas for each subject, which in their turn form the basis for the development of test items and remedial tasks.

Technical and organizational capacity required

Technical and organization capacity requirements are basically similar to those for national assessment programs.

Controversial points

The same kind of controversies might arise as in national and school assessment programs, primarily the criticism that important educational goals would escape measurement. In settings where schools are given complete autonomy in establishing the curriculum, these, and other student assessment instruments, could be seen as letting in centralization through the back door. If one accepts fixed educational objectives and national item banks one should probably also be ready to accept the equalizing tendency that the assessment tools would inevitably have. "Teaching to the test" would only deserve its negative connotation if the test was fallible and the item bank too small. The

flexibility and quality of tests developed according to the current state of the art methodology should be able to prevent this.

3.2.5 Assessment-based school self evaluation

General description

This type of M&E is best perceived as a spin-off of other assessment types. The core idea is that schools use the information from externally initiated assessments or from internal student monitoring systems to evaluate their own performance. There are nevertheless also examples of projects where school self-evaluation appears to have been the primary motive for the development and administering of achievement tests.

Main audiences and types of use of the information

School managers and the school staff are the main category of users. Parents could also be a target group for disseminating the information to.

Following the achievement of cohorts of students in the main subjects would allow schools to monitor their own standards and detect problems in a particular *time x grade x teacher x subject* combination. Follow-up actions might involve adapting the school-curriculum, choice of textbooks, initiatives for counseling and consultation of teachers, and decisions about matching teachers and groups of students.

Technical issues

Psychometric quality of the achievement tests is relevant to the possibilities for their use. The issue of criterion—versus norm referenced testing is relevant in this context as well. Additional technical problems arise when it is the ambition to relate information on process indicators to the assessment results and indices of learning progress (Scheerens & Bosker, 1995). These technical issues concern additional data collection, developing appropriate data-records, and problems of data-analysis.

Required technical and organizational capacity

Until fully computerized forms become available schools would require the assistance of assessment specialists and data-analysts to compute statistics, make comparisons over time, and (possibly) link the information to other data-sources.

At school level specific organizational pre-conditions need to be fulfilled in the sense of established discussion platforms and clear rules about the way the information will be used. Confidentiality is an important issue.

Controversial issues

Controversial issues are similar to other types of achievement-test based assessments. The implied “multi-purpose use” of instruments is not unproblematic. For example, from

the perspective of School Performance Reporting the test results are to be made public, while, for school self-evaluation purposes, schools might give preference to keep part of the information confidential.

3.2.6 Examinations

General description

Examinations are sets of learning tasks or test items and specific procedures to administer these (e.g. written and oral exams, portfolio's showing samples of accomplishments). These are used to determine whether a candidate has the required level of achievement to be formally certified as having successfully completed a program of formal schooling or training.

Main audiences and types of use of the information

Examinations belong to the institutional arrangements of a country and regulate selection for follow-up education and entrance to positions on the labor market.

Technical issues

A major technological question is whether examinations can fully depend on objective and standardized achievement tests, or need other review procedures and demonstration of skills as well. By allowing for the objective scoring of open test items, tests on more general cognitive skills and "authentic testing" the achievement test methodology appears to be "moving up" in taking care of these more complex aspects. Therefore tests will probably play an increasingly important role in examinations. Organizational forms like allowing for a school-based part and a central part of a final examinations could allow for combining more holistic and informal review by school-teams and objective testing (the central part of the exam).

Technical and organizational capacity required

Assuming that examinations will, at least partially, be based on standardized tests, the required technical capacity in a country matches that for other applications of educational achievement tests.

In addition, examinations require committees that take the formal responsibility for each annual version of the examination. Sometimes the educational inspectorate has a function in this as well.

Of course there should also be technical and logistic facilities to score test-forms, possibly combine test-results with the results of other parts of the examination etc. Again state of the art ICT applications, like for example optical readable test-forms, are relevant.

Controversial points

Perhaps the issue of norm-referenced versus criterion-referenced testing applies to examinations more than to other student assessment forms. Traditionally examinations have been norm-referenced. The main draw-back of this being that norms would differ across years and cohorts.

3.3 Forms That are Based on Education Statistics and Administrative Data

3.3.1 System level management information systems

General description

Management information systems depend on indicators. Educational indicators are statistics that allow for value judgements to be made about key aspects of the functioning of educational systems. To emphasize their evaluative nature, the term "performance indicator" is frequently used.

Included in this definition of educational indicators are:

- the notion that we are dealing with measurable characteristics of educational systems;
- the aspiration to measure "key aspects", be it only to provide an "at a glance profile of current conditions" (Nuttall, 1989) rather than in-depth description;
- the requirement that indicators show something of the quality of schooling, which implies that indicators are statistics that have a reference point (or standard) against which value-judgements can be made.

As indicated earlier, indicator systems are based on a particular model of the education system The context-input-process-output model that was used to categorize educational content in Chapter 2 is a useful tool to categorize education indicators. When all indicators are formulated at the level of the national education system this is referred to as a system level indicators system. When multiple aggregation levels are used, and different categories of indicators can be related, because they are collected on the same or connected units, we speak of an integrated multi-level indicator system. Concrete examples of both types of indicator systems will be presented in Part 4 of this book.

Main audiences and types of use of the information

Education indicator systems provide a picture of the overall state of affairs with respect to the functioning of education systems. The term "management information systems" implies that this information is to be used for policy and management decisions. For both variations in indicator sets (system level indicators and an integrated multi-level system) the central government and its bureaucratic apparatus is the main client and user. In the case of a multi-level indicator system the information could also be fed back to lower levels in the system. In principle a MIS can also be designed at the state- provincial or municipal level with corresponding audiences.

According to Table 1.1 in Chapter 1 a MIS has a function for accountability as well as for organizational learning and diagnosis.

Technical issues

Educational statistics form the base-material for education indicators. Traditionally education statistics are more readily available in the area of educational inputs, like financial data, number of teachers, and number of schools. Information on the flow of students through the system, participation rates and success rates per age-cohort are also required. Indicators on the stock of teachers and human resources in general are not usually available, but would also be important. Indicators can depend on "single" statistics or be composites of several statistics, like for example student/teacher ratios or "school overhead", measured as the proportion of administrative personnel relative to all staff or all students.

Integrated multi-level indicator systems have additional technological demands attached to them, like data collection on the same or "nested" units, and an integration of statistics with survey-based and/or student assessment data.

Innovative aspects

Thinking in terms of MIS and indicator systems provide an interesting challenge to national statistical bureaus. Innovation is a slow process of gradually covering all main categories of the theoretical model of input-process-output-context indicators with statistics for which basic data is collected at a regular basis. International education indicator projects, like those of the OECD (INES-project, with Education at a Glance as the central, annual publication) and the European Union (the publication of indicators in "Key Data") can stimulate these developments at national level.

Technical and organizational capacity required

A MIS requires a bureau for Education Statistics with a specialized unit for developing indicators in the domains where traditional statistics do not fully cover all categories of the theoretical model.

In the case of integrated multi-level indicator systems additional methodological skills are required, encompassing survey and assessment methodology.

Controversial points

The level of detail of an MIS should be matched to the pattern of centralization and decentralization of the educational system in a country. Mismatches could lead to controversy over, for example, "centralistic tendencies" in an otherwise decentralized setting.

3.3.2 School management information systems

General description

School management information systems have been inspired by similar systems in private industry. Generally they consist of a careful modeling of information streams and information needs within a company, deciding which data should be available for which purpose on a more or less permanent basis, followed by design and implementation of a computer configuration and software. A school management information system is described as an information system based on one or several computers, consisting of a data-bank and one or several software applications, which enable computer-based data storage, data analysis and data distribution.

A question that could be answered by means of such a school management information system would be: “to which degree has absenteeism decreased after the implementation of specific measures to diminish absenteeism?”

Management information systems have a great potential for supplying important information on a routine basis. At present there are still quite a few practical barriers. One needs to have sufficient and adequate computer hardware and even when professionally developed software packages become available, quite a few specific maintenance functions must be carried out, while new routines and perhaps even new job-functions to guarantee adequate data-entry should be developed.

Main audiences and types of use of the information

Although it would be conceivable to place the use of a school MIS in an accountability context, as for example when a school district or municipality would require specific information yielded, school-based use for purposes of school selfevaluation is more likely. The information could be used for all kind of corrective actions in the running of the school. School-leaders and maybe also the staff-team of the school are the most important user category.

Technical issues

Construction of a computerized school MIS requires considerable development costs. Development consists of providing functional specifications of information flows, developing adequate software, and development of manuals. The total costs of the development of an MIS for secondary schools in the Netherlands, toward the end of the eighties, were about US$ 500,000. Development up to the stage of a commercially exploitable prototype took about 5 years. Currently there are many commercial products available. Empirical studies indicate that the systems are frequently “underutilized” by schools.

Technical and organizational capacity required

Introduction requires adaptation in the school-management style, adequate communication platforms and getting used to the ICT aspects. Possibly new roles or job-functions need to be implemented.

Controversial points

Implementation problems may tip the cost-benefit balance in the negative direction. The technological image and formalized methodology may not match well with the school culture.

3.4 Forms That are Based on Systematic Review, Observations and (Self)Perceptions

3.4.1 International review panels

General description

International review panels consist of education experts from a variety of countries that carry out a systematic review of the functioning of a national education system. The panels are usually supported by a technical staff from the host-country. Primary data that are used as input for the review consist of education statistics, indicators and research reports. In some countries annual indicator-based reports may exist on the "state of education"; sometimes such annual reports are composed by the education inspectorate.

The methodology that is followed by the review panel consists of working though the documentary material, interviewing key-persons and site-visits to schools and other relevant organizations. Examples are the "OECD-examiners" who review the functioning of education systems of member countries.

Main audiences and types of use of the information

Governments and Ministries of Education are the main target group for these international reviews. But their reports usually find their way to other stakeholders inside and outside the education province as well.

Review panels can be seen both from the perspective of accountability, as the education system puts itself open to review and criticism to other sectors of society, and from the perspective of diagnosis and organizational learning as conclusions will most likely be followed up by recommendations for improvement.

Technical issues

Review panels depend to a large extend on the availability and quality of the basic data and documentation. If this is largely absent the method would be difficult to apply.

Technical and organizational capacity required

Countries should be able to put together a good support team that provides the basematerial for the review. Field visits and interviews should be properly planned in order to carry out the visitation as efficient as possible.

Controversial points

Given a systematically conducted review and panel members with strong international reputation governments should be ready to take the results to hart, also in the case of politically unwelcome conclusions.

3.4.2 School inspection/supervision

General description

The core activity consists of school visits carried out by the inspectors/supervisors. The range of aspects of school functioning that is reviewed may vary, ranging from a rather formal contact with the school director to classroom observations and talks with students. In several European countries the work of education inspectorates has become more systematic over the last decades in three ways:

- in the sense of a plan to visit all schools in the country with a specific frequency (e.g. once every two years);
- in the sense of standard setting;
- in the sense of using more systematic, research-like methods of data-collection, i.e. systematic observation in schools and classrooms.

Main audiences and types of use of the information

Inspectorates usually have a dual function. On the one hand they are to inform authorities, like central or regional education officers, on the other hand they are usually also seen as a kind of counselors to schools. Emphases between these two functions may differ between countries. For example, in the United Kingdom inspections are predominantly placed in an accountability context, whereas in the Netherlands accountability has less emphasis and school reports have more of a formative function to the schools.

Technical issues

The definition of evaluation criteria and standards is one technical issue, systematic observation by inspectors is another. As to the former specific consensual procedures are used, involving different kind of educational expertise to define key areas of evaluation (criteria) and, next, norms of acceptable performance (standards) on these criteria. Typically inspectorates will not only consider outcome indicators but they will also look at input and process indicators of school functioning. How to value these processes, when

their explicit association with outcomes is uncertain is a difficult issue in evaluation methodology.

The challenge is to combine systematic procedure and standardization in datacollection with the expert judgement and "clinical look" of school inspectors. Reports may have a quantitative and qualitative component.

Finding a way to inspect all schools in the country with a certain frequency is a technical and practical issue which is of course related to the numerical capacity of the inspectorate and the intensity of school visits.

Technical and organizational capacity required

A school inspectorate, or a structure of school supervisors, which, for example, might be in regional offices or regional educational support centers, is an important institutional and organizational facility in a country's education system. It requires experienced educators with basic knowledge about systematic data collection methods and evaluation methodology in general.

Controversial points

In situations where building up an inspectorate from scratch is considered, given the high costs, a careful analysis of possible alternatives should be considered (e.g. a combination of a national MIS and stimulation of school self-evaluation). A more loosely organized network of School Panel Inspections exists, for example, in Jamaica (cf. Scheerens, 2002).

3.4.3 School self-evaluations, including teacher appraisal

General description

School self-evaluations are *internal* evaluations of the school as a whole, or of subunits within the school, aimed primarily at school improvement. In fact there is a gliding scale from "completely internal" to extensive use of external capacity in school self-evaluation. The decisive point being the condition that the school is the initiator and the prime-audience of the evaluation.

There can be several different methodological emphases in school selfevaluation. In previous paragraphs "assessment-based" school self-evaluation and the use of "school management information systems" were discussed.

A third major methodological emphasis is known under headings like schoolbased review or school-diagnosis. School-based review depends heavily on opinions of school personnel on discrepancies between the actual and an ideal state of affairs in schools. In this way a broad perspective, in which all the main aspects of school functioning can be scrutinised, is possible. Usually, respondents are also asked to indicate whether a certain discrepancy should be actively resolved. This approach to school self-evaluation seeks to gear improvement-oriented action to appraisal. The context of application is usually school improvement, which means that a school-based review is carried out when there is a prevailing commitment to educational innovation.

Advantages of this approach are: a broad scope, a user-friendly technology, an explicit linkage between evaluation and action, and a high degree of participation (all school personnel take part in the review). A definite weakness of school-based review is its dependence on subjective opinions and its (usual) neglect of "hard" factual data on school functioning, most notably output data. Examples of procedures for school-based review are the GRID and GILS-systems (see Hopkins, 1987).

A fourth approach seeks to provide schools with self-evaluation instruments (questionnaires) that meet scientific requirements of reliability and validity. In this way the subjectivity in the self-appraisal can be countered; this is also accomplished by using ratings of different categories of respondents on the same phenomena and by comparing the results. This is an example of "triangulation", a procedure that was developed in qualitative research methodology.

In a fifth approach teams of colleagues from other schools visit schools and do a review of the school's functioning. This approach of "visitation committees" and peer review can also be a formal part of accountability oriented appraisal.

Finally, several of these forms of school self-evaluation can be combined and integrated with one another. For example assessment information on outcomes, administrative data on student background characteristics and self-reports on the functioning of key processes could be carried out in conjunction, possibly relating the various types of information to one another to obtain more insightful diagnoses.

Teacher appraisal can be a specific emphasis in school self-evaluation. The important issue being that the empirical methods of school self-evaluation would offer a more factual basis to assessments by the school director.

Main audiences and types of use of the information

School management and staff teams are the major audience for the results of school self-evaluation. It is not unusual, however, that (parts) of the results are also presented to other constituencies, like higher administrative levels or stakeholders in the local community. Feedback may also be focused at individual teachers, or subteams, like departments within a school.

Information can be used to redesign school development plans, preferred teaching strategy, grouping of students and targets for professional development of teachers.

Technical issues

All aspects of the science and the art of evaluation of human service programs are also relevant in school self-evaluation. How to combine objectivity and commitment? How to deal with demands for openness to external constituencies and demands for confidentiality and mutual trust at the same time? How to make an efficient choice of the many methodological options? How to deal with resistance to and fear of being assessed? An additional practical consideration is how to find the time for school self-evaluation. A viable option seems to be an integration of new assessment and monitoring forms with other developments like curriculum redesign and changed teaching strategies.

Technical and organizational capacity required

School teams need to be trained for most types of school self-evaluation. For some forms in particular, external support is required on a more permanent basis (e.g. testdevelopment, data-analysis, data-feedback).

Carrying out school self-evaluations requires that there are certain communication platforms operational within the school. Support by the school head or director is an important pre-condition.

Controversial points

The key question that remains is the objectivity of self-evaluations. From a very strict methodological position "objective self-evaluation" could even be regarded as a "contradictio in terminis". On the other hand objectivity can be supported by providing instrumentation that meets scientific criteria. The degree to which autonomous functioning of schools is a priority in an educational system and "quality care" is decentralized to the school level, is also relevant to this discussion.

3.4.4 School audits

General description

As educational institutes (schools and universities) are made to function more autonomously, they may become more like private companies in their managerial and organizational characteristics. An example of this would be a stronger emphasis on strategic planning and on scanning the external environment of the school. It is therefore not surprising that approaches used in management consultancy are introduced in schools. A strong point of these approaches is that it is likely to pay attention to issues that were kept largely unnoticed by the educational province, such as external contacts, anticipation of developments in the relevant environment, and flexibility in offering new types of services.

Screening the organization for quality in its internal and external functioning can be formalized on the basis of quality systems and norms for organizational accreditation like the well-known ISO norms. In this way schools and universities can be formally accredited.

Since this practice is still fairly uncommon, it will not be further developed here.

3.4.5 Monitoring and evaluation as part of teaching

Informal "formative" assessment of students 'performance and progress has always been part of regular teaching. Students do assignments that are marked, and teaching methods contain progress tests.

Similarly teachers "keep order" and monitor the behavior of students in classrooms. This aspect of normal teachers' work should not be overlooked, as it is can be seen as the basis for the application of more formalized forms of assessment and monitoring. To a

degree teacher have always been "reflective practitioners" and evaluating student performance was not invented with the first multiple-choice test.

This point is only made as a reminder that the principle of really-testing and feedback that is at the center of the motivation to enhance and stimulate educational M&E has always been there as an important principle of "good practice" in teaching.

3.5 Program Evaluation and Teacher Evaluation

3.5.1 Program evaluation

General description; the distinction between monitoring and program evaluation

First of all it is important to point at a gradual difference between monitoring and evaluation. In the case of monitoring at various phases of the progression of project events descriptive information is provided. This descriptive information, when compared to the intended progression of events, can be used for value-judgements and provide crude indications about where corrective actions are required. Macro level indicator systems which provide information on the use of project inputs, outputs and outcomes are appropriate tools for monitoring.

In the case of evaluation, in the sense of program evaluation, there is an additional ambition concerning causality. *Can the outputs that are measured be attributed to the project, or are they due to other circumstances?* Program evaluation requires a refinement in methodology concerning the unequivocal attribution of measured outcomes to the project activities. This refinement can take the form of controlling for biased interpretations of measured outcomes (i.e. selection bias) or for pseudo "treatments" or faults in the implementation of the project. (See the literature on the "internal and external validity of (quasi)experimental designs).

To the degree that indicator systems are more specific, use disaggregate data, and are more comprehensive (in the sense that input, process, output and outcome indicators are available on the same project) they allow for a type of monitoring that approaches the ideal of program evaluation. This means that such more comprehensive indicator systems can provide hints about why projects do or do not reach these objectives and in this way are more informative for corrective actions. (See for more information on comprehensive, multi-level indicator systems, Part 4 of this book).

The implication is that integrated, multi-level education indicator systems, although less perfect than carefully designed field-experiments, may provide a viable alternative to fully fledged program evaluation.

In preparing the design of program evaluation a systematic analysis of the intended program in terms of goals and means and assumptions about the causal relations between these is important to the choice of variables. Such analyses have evaluative relevance in their own right as they provide indications about the degree of realism and likely success of the program. Such pre-analyses are sometimes referred to as *analytical evaluation;* in other contexts they are seen as part of *evaluability analyses* (cf. Smith, 1989).

Main audiences and types of use of the information

Depending on formative or summative orientation in program evaluation the results are to be used to modify aspects of program implementation or for major decisions about program continuation, respectively.

Technical issues

Major technical issues in program evaluation are:

- establishment of a credible causal model for the program;
- a valid operationalization and measurement of program goals;
- a design that guarantees internally and externally valid conclusions about the attribution of effects to the program.

Technical and organizational capacity required

Teams to carry out program evaluations would require:

- a senior researcher with both research-technical and educational expertise as evaluation coordinator; he or she should also take responsibility for communication with all actors and stakeholders involved;
- one or more researchers to choose and develop instruments, plan and monitor data-collection and analysis and produce reports in conjunction with the coordinator;
- a data unit responsible for the logistics of data-collection and -retrieval, datacleaning and -analysis.

Controversial points

Results of program evaluations may lead to political disputes when the results are critical, the stakes in the program that was evaluated high, and the credibility of the applied research-methodology less than optimal.

3.5.2 Teacher evaluation

Traditionally quality control concerning teachers has depended on professionalization and certification of teachers. This type of “input”-control has been one of the most important measures for quality care in education for a long time; particularly when combined with another type of input control, namely centrally standardized curricula.

This model is still predominant in quite a few countries; the best example in Europe being Germany. The relative autonomy and exemption of external interference that teachers enjoyed in this context was often enforced by teachers unions resisting more flexible conditions of labor and a more performance oriented management style.

This traditional state of affairs is slowly eroding under the influence of a higher sensitivity to externally felt needs for change and adaptation in education and as a consequence of decentralization policies. The combined impact of these influences is a higher scrutiny concerning the qualifications and actual performance of teachers. On the

one hand this has led to a sharpening of input control measures, like in the case of those US States which have implemented quality standards for teachers. On the other hand there is a gradual development in assessing “on the job performance” of teachers. The most visible form of this, at least in some European countries (the UK and the Netherlands, for example), is that school inspectors systematically evaluate samples of lessons. Sometimes just to get an overall picture of the quality of education in a particular school, but sometimes also to evaluate individual teachers.

PART 2
Theoretical Foundations of Systemic M&E

4
The Political and Organizational Context of Educational Evaluation

4.1 Introduction

In this chapter evaluation is seen as one of the rational techniques of policy-analysis. To the degree that the actual political and organizational context of evaluations differs from the rational ideal specific measures have to be considered to maintain standards of accuracy and utility. Contextual problems of evaluations are discussed when considering phase models of educational reform and articulation of the decision-making context. In its turn the relevant decision-making context of evaluations depends on the patterns of centralization and decentralization of educational systems. In the final sections of the chapter various strategies for improving the institutional, organizational and technical context (in the sense of a technical infrastructure for educational monitoring and evaluation) are discussed.

4.2 Rationality Assumptions Concerning the Policy-Context of Evaluations

In many applications of educational evaluation, e.g. program evaluation or school evaluation, the evaluation object or evaluandum, is a real-life setting in which (educational) goals are striven for through specific practical activities. Such practical situations can be abstractly described as a set of means and goals.

Goals are particularly important for all kinds of evaluations ranging from program evaluations and school evaluations to the construction of examinations. Goals can be seen as “desired states” or “ideal type processes”, which in their turn can be used as targets and evaluation standards. For example: a certain level of attainment on a math score by at least 80% of the pupils, or use of the computer for at least 20 minutes during 80% of the language lessons. Moreover, goals need not necessarily be defined in such a precise, operational and quantitative form. Even when there is just a general notion of the dimensions on which an existing situation should be improved after a period of program implementation, or, in our case, schooling, we could still see the situation as goal-oriented and assessable. In the latter case an expert committee could be used to make the assessment. Generally, when the evaluation criteria remain more global and “open”, the requirements on the substantive expertise of evaluators should be particularly high, as they could be seen as replacing the rigor of otherwise applicable structured and standardized instruments.

The presence of goals, specific or general, is an important feature of what can be referred to as the formal rationality of the evaluation setting. Where we could take "evaluation setting" as both the evaluation object and the larger context in which this object and the evaluation itself is taking place. Evaluation itself can be seen as part of the rationality model applied to policy programs or to the functioning of schools.

The main features of this rationality model can be stated according to the following points (note that the concept of "program" which is frequently used in the points stated below should be interpreted in a broad sense, including, for example, a particular phase during regular schooling):

- the program to be evaluated has goals, and the evaluation can be guided by means of these goals;
- the program itself is to be seen as a set of means, for which there exist some reasoning with respect to the likelihood that they will indeed lead to goal attainment;
- planned, or "blue print" programs are also implemented according to plan;
- evaluation has the general form of empirical investigation of whether goals are attained on the basis of the program, i.e. the implemented set of means;
- evaluation activities can be carried out in a relatively undisturbed and unbiased way, free from all type of influence from parties with certain interests, and be conducted according to professional norms (i.e. standards of good evaluation practice);
- the results of the evaluation will be used for decision-making, which may be of a "formative" or "summative" nature, and in this way practice will be improved.

Oftentimes the last point does not occur so straightforwardly. Evaluation use is often of a less "linear" and "instrumental" nature but rather a gradual and fuzzy process of influencing conceptions of relevant actors.

In a more general way one could say that to the degree that the evaluation setting departs form the rationality model, evaluation, in the sense of systematic, disciplined inquiry, will become more complicated.

Turning back to the goals requirement, it is one thing to say that goals may be general. But, in a situation where goals are contested among relevant stakeholders, such as, for example, teachers and head teachers, it would be more difficult for evaluators to find a point of reference for designing an evaluation. In the context of large-scale policy evaluations the situation that interested parties differ about the program goals is not at all unlikely. Sometimes, there can also be large discrepancies between the official goals and the "real" goals of stakeholders. For example, in an experimental program for adult education, the official goals stated by the Ministry of Education were stated in terms of learning gains among participants, but for the teachers in the experimental program, prolonged employment and long-term tenure appeared to be more important (Scheerens, 1983).

In the context of policy-evaluations evaluators may find themselves in the midst of a totally politicized context, where partisans will try to use the evaluation to enforce their own position. In such situations a lot will depend on the professional standing of the evaluations, and the organizational independence of the way they can operate. In (internal) school evaluations the situation may not be so openly politicized, though nevertheless, important differences of goals and priorities in a certain domain of schooling may occur as well. And, with respect to external school evaluation, there may

also be differences about the key-function of the evaluator: to inform higher administrative levels, to enhance "consumerism" or to inform the school personnel about strong and weak points of school functioning.

The practical implication of all this for evaluators is that the initial stage of setting the priorities and choosing evaluation criteria and standards is particularly important. Activities should not just be seen as analytic but also, maybe even more so, as practical and "managerial", in attempting to come to terms on evaluation priorities with stakeholders, surmounting resistance and building commitment.

The second aspect of the rationality model, the existence of a somewhat explicit rationale about the "means" of the program being adequate to reach the goals, has given rise to a particular type of evaluation, namely "analytic evaluation". Analytic evaluation is meant to articulate the basic means-to-end structure of a program. Sometimes this process is described as "making explicit the *program theory*", or reconstructing the "underlying program logic" (Leeuw, Van Gils & Kreft, 1999).

Contrary to the rational ideal, in actual practice, the link between goals and means can be considerably loose. Means, for example, can be chosen because they are really ends in themselves—"the medium is the message", because they serve the more particularistic objectives of some stakeholders, or just because the program was not well prepared or designed. In such situations evaluators can, in principle, save a lot of time, efforts and ultimately money, by pointing out such weaknesses. In "analytic evaluation" the evaluator uses analytic methods (like review of existing research findings on the issue) to get an indication on how likely the proposed methods will lead to goal attainment.

According to the third characteristic of the rationality model, planned programs are also "really" implemented. Again, practice shows that this is not always the case and partial or even no implementation at all may be the reality. In case program objectives are assessed and program implementation has failed, the evaluation is called a "non event evaluation". In order to prevent this, it is preferable to include an implementation check in the evaluation design. This can be done by means of measuring process indicators or by means of direct observations.

If the program, particularly the means, methods and approaches comprising the program are less straightforward, and implementation is more characterized as a process of "mutual adaptation" between ideas and individual preferences of practitioners, checking implementation becomes more complex. In such situations observational studies may work more like constructing "post hoc" program variants, which then may be associated with outcome measures later on.

The fifth rationality assumption about the evaluation setting is that evaluators will be in a position to carry out their professional job in relatively undisturbed way.

As we already saw when discussing the situation at the outset of program evaluations, where evaluators may become emerged in political debates among partisans, this condition too, is not always met in actual practice. But also when it comes to choosing evaluation methods and carrying out data collection this condition may be violated. When the stakes of the evaluation are, rightly or wrongly, considered high by practitioners, they will not have a neutral stance with respect to the data collection methods that are proposed by the evaluator. One could look upon evaluation methods as varying on a continuum running from "evaluator control" to "practitioner control". Participatory observation, methods like a teaching writing a "log" of each lesson and open interviews

are examples of methods which are very much under the control of the respondents. Standardized tests and external observations are largely outside the control of the practitioner. Scheerens (1983) describes a setting where evaluation apprehension of practitioners led them to renounce objective, evaluation-controlled measures, and plead for more open methods in which they themselves were the main providers of information.

In many cases a clear exploration of the purposes of the evaluation will help to overcome resistances. For example, in program evaluations and school evaluations organizational functioning rather than individual functioning of teachers is the evaluation objective. Nevertheless teachers may still think that they are the evaluation object, and they would need to be told explicitly that this is not the case.

Although most of the features discussed in this section are more prominent and are documented within the realm of program evaluation, the recommendations that were provided are also relevant for other kinds of educational evaluation, like school evaluation. The major recommendations are:

- to analyze program objectives carefully and enter a process of illumination of objectives among stakeholders, preferably resulting in overt commitments to goal statements and effect criteria that will be ultimately used in the evaluation;
- to critically assess the attainability of means-end relationships, in other words the likelihood that proposed program means will lead to goal attainment, preferably before empirical evaluation activities start;
- to empirically check the implementation of the program or set of activities that is to be evaluated;
- to be prepared for politically inspired negotiations about more or less reactive data collection methods and also for investing time and energy in communication and presentation of the intended evaluation activities and their objectives;
- to invest in communicating the evaluation results as a general process of illuminating issues to stakeholders, which may or may not lead to immediate impact on decision-making.

4.3 Gearing Evaluation Approach to Contextual Conditions; the Case of Educational Reform Programs

In this section the emphasis is on the procedural dimension of educational reform programs and the decision-making context. Procedural reform strategy depends on the sequence of phases and on the locus of decision-making concerning the reform program, distinguishing a top-down versus a bottom up approach in the management of the reform. When it comes to a proper gearing of monitoring and evaluation to decisions regarding the project in its various phases two aspects of this decisionmaking context are important: the overall clarity and rationality of this context (see the previous section) and the division of decision-making authority over hierarchical levels of the education system.

4.3.1 Phase models

In their paper on "Monitoring and evaluation in World Bank education operations" Scheerens, Tan and Shaw (1999) provide a framework that depends on a phase model of projects.

A "Bank-financed project" is described as the provision of funding for a set of activities or inputs to produce outputs that are expected eventually to yield some desired educational, social or economic outcomes—such as broader access to schooling, greater survival rates, better student learning, expanded employment, higher earnings and so on. Figure 4.1 provides a schematic model of the progression of events.

> "In the figure each phase of project development is linked to the next by various processes and all the phases take place in an overall social, economic and policy context. The context is important not only because it establishes a baseline against which progress can be assessed, but also because it affects the overall socioeconomic, administrative, management and incentive structures within which the project events unfold. The processes are similarly relevant, particularly those relating to the management of processes at the level of schools and classrooms."

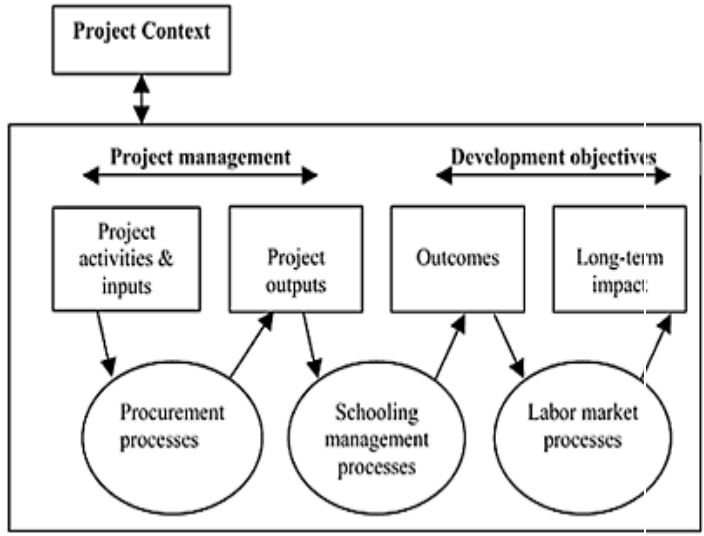


Figure 4.1 Schematic model of the progression of project events.

"Each of the many distinct phases and processes that occur between the provision of Bank funding for a project and the achievement of its intended development objectives can be the focus of monitoring and evaluation work. The diversity creates substantial room for confusion in discussions about such work in the context of Bank-financed operations.

> In particular, because people may focus on different parts of the schematic model, it is conceivable for them to mean and expect quite different things when they talk about M&E activities. For example, people interested in how project funds are used may be interested mainly in monitoring the procurement processes, whereas educators are much more concerned about monitoring and evaluating the impact of the project on educational outcomes. Even among educators, some may place greater emphasis on assessment of the schooling processes than on the tangible inputs that create the environment for learning. Apart from the differences in focus, other sources of misunderstanding and cross-communication relate to such issues as: the choice of indicators, the level of aggregation in the information, the timing and periodicity of reporting, the audience for the information, the allocation of responsibility for M&E activities and reporting, and so on." (ibid).

Figure 4.1 depicts a sequence of project activities. Monitoring emphasis differs according to each of the subsequent phases:

Phase 0. In a "zero" phase (not included in Fig. 4.1) where the overall project rationale and its feasibility are analyzed, so called "risk" indicators (see World Bank, 1996) answer the question *whether the project design is "sound and feasible"*.

Phase 1. During the procurement process (see Fig. 4.1) the core question is *how well project funds are used for the activities that should take place.*

Phase 2. *The next phase is the stage of actual project implementation. Here the immediate "outputs "or project activities are the object to be monitored. Examples are:* Have intended teacher training courses actually taken place, and were enrolments up to standard? Were new curricula and textbooks actually produced in time and are they used by the teachers? Have the intended number of new school buildings actually been built? Have student enrolments increased as planned?

Phase 3. The next question in the sequence is *whether the direct intended outcomes of the project have been attained*. Relevant outcome indicators are: success rates in examinations, drop-outs, subject matter mastery at the end of a period of schooling as measured by means of standardized achievement test.

Phase 4. In this phase longer-term outcomes and project impact are the objects to be monitored. Some examples are:

- position and success rates of students in *tertiary* education as an impact measure of a project in *secondary* education;
- the labor market position of graduates;
- the impact of school improvement projects on indicators of the functioning of local communities;
- improved enrolments in secondary education by pupils from poor rural areas.

The phase model is also used to explain a difference in emphasis between monitoring on the one hand and evaluation on the other. When the different foci which depend on the sequence of phases in project implementation are assessed in a predominantly descriptive way a *monitoring* orientation predominates. In case the ambition goes further in the sense of explaining and causal attribution the orientation becomes more comparable to the logic

of *program evaluation*. In that case it would be necessary to "drop down" to the various processes depicted within circles in Figure 4.1. This would mean additional measurement and description (of the various processes) as well as attributing outputs, outcomes and impacts to these process arrangements

From the above exposition it is clear that World Bank education projects are described according to the principles of planned change in social systems. The education innovation literature, starting with the "Rand Studies" (cf. McLaughlin, 1990) added a dimension to the sequences of events which reflects a gradual adaptation and implementation of the reform by practitioners. According to the underlying view project implementation is not simply a matter of the "fidelity" in following the schemes and plans of external change agents but a process of "mutual adaptation" between planners and practitioners and between newly presented material and interpretation by professionally autonomous professionals. Usually the following sequence of stages is distinguished: initiation, adoption, implementation and institutionalization of the reform.

The most important implication for the design of monitoring and evaluation activities is the realization that implementation cannot be taken for granted and should be an object of monitoring in its own right, for example, by checking whether intended innovative procedures are actually brought into practice. In fact the various assumed stages in the gradual adoption and implementation of reforms by the various relevant actors could be studied and monitored as an evaluation object in its own right.

Although national educational reform projects will tend to have a distinct set of centrally arranged inputs, the dimension of "top-down" versus "bottom up" development of projects is nevertheless relevant. This is the case because there may be differences in degree as to which parts of the reform are expected to be initiated and managed from lower levels in the system.

4.3.2 Articulation of the decision-making context

Monitoring and evaluation belong to the category of rational techniques of policyanalysis. In all applications, even if they only reflect a partial aspect of the rationality paradigm (see previous section), there is the assumption that there is some kind of designated decision-making structure and that the monitoring and evaluation results are actually used for decision-making.

In the earlier cited paper by Scheerens, Tan and Shaw the following tentative overview of audiences for monitoring and evaluation within the context of World Bank funded education projects is given.

Figure 4.2 gives a schematic overview.

> "Two final observations with respect to the audiences of M&E activities are in place. Firstly, there is a distinction between control vs learning "modes" in the use of evaluative information. Generally control-functions are served on the basis of the various monitoring activities, categorized according to the phase of project preparation and implementation. Learning functions are served by evaluation and "effectiveness" studies as summarized in the lower part of Figure 4.2. Because input/process/outcome relationships are central in these latter types of

evaluation, they will generally provide more specific diagnoses as well as handles for improvement. Secondly, the recognition of "outside" audiences, points at the possibility to institutionalize evaluation procedures after projecttermination. This could be seen as a very important spin-off of World Bank education M&E activities, because it could lead to a sustained strengthening of the educational evaluation function in recipient countries" (cited from Scheerens, Tan & Shaw, 1999).

Type of Monitoring	**Bank staff**	**Recipient country**
phase 0 (risk indicators)	regional officers country directors task managers	Ministry of Education (MOE)
phase 1 (accounting for funding)	Procurement officers (??)	MOE
phase 2 (implementation/output indicators)	country directors task managers	MOE
phase 3 (outcome indicators)	regional officers country directors	MOE
phase 4 (impact indicators)	senior management regional directors country directors	MOE and other Ministries
Type of Evaluation		
Effectiveness of procurement processes Effectiveness of schooling and project management processes	Procurement officers task managers staff departments research departments	MOE MOE, regional officers school leaders & teachers, parents MOE, other Ministries
Effectiveness of school-labor market transmission processes	country directors	

Figure 4.2 Audiences for World Bank education M&E activities.

Leaving aside the specific context of World Bank projects a more general treatment of educational monitoring and evaluation and types of decisions can be made. By listing decisional contexts that follow the levels of education systems from high (central) to low (classroom and student level) a stylized overview of types of decisions with corresponding actors and other stakeholders can be given. This is attempted in Table 4.1, below.

As announced, the table provides a stylized overview. The implied assumption that evaluative information is used by decision-makers in a straightforward and linear way is challenged by the results from empirical studies that have investigated actual use, mostly in the context of program evaluation. As stated in the first section of this chapter, according to the rational ideal, evaluation would have a natural place in a clear-cut decision-making setting, where scientific methods and scientific knowledge are used to guide political decision making. As authors in the realm of studies about the use of

evaluation research have shown (e.g. Caplan, 1982; Weiss, 1982; Weiss & Bucuvalas, 1980) the assumptions of evaluation within a rational decision-making context concerning the decision-makers being clearly identified, and evaluation results being used in a direct linear way, may *not* be fulfilled.

Table 4.1 Major Decision Areas, Decision-makers and Other Stakeholders in Use of Information From Educational M&E.

Decisions	**Decision-makers**	**Other stakeholders**
(In case of evaluation of the success of an educational reform program) adaptation of program implementation, determining continuation or termination, conditions for sustainability of program	• Donor agencies	Elected officials and top level education officers (MOE) in borrowing country
Reconsideration of national educational policy agendas	elected officials and top level education officers deciding on financial inputs, and revision of the national policy agenda using system level indicators on inputs and outcomes, possibly using international benchmarking information as well	Taxpayers Social partners (industry, unions) Educational organizations
Reform of national curricula	• Same as above	As above. Subject-matter specialists Assessment specialists Educational publishers
Restructuring of the system in terms of functional decentralization	• Same as above, also including information of educational structures in other countries	Administrators at all levels of the system Social Partners Educational organizations
Reconsideration of regional and local educational policy-agendas	Municipal or school district level educational authorities deciding on levels of school finance, resources, and substantive educational priorities (again depending on their discretion in these domains) of the schools in the community, using information from school level context, input, process and output indicators	Local community Local/regional pressure groups Local industry Teacher unions School representatives Educational organizations

School development planning and school improvement activities	school managers and teachers using school-based information on inputs, processes and outputs, compared to regional or national averages to monitor or adapt overall school policy, the school curriculum, school organizational conditions and teaching strategies	Parent association Student representatives Local community
Choice of teaching strategies and individualized learning routes for students	teachers use detailed information from student level monitoring systems to monitor or adapt their teaching and pedagogy with respect to groups and to individual students	Parents Students
School choice	• parents, students	Schools, local authorities

In actual practice several or all of these assumptions may be violated. As goals may be vague, or contested among stakeholders, the assumption of "one" authoritative decision maker, either as a person or a well-described body, is also doubtful, even in the case where the decision makers are governmental planning officers.

As far as the use of evaluation results is concerned, empirical research has shown that linear, direct use is more the exception than the rule, "research impacts in ripples, not in waves" says Patton (1980) in this respect. Authors like Caplan (1982) and Weiss proposed an alternative model of evaluation use, which they consider more realistic. According to this view evaluation outcomes gradually shape perspectives and conceptual schemata of decision-makers and has more an "illuminative" or "enlightenment" function than an authoritative one.

The decision-making context is likely to be less rational (see the first section of this chapter). Instead it may be diffuse while political aspects may have an impact on the use of evaluations *and* on the very conditions in which it can be applied as the impartial, objective devise it was meant to be.

Although the above considerations most directly apply to program evaluations, they are also likely to play a role, when evaluation has the nature of regular, periodic assessment of the running of a complete educational sub-sector, as with national assessment programs, educational indicator systems, or periodic evaluations carried out by inspectorates. For example when assessment programs are conducted in a setting where the stakes are high, e.g. by making school-finance contingent on performance levels laid bare by the assessments, strategic behaviour is not unlikely to occur. Examples are: training students in doing test-items, adapting classrepetition policies so that a more select group of students actually goes in for testing, leaving out the results of students that score less well.

As Huberman (1987) has shown, the degree to which evaluation can play its rational role is dependent on various structural arrangements:

- the institutionalisation of the evaluation function, e.g. whether there is an inspectorate with a distinct evaluation function, whether there are specialised research institutes;
- the scientific training and enculturation of the users of evaluation results;
- the degree to which evaluators are "utilisation focused" and actively try to act on the political realities in the evaluative setting.

The research literature in question has also yielded a set of conditions that matter for the use of evaluative information. The following conditions, all of which are amenable to improvement, have been mentioned:

- perceived research technical quality;
- commitment of audiences and users to the aims and approach of the evaluation;
- perceived credibility of evaluators;
- communicative aspects (like clarity on aims and scope; brevity of reports);
- political conditions (e.g. evaluations that bring "bad news" on politically sensitive issues are in danger of being ignored or distorted).

Conditions that impinge on the actual use of evaluative information relate to an area that is of enormous importance to the question to what extend the increased range of technical options in educational assessment, monitoring and evaluation will actually make true its potential of improving the overall performance of systems. This is the domain of the pre-conditions that should be fulfilled for an optimal implementation and use of M&E. It is dealt with in more detail in a subsequent section.

4.3.3 Monitoring and evaluation in functionally decentralized education systems

When discussing the use of M&E results for different type of decisions made at different levels of the education system (see Table 4.1) a more systematic treatment becomes possible when a concise conceptual framework is employed which specifies levels and domains of decision-making in education. We therefore turn to the conceptualization and measurement of functional decentralization of education systems. The exposition depends largely on work on "locus of decision-making" carried out within the context of the OECD-INES project, by one of the Networks of this project (Network C).

The OECD-INES procedure to measure "locus of decision making" distinguishes *three* facets of the rather crude distinction between centralisation and decentralisation:

- the tier or administrative level where a decision is taken; this dimension was referred to as the *locus of decision-making;*
- the amount of discretion, or the degree of autonomy of decision-making at a particular administrative level; this facet was called the *mode of decision-making;*
- the particular element of educational administration a decision belonged to; this facet was referred to as the *domain of decision-making*.

These three facets can be related to existing categorisations in the relevant literature, although the use of central concepts is by no means consistent among authors and publications. Our three-dimensional conceptualisation is compared to the terminology as clarified by Bray (1994, p. 819) in an analysis of alternative meanings of centralisation and decentralisation.

The distinction between levels confirms to the concept of *territorial decentralisation,* defined as "the distribution of powers between different tiers of government".

Degrees of autonomy in decision making at a particular level are reflected in terms that refer to an increase in discretion. Again following Bray, *deconcentration, delegation*

and *devolution* are modes of decision making in which an increased amount of decision-making authority resides at a lower level.

"*Deconcentration* is the process through which a central authority establishes field units, staffing them with its own officers".

"*Delegation* implies a stronger degree of decision making at the lower level. However, powers in a delegated system still basically rest with the central authority, which has chosen to "lend" them to a local one".

"*Devolution* is the most extreme form of decentralization. Powers are formally held by local bodies, which do not need to seek approval for their actions" (ibid, p. 819).

In the operationalization of this continuum of increasing autonomy, these abstract definitions were avoided and respondents were asked to indicate whether decisions could be taken within the framework determined by a higher level, in consultation with a higher level or in full autonomy.

In order to determine elements or *domains* of educational administration, many categorization schemes are available in the literature (e.g. Bacharach et al., 1990; James, 1994; Rideout & Ural, 1993; Winkler, 1989). The common core of these categorizations are three main areas:

a. an educational domain (goals, methods, curricula, evaluation procedures);
b. an organizational, managerial and administrative domain (including human resource management, groupings and assignment and foundational regulations);
c. a dimension concerning finance and the way financial resources are applied.

In the operational classification that was chosen four main categories were used, by splitting up area b (organisational) into two domains "planning structures" and "human resources", and including areas a and c.

The distinction between domains of decision-making in educational systems bears some resemblance to Bray's use of the term "functional decentralisation" as cited from Rondinelli. "Functional decentralisation refers to the dispersal of control over particular activities" (Bray, 1994, p. 819). The main issue is the recognition that educational systems maybe centralised in some domains of decision-making but not in others.

To learn more about educational decision-making in OECD countries and to systematically compare decision-making processes across countries, an instrument was developed that examined the locus of decision-making in four important domains. As stated above, these domains were: (1) the organization of instruction; (2) personnel management; (3) planning and structures; and (4) resource allocation and use. Within each of these four domains, between seven and 15 decisions were examined. In the domain entitled, "organization of instruction," for example, the instrument focused on decisions about such matters as textbook selection, grouping of pupils for instruction, and assessment of pupils' regular work. In "personnel management," questions were asked about hiring and dismissal of teachers and other school staff, duties and conditions of service, and the setting of salary schedules. In "planning and structures," the focus was on creation and abolition of schools and grade levels, the design and selection of programs of study, course content, and policies regarding credentials. Finally, in the area of "resource allocation and use," the instrument focused on decisions about the allocation of resources for staff and materials, and the use of financial resources for these purposes.

Each of the questions in the instrument was designed to identify the level at which decisions are made in the governmental system (the "level" of decision making) and the way decisions are made (the "mode" of decision making). Six "levels" of decision-making were set out in the instrument. These include the following: (1) central government; (2) state governments; (3) provincial/regional authorities or governments; (4) sub-regional or inter-municipal authorities or governments; (5) local authorities or governments; and (6) schools. Three "modes" of decision-making were examined in the instrument. Decision could be made by an authority (1) autonomously, (2) within a framework established by another level within the system, or (3) in consultation with other levels in the system. Based on the instrument, it was possible to determine how centralized or decentralized decision was overall, in each of the four domains, and for individual education decisions.

Finally, it should be noted that the instrument included questions about decision making at three different education levels: primary education, lower secondary education, and upper secondary education. Within upper secondary education, questions were asked separately for general education and vocational education.

The decision-making survey was administered in the spring of 1998 to panels of national experts. For each level of education, countries assembled two 3-person panels with representatives from each of the following government levels: (1) highest level (central government); (2) middle levels (state governments, provincial/regional authorities or governments, sub-regional or inter-municipal authorities or governments, local governments); and (3) schools.

The two panels constituted for each education level went through the instrument question by question and attempted to arrive at consensus on the "level" and "mode" of decision-making on each question. The responses of each panel were then reviewed by each country's representative in INES Network C to determine whether there was consistency in the panels' responses to the each question. In cases where the responses differed, the Network C representative used source documents and consultation with the National Coordinator of the INES Project to reconcile these differences. Following the administration of the questionnaires by each country, completed instruments were sent to the survey coordinator, who entered countries' responses into a database. These responses were then used to calculate the indicators on decision-making, which were published in the 1998 edition of Education at a Glance. The data-set is sufficiently rich to calculate additional indicators. Examining locus of decision-making with respect to domains and sub-domains is one of the most interesting possibilities. Illustrative findings are presented in Figures 4.3 and 4.4.

In Figure 4.3 countries are compared with respect to the percentage of decisions taken at the school level. This percentage can be seen as a rough measure of school autonomy: the higher the percentage of decisions taken by the school the more an educational system is decentralized. A somewhat "stricter" definition of school autonomy is used in Figure x, where the percentage of decisions taken by schools in *full autonomy* are compared across countries.

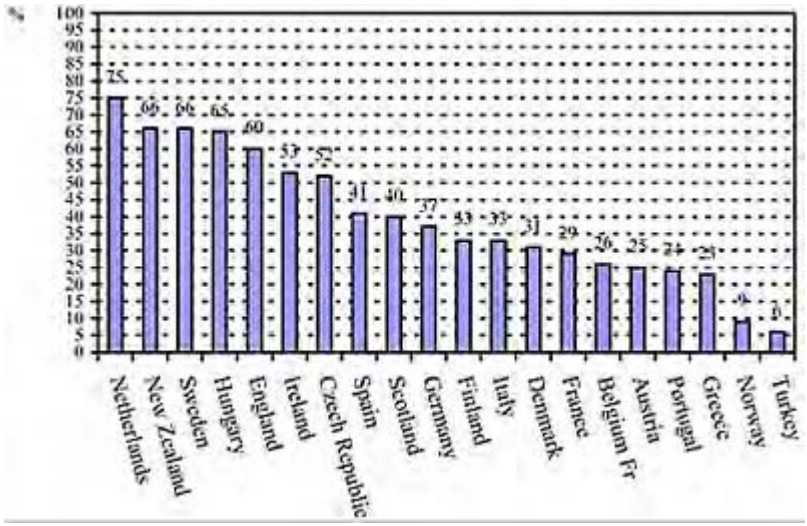


Figure 4.3 Percentage of decisions taken by schools across countries at lower secondary level (Source: OECD Network C).

With the Czech Republic as the main exception, the grouping of countries in Figures 4.3 and 4.4 is rather similar, which means that generally in those countries where the school takes relatively many decisions schools also take the largest percentage of decisions autonomously. In the case of the Czech Republic 37% of the 52% of decisions taken by the school are taken within a framework set by a higher level and 12% is taken after consulting other levels. The group of OECD countries where schools have considerable impact on decisionmaking composes of the UK, Sweden, the Netherlands, Hungary, New Zealand and Ireland. At the other extreme we find countries like Turkey, Norway and Portugal.

There are three reasons to look somewhat more closely at school evaluation, when considering the possible impact of decentralization policies on the improvement of school functioning:

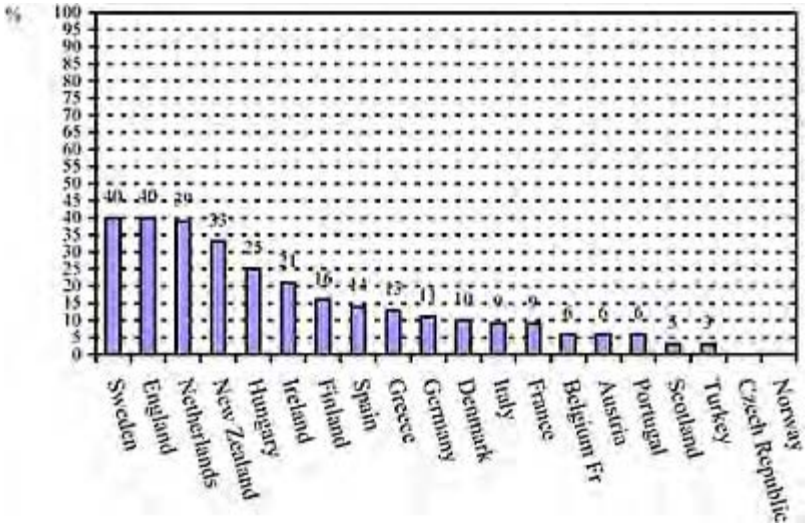


Figure 4.4 Percentage of autonomous decisions taken by schools across countries.

a. evaluation and assessment of pupil functioning can be seen as a functional domain that can be decentralized in its own right (evaluation as part of decentralization policies);
b. evaluation and assessment can be kept centralized as an overt policy to monitor output while freeing processes (evaluation as a counterbalance to decentralization in other domains of schooling);
c. evaluation and assessment together with feedback and use of evaluative information can be considered as a potentially effectiveness enhancing mechanism, and thus as an important lever of school improvement.

In the OECD locus of decision-making questionnaire there are three items which refer to pupil assessment. In all countries the school is responsible for the assessment of pupils' regular work, and in the majority of these the school is solely responsible for this task.

Setting examinations (i.e. determining the contents of examinations) and "credentialling" (making practical arrangements for examinations, i.e. marking procedures) is mostly the responsibility of the central level.

Analyzing patterns of functional decentralization of education systems that take part in reform programs is useful to obtain a clear picture of the responsibilities for program administration and implementation of the various administrative levels. Of course if decentralization is a substantive part of the reform program, assessing its manifestation at various phases of the program's development would be a logical part of the overall evaluation.

Patterns of functional decentralization also provide handles to determine the direction of feedback loops, when it comes to the utilization of evaluation results. Both in the context of reform program operation and in the context of more permanent monitoring of the system the following rule of thumb could be proposed. *Information that is yielded by*

monitoring and evaluation should be fed back primarily to those organizational levels and administrative bodies that have the discretion to rule on the information. An implication of this rule would be that, in a situation where schools are autonomously responsible for classroom instructional processes, evaluative information on these processes should only be fed back to the schools' management and staff.

4.4 Creating Pre-Conditions for M&E

In this part of the chapter political, institutional, organizational and technical preconditions for proper application of M&E in education are defined and clarified. Analyzing these pre-conditions is relevant in targeting areas for organizational development and capacity building.

Of course the application of monitoring and evaluation requires technical capacity. But apart from technical capacity even more basic pre-conditions that are relevant to the actual chance of success of M&E initiatives should be considered as well, these are: the "political will" and commitment to the aims of M&E and the *institutional and organizational capacity* for educational evaluation. To the extent that pre-conditions in these areas are not fulfilled the implementation of M&E is constrained or may not get off the ground at all. Dealing with these constraints and pre-conditions involves assessing them at the outset of developing systemic M&E, improving them where possible and (to the degree that this is not possible) adapting the M&E ambitions.

Evaluation occurs on the borderline between systematic inquiry, guided by the principles of scientific method, and the practical world of policy-making, governance and management. The common ground between these two worlds depends on certain theoretical assumptions about principles of rationality (see section 4.2).

4.4.1 Political will and resistance

Since evaluations will ultimately lead to value judgements they make for a politically sensitive endeavor. Political involvement could mean various things, ranging from a rational and technological assessment of costs and benefits, matching the M&E agenda to the issues that are of central importance from a particular ideological point of view to avoiding and blocking M&E when it touches on politically sensitive areas.

Political sensitivity goes beyond the level of national policy, however, and occurs at all levels of project implementation where vested interests are at stake. Even if this is not "objectively" the case people may feel threatened by evaluations. Education in many societies has been a relatively closed system, with little outside interference on what goes on in schools and classrooms. Teachers are prototypes of autonomous professionals and for schools as organizations a specific organization model has been invented: the professional bureaucracy. One of the characteristics of this model is that the autonomous professionals tend to resist close supervision and systematic review of their work.

Aspects of political commitment and resistance to M&E that should be considered are:

- the degree to which the political top of the education system supports particular M&E activities and provisions;

- perceived threats at the political top or the top of the government bureaucracy for the possible outcomes of new or improved M&E provisions and activities;
- the degree to which M&E becomes part of a political controversy between ruling party and opposition;
- the stability of the political top during the period when the M&E activity has to be realized;
- a certain antagonism in supporting M&E project managers, based on perceived risks of creating a "critical conscience" in the education system;
- how teachers and school directors perceive of the "stakes" involved in M&E;
- how "user friendly" or "alien" the methods of M&E are perceived by school staff;
- in a more general sense the incentives and "disincentives" of actors like local administrators, school managers, teachers and students to participate in or be subjected to M&E activities (at the cost side investment of time, loss of status, fear of being criticized, fear of weakening or loss of position should be considered);
- the position of stakeholders in a particular M&E activity in terms of commitment, resistance and "political" preferences for certain methodological approaches (for example when teachers only tolerate qualitative, participatory methods, in other words methods in which their way of seeing things is clearly represented, and resist more external and objective methods of data collection, this could be more than a methodological preference, and be a sign of resistance to critical review).

Since educational evaluation depends very much on the cooperation of people in the situation that is object of evaluation it is particularly vulnerable to distortions and manipulations at this level as well.

Strategies that should be considered for improving political commitment are:

- persuasion, by clearly stating the objectives and clarifying the methods of M&E, also guaranteeing safeguards against possible harmful side-effects, guaranteeing anonymity etc.;
- providing incentives for participation in M&E activities, preferably stimulating intrinsic motivation by exploiting spin-off of M&E to the benefit of actors (for example by making a special effort in feeding back information to schools);
- coercion and close supervision by the government, by making other aspects of reform (like provision of better equipment, teacher training schemes and new curricula) contingent on cooperation with M&E activities.

4.4.2 Institutional capability for M&E

Institutions are "the rules of the game" in a society. They should be distinguished from organizations, which structure "the way the game is played" (Berryman et al., 1997). Examples of institutions are the legal system, property rights, weights and measures and marriage. But the rules of the game may also be less formal and depend on convention and implicit norms.

In assessing the institutional capability for M&E in a country instances of an *"evaluation culture and tradition"* should be looked for.

Indicators of the degree to which a country has a strong or weak evaluation culture that could be considered are:

- whether or not quality and safeguarding quality in education is mentioned in the constitution or other legislation;
- the elaborateness of the system of examinations and certification in education;
- evaluation history (e.g. for how long have educational programs been empirically evaluated, and with what degree of success);
- instances of real use (e.g. evident from referring to evaluation results in public media) vs symbolic use of evaluations;
- participation in international assessment surveys;
- emphasis on accountability in education in public and political debate.

Institutional capability for M&E is most realistically addressed as an assessment activity in order to obtain a notion of the general climate in which M&E activities in a country will "land". Institutional development in this domain is an endeavor that appears to go beyond the conduct of a particular reform project and should be embedded in a more general and long term country strategy.

4.4.3 Organizational and technical capacity for M&E

Questions about organizational capacity for M&E in a country first of all regard the issue of whether important technological functions have an "organizational home" in the country. For example, initiating a national assessment is the more of a heavy task when there exists no organization that has specialized in the development of educational achievement tests in the country. The same applies when external supervision of schools is considered at a fairly large scale and the country has no educational inspectorate.

Further criteria in determining the organizational capacity concern the wellfunctioning of organizations in terms of effective leadership, ability to mobilize financial, material and human resources and appropriate work practices (Orbach, 1998).

For organizations concerned with educational M&E additional criteria for wellfunctioning are professional standing and a credible degree of impartiality.

Organizational capacity building for M&E should start from a careful analysis of the planned M&E approach and technology and the mix of skills and expertise needed to carry it out successfully. Next, available organizational "homes" should be examined for gaps between the required and available skills and general organizational well-functioning. If no such "homes" are available the creation of new units should be considered. In case of gaps several options are to be considered: narrowing down of the M&E objectives or changing and improving current practices, e.g. by means of training, and provision of additional resources, human resources (e.g. external consultants) in particular.

As far as *technical capacity* is concerned the required set of skills for successfully carrying out M&E activities depends on the priorities and ambitions of the M&E plan, in the sense of the M&E objectives, general approach and specific methods. Although these are likely to be given most of the attention it should be noted that the required skills do not just pertain to research methodological and technological skills but also to communicative skills and substantive educational knowledge.

Issues of organizational and technical capacity for M&E were documented more specifically in Chapter 3, in referring to the technical and organizational requirements

needed for each of the 15 specific types of M&E that were distinguished in the first chapter).

4.5 Conclusion: Matching Evaluation Approach to Characteristics of the Reform Program, Creating Pre-Conditions and Choosing an Overall Strategy for Systemic M&E

In this final section, first of all, some conjectures are made about fitting m&e strategies to the various contextual aspects discussed throughout the chapter. The idea of "fitting arrangements" reflects a contingency approach, which comes down to the assumption that there is no ideal evaluation strategy that would be optimal for all contexts. At the end of the section the contingency view is challenged in considering whether certain elements of m&e strategies would deserve universal implementation.

Overall "rationality" in the decision-making context

To the degree that the decision-making context surrounding the reform program is fuzzy, i.e. when reform departs from the planned change model to incremental reform, monitoring, inspection and school self-evaluation rather than program evaluation become the most appropriate strategies. If the reform program itself is a loose collection of initiatives and interpretations from individual schools, perhaps as a consequence of a bottom up innovation process, "strong" experimental or quasiexperimental research designs may be difficult to implement and more "ad hoc" designs could be the only option.

Patterns of functional decentralization

The client orientation of M&E should depend on the decision-making authority of actors in the education system. A tentative principle that was proposed states that evaluative information is only made available to actors that have the discretion to act on the basis of this information.

The phase of the reform program

In the early (initiation/adoption) phases of the program pre-evaluative studies describing practitioners' reactions could be considered. In these phases evaluation feasibility studies could be useful as well.

At the implementation phase, process evaluation in the sense of treatment implementation checks, are of great importance, particularly when programs have a broad scope, wide coverage and degrees of leeway in the interpretation of program contents.

At the phase of program institutionalization program outputs should be monitored.

Contingency or uniformly recommendable strategies?

The above conjectures all express a contingency approach: the appropriateness or the efficiency in the choice of monitoring and evaluation strategy depends on characteristics of the reform context. To challenge this perspective one could raise the question whether some evaluation approaches might be preferable in all possible contexts of educational reform. The obvious candidate for this qualification would be output assessment.

The fields of educational evaluation in the sense of measuring student performance on the one hand and educational evaluation in the sense of program evaluation on the other hand have developed as two relatively separate fields. It is quite clear that integration is required and that program evaluations in education should use assessment of student performance, in basic subjects and key competencies, as central criteria. In periodic monitoring of ongoing functioning of the system or long-term education reform student performance assessment is also likely to be chosen as the first priority. In cases where national student assessment programs are not available and international assessment, for one reason or other, are not used, other types of outcomes, like pass/fail rates, participation in further education and drop-out rates, would be the only option, still maintaining the prevalence of outcome indicators.

In all reform programs, where some kind of curriculum revision is at stake, it would also be likely to use student achievement assessments in the particular curricular domain as effect criteria.

In school self-evaluation pupil-monitoring systems could also play a central role in the future, combining a data-driven and result oriented school management strategy with monitoring and diagnosis at classroom level.

So, in conclusion, it seems that one comes a long way in ascertaining the usefulness of student performance assessment as the backbone of most contexts of educational reform and evaluation. Only in early, formative phases of program development are other, more process-descriptive, approaches sufficient. More fully fledged monitoring and evaluation approaches would seek to causally relate input/process information to assessment results. Sometimes a case can be made for using “process” indicators as “substitute” outcome indicators (Oakes, 1989; Scheerens, 1990).

Given the “ingredients”, including the analysis of the various kind of “preconditions” that were presented in this chapter building a general strategy for determining the priorities for systemic M&E in a country could subsequently address the following issues:

- describe the available provisions for M&E in the country with respect to the three basic functions that were discussed (certification/accreditation, accountability, self-improvement of units); conclusions in terms of the degree of development of provisions serving each of these functions;
- analyze patterns of functional decentralization in the country; conclusions about M&E provisions that “fit” with a certain pattern (for example, as illustrated with the case of the North African country, M&E types oriented towards selfimprovement of units, like school self-evaluation, only make sense if schools have a certain degree of discretion to do something with the results; asking schools to report about their performance and services to parents is more relevant in a context where there is free choice of schools etc.);

- target M&E provisions to substantive priorities in educational policy; for example when the priority for the next five years is on the improvement of lower secondary education, this could be the first sub-sector to optimize M&E provisions;
- check political, institutional and organizational pre-conditions to determine the degree of effort to realize the preliminary set of priorities.

These steps would result in a set of general priorities, which should be further specified on the basis of a more detailed analysis of technical options and possibilities for synergy between them.

5
Evaluation as a Tool for Planning and Management at School Level

5.1 Introduction

Evaluation and monitoring can be examined from a purely technical and methodological perspective. In this context, however, the emphasis is placed on the applied nature of M&E. This means that the integral function of M&E within the context of educational planning and management is at the center of attention. Speaking of an "integral function" of M&E is supported, first of all, by the realization that the evaluation and feedback function has a key role in meta-theories in the field of planning and management that all depend on the rationality paradigm. In this section a brief excursion will be made to explain the role of the evaluation function in each of these planning and management theories. After a reconsideration of more prescriptive applications of the rationality principles already introduced, three more specific interpretations will be discussed: synoptic rational planning, creating market mechanisms (or *choice*) and cybernetics. This latter mechanism is further elucidated by considering what is indicated as "retroactive planning" and placed in the context of organizational learning and the image of the learning organization. A critical look is taken at the validity and applicability of these theoretical constructs to the reality of schooling, by making a comparison with more traditional perspectives (of the school as a professional bureaucracy) and results from empirical school effectiveness research. The conclusion is that although schools do not confirm to the core principles of "learning organizations" in all and every way it is still a heuristically relevant image for the field of school improvement. Its key-mechanism of evaluation and feedback is considered of central importance to improving the responsiveness and instrumental effectiveness of schools.

5.2 The Rationality Paradigm Reconsidered

In the previous chapter the rationality paradigm was used as a point of departure for descriptively characterizing typical settings in which educational M&E takes place. The major conclusion being that settings usually differ more or less dramatically from the rational model and that this poses problems for the application of M&E.

In this chapter the same paradigm is considered in a more prescriptive way, namely as an ideal type of "good" policy-making and effective management. The role that monitoring and evaluation is supposed to play in such forms of "good policy-making" and effective management being the central topic. In considering the rationality paradigm from this perspective we are once again confronted with limitations. Considering these limitations has led to modified (still prescriptive) models of the "pure" rationality model.

The role of monitoring and evaluation in these modified versions of the pure rationality model, will be considered as well.

The rationality paradigm lies at the basis of theories on planning and public policy making, micro-economic theory, "organizational learning" theory and even contingency theory.

The basic principles of the rationality paradigm are:

a. behavior is oriented toward preferred end states (as reflected in goals or individual well-being);
b. in situations where there is a choice between alternative ways to attain the preferred end states, an optimal choice is made between these alternatives, which means that profit, well-being, or other preferred end states are maximized given the alternatives and constraints;
c. in organizational settings the alignment of individual preferences and organizational goals is a major issue.

The rationality paradigm is applied in formal and less formal models of planning, control, design and feedback and is attached to different units: organizations as a whole, sub-groups or departments and individuals. Apart from this, procedural vs. structural interpretations may be distinguished, the first referring to organizational processes and the latter referring to the design (division and interconnection) of units and sub-units.

A further important distinction has to do with the question whether goals are considered as "given" to the social planner or designer, or that the process of choosing particular goals is seen as part of the planning process. In the first case the approach is "instrumental", whereas the term "substantive rationality" (Morgan, 1986, p. 37) is sometimes used for the latter. Stated more popularly the instrumental approach is inherent in the phrase "doing things right" whereas the substantial perspective asks the additional question of about "doing the right things".

5.2.1 Synoptic planning and bureaucratic structuring

The ideal of "synoptic" planning is to conceptualize a broad spectrum of long term goals and possible means to attain these goals. Scientific knowledge about instrumental relationships is thought to play an important role in the selection of alternatives. Campbell's (1969) notion of "reforms as experiments" combines a rational planning approach to social (e.g. educational) innovation with the scientific approach of (quasi-) experimentation.

The main characteristics of synoptic planning as a prescriptive principal conducive to effective (in the sense of productive) organizational functioning, as applied to education, are:

- "proactive" statement of goals, careful deduction of concrete goals, operational objectives and assessment instruments;
- decomposition of subject-matter, creating sequences in a way that intermediate and ultimate objectives are approached systematically;
- alignment of teaching methods (design of didactical situations) to subject-matter segments;
- monitoring of the learning progress of students, preferably by means of objective tests.

As stated before, given the orientation towards the primary process, inherent in economic rationality, the synoptic planning approach in education applies most of all to curriculum planning, design of textbooks, instructional design and preparation of (series of) lessons.

When the ideal of rational planning is extended to organizational structuring, related principles about "controlled arrangements" are applied to the division of work, the formation of units and the way supervision is given shape. "Mechanistic structure", "scientific management" and "machine bureaucracy" are the organizational-structural pendants of rational planning (cf. Morgan, 1986, Ch. 2). The basic ideas go back to Max Weber, who stated the principles of bureaucracy as "a form of organization that emphasizes precision, speed, clarity, regularity, reliability, and efficiency achieved through the creation of a fixed division of tasks, hierarchical supervision, and detailed rules and regulations". Although Mintzberg's conception of the professional bureaucracy, applicable to schools and universities, is often treated as the complete antithesis of classical bureaucracy, it should be underlined that the basic notion of standardization and predictability of work-processes, be it with a considerable bandwidth of individual leeway, is retained.

5.2.2 Creating market mechanisms: alignment of individual and organizational rationality

A central assumption in the synoptic planning and bureaucracy interpretation of the rationality paradigm is that organizations act as integrated purposeful units. Individual efforts are expected to be jointly directed at the attainment of organizational goals. In the so-called political image of organizations (Morgan, 1986, Ch. 6) this assumption is rejected, emphasizing that "organizational goals may be rational for some people's interests, but not for others" (ibid, p. 195). The fact that educational organizations consist of relatively autonomous professionals, and loosely coupled sub-systems is seen as a general condition stimulating political behavior of the members of the organization.

In public choice theory the lack of effective control from democratically elected bodies over public sector organizations marks these organizations as being particularly prone to inefficient behavior, essentially caused by the leeway that is given to managers and officers to pursue their own goals besides serving their organization's primary mission[1].

Public choice theory provides the diagnosis of instances of organizational ineffectiveness, such as goal displacement, over-production of services, purposefully counter-productive behavior, "make work" (i.e. officials creating work for each other), hidden agendas and time and energy consuming schisms between sub-units. When discretional leeway of subordinate units goes together with unclear technology this too adds to the overall nourishing ground for inefficient organizational functioning; see Cohen, March and Olsen's famous garbage can model of organizational decisionmaking (Cohen et al., 1972). Not only government departments but also universities are usually mentioned as examples of types of organizations where these phenomena are likely to occur. Market mechanisms and "choice" are seen as the remedy against these sources of organizational mal-functioning.

[1] A more extensive treatment of the implications of public choice theory for school effectiveness research is given elsewhere, Scheerens, 1992, Ch. 2.

Notes of criticism that have been made with respect to the propagation of choice are that parents' choices of schools are based on other than performance criteria (Riley, 1990,

p. 558), that "choice" might stimulate inequalities in education (Hirsch, 1994) and that completely autonomous primary and secondary schools create problems in offering a common educational level for further education (Leune, 1994).

The alleged superiority of private over public schools is the most supportive piece of empirical effectiveness research for the claims of public choice theory, although the significance of the results in question is much debated (Scheerens, 1992). At the macro level there is no evidence whatsoever that national educational systems with more autonomy of schools perform better in the area of basic competencies (Meuret & Scheerens, 1995).

5.2.3 The cybernetic principle: retroactive planning and the learning organization

A less demanding type of planning than synoptic planning is the practice of using evaluative information on organizational functioning as a basis for corrective or

improvement-oriented action. In that case planning is likely to have a more "step by step", incremental orientation, and "goals" or expectations get the function of standards for interpreting evaluative information. The discrepancy between actual achievement and expectations creates the dynamics that could eventually lead to more effectiveness. In cybernetics the cycle of assessment, feedback and corrective action is one of the central principles.

Evaluation-feedback-corrective action and learning cycles comprise of four phases:

- measurement and assessment of performance;
- evaluative interpretation based on "given" or newly created norms;
- communication or feedback of this information to units that have the capacity to take corrective action;
- actual and sustained use (learning) of this information to improve organizational performance.

In the concept of the learning organization procedural and structural conditions thought to be conducive of this type of cycles are of central importance. Examples are: the encouragement of openness and reflectivity, recognition of the importance of exploring different viewpoints and avoiding the defensive attitudes against bureaucratic accountability procedures (Morgan, 1986, p. 90).

When the cybernetic principle is seen as the basic regulatory mechanism there is room for autonomy and "self-regulation" at lower levels in the system. This is a particularly helpful phenomenon in *education* systems, given the usually large degree of professional autonomy of teachers, and tendencies to increase school autonomy as a result of decentralization policies.

5.2.4 The importance of the cybernetic principle

From a theoretical point of view the cybernetic principle of evaluation-feedback-action is very powerful as an explanatory mechanism of organizational effectiveness. It should be noted that evaluation and feedback, apart from being the central mechanism in the interpretation of the rationality paradigm described under the heading of retroactive planning, also have a place in synoptic planning *and* in the perspective from public choice theory. In the former case evaluations are most likely to be used for *control* purposes, while in the latter case there would be an emphasis on positive and negative *incentives* associated with review and evaluations. From the organizational image of the learning organization, adaptive and learning implications of evaluations are highlighted.

Education reform programs will usually be designed according to the synoptic planning model. In such situations evaluation and feedback are used in interaction with more proactive planning techniques. The prototype image of gearing planning and evaluation is program evaluation, preferably designed by means of experimental or quasi-experimental models. In the case of reform programs centered around the enhancement of direct democracy, choice, community control, more continuous information provision to constituencies on the basis of monitoring of critical outcomes and processes has prominence as M&E approach. According to the principle of retroactive planning and cybernetics the evaluation mechanism *is* in fact the motor and core of the reform. Designing reform programs as being driven by the evaluation function is an approach that deserves more attention as a possibly more efficient reform strategy as compared to the classical synoptic planning approaches.

In the subsequent section a closer look will be taken at the cybernetic principle in self-regulating systems and “learning organizations”.

5.2.5 Retroactive planning

The idea of “retroactive planning” is best clarified by zooming in on the ways it differs from “synoptic rational planning”, while still clearly being part of the overall rationality paradigm.

The pure rationality model (Dror, 1968) formally enables the calculation of the optimal choice among alternatives after a complete preference ordering of the end states of a system has been made. This ideal is approached in mathematical decision theory, as in game theory where different preference orderings of different actors can also be taken into account. For most “real life” situations of organizational functioning the assumptions of pure rationality are too strong, however. Simon’s (1964) construct of “bounded rationality”, modifies these assumptions considerably by recognizing that the information capacity of decision-makers is usually limited to taking into consideration just a few possible end states and alternative means.

Cohen, March and Olsen (1972) and March and Olsen (1976) go even further in criticizing the descriptive reality of the pure rationality model. Cohen et al. (1972) describe organized anarchies as characterized by “problematic preferences”, “unclear technology” and “fluid participation”. With respect to problematic preferences, they state that the organization can “better be described as a loose collection of ideas than as a coherent structure; it discovers preferences through action more than it acts on the basis of preferences” (ibid, p. 1). Unclear technology means that the organization members do

not understand the organization's production processes and that the organization operates on the basis of trial and error, "the residue of learning from the accidents of the past" and "pragmatic inventions of necessity". When there is fluid participation, participants vary in the amount of time and effort they devote to different domains of decision making (ibid, p. 1).

According to Cohen et al., decision making in organized anarchies is more like rationalizing after the fact than rational, goal-oriented planning. "From this point of view, an organization is a collection of choices looking for problems, issues and feelings looking for decision situations in which they might be aired, solutions looking for issues to which they might be the answer, and decision makers looking for work" (ibid, p. 2). They see educational organizations as likely candidates for this type of decision making. In terms of coordination, organized anarchies have a fuzzy structure of authority and little capacity for standardization mechanisms.

March and Olsen (1976) describe their reservations with respect to rational decision making in terms of limitations in the complete cycle of choice (see Figure 5.1, where this cycle is depicted).

The relationship between individual cognitions and preferences on the one hand and individual action on the other is limited, because of limitations in the capacity and willingness of individuals to attend to important preferences and because of discrepancies between intentions and actions: "…the capacity for beliefs, attitudes, and

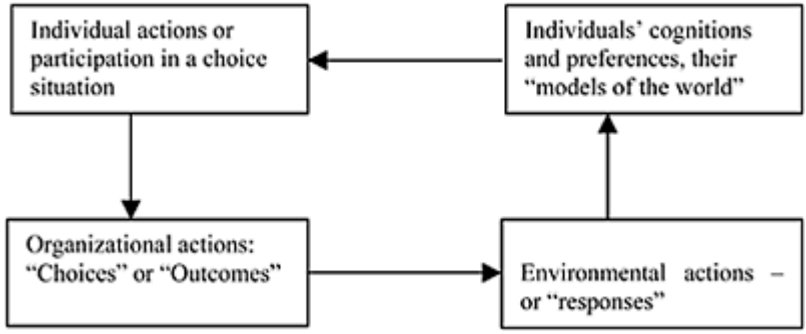


Figure 5.1 The complete cycle of choice, cited from March & Olsen (1976).

concerns is larger than the capacity for action" (ibid, p. 14).

At the same time there may be a loose connection between individual action and organizational action, because internal individual action may be guided by other principles than producing substantive results (e.g. allocating of status, defining organizational truth and virtue). In the same vein they observe that actions and events in the environment sometimes have little to do with what the organization does and that it is sometimes hard to learn from environmental response.

Despite all these limitations on the descriptive reality of rational decision making and planning in organizations, even the most critical analyses leave some room for shaping

reality somewhat more to the core principles. The first type of activity which could bring this about is synoptic planning.

The earlier stated ideal of "synoptic" planning, namely to conceptualize a broad spectrum of long term goals and possible means to attain these goals, contains the basic logic of planned change. In models of planned change the various aspects of synoptic planning are usually structured as phase models (compare the previous chapter); the following description of the different phases is partly based on Ackoff, 1981, 74, 75).

In a *first phase* there is a reflection on values and normative aspects that should be attained through social programs or specific organizational behavior. This first phase can also be taken as the phase of defining the problem domain, in the sense of a system of threats and opportunities that face the organization.

In the *second phase* ends planning takes place in the sense that goals end objectives are specified.

In the *third face* means-planning takes place, where ideally there should be a rationale for selecting the means (examples in education are results of empirical educational effectiveness or practical experience on "what works" in education).

In a *fourth phase* resource planning is focused at determining "what resources will be required, when they will be required, and how to obtain those that will not otherwise be available" (ibid, 75).

In a *fifth phase* design of implementation and control determines "who is to do what, when, and where, and how the implementation and its consequences are to be controlled, that is, kept on track" (ibid, 75).

In a *sixth phase* (which, by the way, is not specifically mentioned by Ackoff), monitoring and evaluation, which can be seen as part of the control processes, are used for feedback and possible modification of means, goals or even values.

The feedback mentioned in this last phase turns the sequence in steps in fact into a circle that can go on and on. Many authors, including Ackoff, do not take the sequence of phases too seriously and say in fact that they make take place in any order. Others, however, see the way one "steps into" the planning, implementation and feedback circle as non-trivial. Borich and Jemelka (1981) see the planned change process as society's attempts to "maintain equilibrium when the system threatens to become disadvantageously influenced by forces whose effects were previously neglected or would have been difficult to predict" (ibid, 216). They see a qualitative difference, however, in two ways of regaining equilibrium. The first being the traditional one where goals are formulated to determine behavior, and which one could see as a proactive orientation (J.S.) the second emphasizing that behavior provides impetus for goals, which they see as a more *retrospective* orientation. They illustrate the difference in these two orientations with a citation from Weick (1969):

> "This sequence in which actions precede goals may well be a more accurate portrait of organizational functioning. The common assertion that goal consensus must occur prior to action obscures the fact that consensus is impossible unless there is something tangible around which it can occur. And this "something tangible" may well turn out to be actions already completed, Thus it is entirely possible that goal statements are

> retrospective rather than prospective. (Weick, 1969, *The Social Psychology of Organizing*, Addison-Wesley, p. 8)

Borich and Jemelka (1982) use this view on retrospective analyses to explain two different views on program evaluation, which they indicate *as forward evaluation* and *backward evaluation.*

Forward looking evaluation is the traditional view, and means assessing discrepancies between a program's objectives and outcomes.

> "Rather than focus on discrepancies between program objectives and outcomes, backward evaluation has as its goal a statement of the values reflected by the program" (ibid, 220).

Once these values are made clear they can be compared to the official or externally established values.

The question should be raised what is gained by emphasizing the retrospective or rather the retroactive view on planned change. According to March and Olsen (1976), learning from experience meets the same fundamental limitations as rational planning.

When goals are ambiguous, which these authors assume, so are norms and standards for interpreting evaluative information. Another limitation is to determine the causality of observed events. They discern four major limitations to organizational learning:

- *role-constrained experiential learning,* if evaluative information is contrary to established routine and role definition it may be disregarded and not frustrated into individual action;
- *superstitious experiential learning;* in this case organizational action does not evoke an environmental response (i.e. is ineffective);
- *audience experiential learning,* when learning of individual organization members does not lead to organizational adaptation;
- *experiential learning under ambiguity;* in this situation it is not clear what happened or why it happened (ibid, p. 56–58).

Even though it is correct that in determining what to evaluate one is forced to address the same kind of selection about what is valuable as in the case of stating objectives, there are at least practical advantages in choosing the retroactive approach:

- on the basis of available instruments a broad scan of current functioning could be made, through which a more efficient selection process on what areas are relevant to address might be expected;
- the difficult and often lengthy deductive process of operationalizing goals can be avoided;
- the basis of subsequent discourse about objectives and means becomes firmly rooted in empirical evidence and is therefore likely to be more concrete and to the point as compared to choosing the "deduction from goals" route;
- particularly when evaluation is of the monitoring type, which means a regular description of the system on key features, the retroactive approach could be viable in creating organizational learning and improvement.

In Table 5.1 the differences between proactive synoptic planning and retroactive planning are summarized.

In educational settings retroactive planning can take place at the level of the national educational system, at regional or district level, and at school level.

At each of these levels a particular type of evaluation or assessment should be put in place as the basis for retroactive planning.

At national level a national assessment program, possibly embedded in a larger system of indicators, would be the most likely instrument. At intermediate levels some kind of indicator or monitoring system could play this role, while at school level school self evaluation instruments are to be seen as the basis for retroactive planning.

The argument for retroactive planning, which one could also refer to as "evaluation centered planning" can be made more forcefully when considering the view on organizational functioning that is inherent in the concept of the learning organization.

Table 5.1 Schematic Comparison of Synoptic and Retroactive Planning.

characteristics	**synoptic planning**	**Retroactive planning**
initial activity	formulate encompassing goals	Assess organization's functioning
choice of means and methods	deduce from scientific knowledge	Induce as improvement of weaknesses in current functioning
scope	a broad scope encompassing all major aspects of the organization	a partial "piecemeal" approach
time-frame	long term	short term
organizational structure	bureaucracy	Learning organization
organizational participation	top-down	Participative

5.3 The Organizational Structural Dimension

In the remaining paragraphs the focus will shift from procedural variations of the rationality model to organizational structures. Various models of the school as an organization are discussed, for their implications of the application of monitoring & evaluation.

5.3.1 Organizational learning in "learning organizations"

Organizational learning can be defined in three different ways:

a. as the sum total of individual learning of the members of the organization
b. in the sense of enhancing the organization's instrumental effectiveness (single loop learning)

c. in the sense of enhancing the organization's external responsiveness (double loop learning)

Re a) *individual learning.* Particularly when organizations are knowledge-intensive, as is the case with educational organizations, there is the strong expectation that workers will keep their knowledge and skills "up-to date". In the corporate world rapidly changing technology and markets are the basic motives for training and human resource development (hrd) activities. In this setting there is a growing interest in a conception of hrd that depends less on formal training, but situates learning in the working place itself, in "learning by doing", subsequently integrating training responsibilities in management functions throughout the organization.

Of course something "extra" is required to convert individual learning into organizational learning. All co-ordination mechanisms that are known from the organiztion literature can play a role in orchestrating individual learning in a way that the benefits for the organization as a whole are maximized. Examples are: a clear mission and result orientation of the organization, organizational structures that enable exchange between units and sub-units, facilitation and supporting technology (i.e. "group-ware") for communication and collaboration between members of the organization and even standardization of outcomes and processes. This latter coordination mechanism does not fit in so well with the expectations of flexibility and a more "organic" functioning of "learning organizations", however.

Re b) *organizational learning as single-loop learning.* The concepts of single- and double-loop learning, as introduced by Argyris (Argyris & Schön, 1978) form the core of the theoretical basis of learning organizations. Single-loop learning rests in an ability to detect and correct error in relations to a given set of operating norms (Morgan, 1986, 88). In its turn single-loop learning should be seen against the conceptual background of cybernetics ("steermanship"), which sees the self-regulation of organisms and organizations as based on processes of information exchange involving negative feedback. Learning in this sense is characterized as a gradual shaping of behavior, constantly correcting for mistakes or sub-optimal solutions. In social contexts "right" and "wrong" are determined by agreements and norms, hence the qualification of the kinds of norms that are central in single-loop learning. When these are defined as the operating norms, they should be taken as the preferred end-states of an organization's primary process, or the objectives of the organization's core business. Single-loop learning takes these objectives as given and concentrates optimal selection of means and technology to attain these objectives. This instrumental perspective is quite similar to the approach of school effectiveness research, in which scientific methods are used to find out which organizational and instructional conditions are most effective in realizing key-outcomes. In a less stylized form the day to day running of an organization can also be seen as guided by this instrumental approach. In case of organizations with "unclear" technologies, such as schools, such a trial-and-error approach to improving the effectiveness of the primary process appears quite relevant, at least in theory. In actual practice such organizations are also likely to have quite a few barriers that work against a learning orientation (see the subsequent section on schools as professional bureaucracies). Single loop learning emphasizes the need for information that can shape a gradual improvement of primary and supporting organizational processes in obtaining basic outcomes.

Re c) double-loop learning. “Double-loop learning depends on being able to take a “double look” at the situation by questioning the relevance of the operating norms) (Morgan, 1986, p. 88). So, double loop learning does not take pre-fixed operating norms (or objectives) for granted, but makes them the object of analysis and reflection. The basic motive to choose this approach is grounded in an open-systems view of organizations, where situational conditions set the stage for defining what organizational effectiveness means. Contingency theory has provided further insight in the kind of situational conditions that matter: changes in the predictability of the environment and the nature of the organization’s technology being the most prominent types of “contingency factors”. The more dramatic the dynamics of these situational conditions, the stronger the need for critical review of the organization’s operating norms and “double-loop learning”. The type of analysis and information gathering that is required for double loop learning cannot stop at an internal review of “instrumental effectiveness”, but also needs an external scan of situational conditions. The emphasis on monitoring with an open mind about operating norms and objectives resembles the orientation of “backward evaluation” and “retroactive planning” described in the previous section. Analysis of the organizational structures that facilitate or hinder organizational learning in the sense of double-loop learning form the basis of further clarification of the fashionable term of the learning organization. Before doing so, it is important to realize that the relevance of this concept, particularly as far as double-loop learning, strongly depends on the dynamics of situational factors. We will turn back to this issue after a closer look at the nature of educational organizations (i.e. schools and universities).

Morgan (ibid, p. 89, 90) mentions three types of failures of organizations in implementing double-loop learning:

Firstly, formal planning approaches including organizational goals and objectives, clearly defined roles and bureaucratic structure with pronounced hierarchy, create fragmented structures “that do not encourage employees to think for themselves”. Fragmented operation of the organization is further seen as to be encouraged by political processes in which each sub-unit pursues its own goals and means are treated more or less as ends in themselves (ibid, p. 89). It is interesting to note that the author judges highly sophisticated single-loop learning systems in such bureaucratic contexts as actually preventing double-loop learning, “since people are unable or not prepared to challenge underlying assumptions” (p. 90).

Bureaucratic accountability systems, where people are held responsible for their performance within a system that rewards success and punishes failure, is seen by Morgan as a second barrier to double-loop learning. He sees such systems as fostering defensiveness of employees and as an incentive for covering up and “impression management” (make situations look better than they actually are). He also criticizes the tendency to oversimplification as complex issues are difficult to address in such a context.

The third barrier to double-loop learning, mentioned by Morgan is the tendency of organizations to rationalize and meet problems with rhetoric. Organizations develop “theories in use” that may be socially reinforced to constructions that are insufficiently rooted in reality.

According to Morgan these barriers can be overcome by encouraging openness and reflectivity, a divergent thinking approach to the analysis and solution of complex

problems, which means that the importance of exploring different viewpoints is underlined. In the third place rational planning approaches that "impose" goals, objectives and targets should be avoided and instead "means where intelligence and direction can emerge from ongoing organizational processes" should be fostered. In short Morgan sees organic structure, a bottom up participatory approach and less formal ways of planning and reflection as core conditions for double-loop learning.

He completes the picture on organizational structures that enhance double-loop learning by referring to some concepts from systems-theory.

The principle of *holographic systems* means that each part comprises a complete image of the whole. This metaphor emphasizes a certain redundancy in functions across sub-systems and implies a more diverged authority systems; self-steering work-teams can be seen as practical examples. Further following the metaphor of the organization as a human brain, strong interconnectivity between the sub-units is emphasized. The principle of *requisite variety* places some boundary on the amount of redundancy (the degree to which units should be able to fulfil similar functions as others) in stating that "the internal diversity of any self-regulating system must match the variety and complexity of its environment". A practical implication is that organizations should pay close attention to the boundary relations between organizational units and their environments.

Apart from these two characteristics that bear on the structure of the organization there are two other principles that refer more to the procedural dimension of organizational functioning: the principles of *minimum critical specification* and *learning to learn*.

The principle of minimum critical specification bears some resemblance to the idea of subsidiarity, which popularly stated comes down to the principle that higher levels of an organizational structure should not do things that can also be carried out at a lower level. Similarly the principle of minimum critical specification speaks for limiting the pre-specification of organizational arrangements and processes to the maximum. In this way "minimum critical specification suggests that managers and organizational designers should primarily adopt a facilitating and orchestrating role, creating 'enabling conditions' that allow a system to find its own form" (ibid, p. 101). Flexibility in organizational functioning is likely to result from such minimal management, which in its turn is seen as a favorable context for "inquiry driven action". The principle of *learning to learn* should prevent flexibility turning into chaos, and it is here that we are back with the organization's capacity for single- and double-loop learning.

How useful are the concept of organizational learning and the metaphor of the learning organization for understanding the functioning of monitoring and evaluation in educational organizations? Dealing with this question will be postponed until a closer look is taken at the specific characteristics of such organizations. This will be done by examining yet another metaphor of organizations: the professional bureaucracy as well as a perspective on school management that was generated by empirical school effectiveness research.

5.3.2 Management in the school as a "professional bureaucracy"

The concept of the school as a professional bureaucracy was developed by Mintzberg (1979). The main characteristics of the professional bureaucracy are the following ones:

- the internal cohesion of the organization depends predominantly on the standardization of skills of the functionaries—teachers in our case—which is based on long specialized training;
- a large degree of professional autonomy of the teachers, whereby loyalty towards the organization has to compete with loyalty to the profession and loyalty to the "client";
- a relatively underdeveloped interest in the external environment; the basic assumption in the professional bureaucracy is that the environment may be complex but is, at the same time, stable;
- a specific role for leadership and management which is seen as mostly administrative and not substantive in the sense that school-leaders are expected to give direction to teachers, but rather to play a submissive and supportive role;
- technology in the professional bureaucracy has on the one hand the nature of a "well-stocked tool box", but on the other hand holds the challenge of adapting these standard tools and solutions to ever changing circumstances in the work with clients (in this case pupils);
- there is little readiness and openness for change and opposition against rationalization of the work and monitoring performance among the professionals;
- recruitment of personnel is the most important control measure within the organization; within the framework of the profession as such adaptation of the initial training is the most important control mechanism.

The concept of the school as a professional bureaucracy is related to Weick's image of the school as a "loosely coupled organization". "Loose coupling", according to Weick, refers to a relatively small interdependence among sub-systems like teachers among themselves, and head teachers and teachers. At the same time there is also little cohesion between aspects of the organization's functioning, like the coupling of means and goals and between decisions planned and actual implementation. An example of this last phenomenon is the well-known situation (at least in the Netherlands) where schools have a nicely phrased "school work plan", which is safely put away in a cupboard and bears little relationship to what is actually happening (cf. Van der Werf, 1988). As far as technology is concerned Weick emphasizes the "fuzzy" technology of schools, where there is little consensus about goals and means and evaluation of central means-to-end relationships is difficult.

What kind of school management would fit in an organizational structure as depicted in the images of the professional bureaucracy and the loosely coupled system? The general answer would seem to be that such structures require only minimal management. In such a structure little need is felt for long-term planning and strategy development. In the prototype form there would be no intermediary structures and hence no middle-management. Operational management is firmly in the hands of the professionals (teachers) in the operating core (the classroom) of the organization. The metaphor of the teacher as the King/Queen in his/her classroom comes to mind. Monitoring and performance control will tend to be seen as threats to the professional autonomy. The—according to this theoretical image—most potent management domain, that is human resource management, is in the actual practice of most countries, strongly limited because of fixed conditions of labor.

In short, the image of a professional bureaucracy is adamant in warning us for the limitations of trying to develop a type of management in schools that touches the primary

process of teaching and learning. Interestingly this orientation is exactly the central focus in the concept of "educational" or instructional leadership, as developed in the context of effective schools research.

5.3.3 Educational leadership as a characteristic of "effective schools"

Empirical school effectiveness research basically addresses the question which organizational and instructional conditions explain why some schools have better results, in terms of student achievement than others, after taking differences in the student intake between schools into account. A more complete review of the methodology and knowledge-base from school effectiveness research will be given in Part 4 of this book. In this context, however, one of the conditions that was generated from school effectiveness research, dealing with educational leadership and management will be referred to briefly. This, because it is considered relevant in obtaining a more balanced overview of conceptions of schools as organizations for the main question of this chapter on the role of educational monitoring and evaluation for the functioning of schools as organizations.

In the operational definitions and instruments concerning educational leadership a general division into two conceptions can be made:

a. general leadership skills applied to educational organizations:

- articulated leadership
- information provision
- orchestration of participative decision making
- coordination

b. instructional/educational leadership in a narrower sense, i.e. leadership directed at the school's primary process and its immediate facilitative conditions:

- time devoted to educational versus administrative tasks
- the head teacher as a meta-controller of classroom processes
- the head teacher as a quality controller of classroom teachers
- the head teacher as a facilitator of work-oriented teams
- the head teacher as an initiator and facilitator of staff professionalization

Of these two dimensions, the second, namely leadership focused on the school's primary process, should be considered as central. The other dimension addresses the specific demands required for leading and controlling organizations in which professionals at the operating core need to have a considerable degree of autonomy. As a whole educational leadership can be seen as a phenomenon that needs to strike a balance between several extremes: direction versus giving leeway to autonomous professionals, monitoring versus counseling and using structures and procedures versus creating a shared (achievement-oriented) culture. Sammons, Hillman and Mortimore (1995) in this context refer to the leading professional.

The system-theoretical concept of meta-control is perhaps the most suitable to express the indirect control and influence an educationally or instructionally oriented school leader exercises on the school's primary process. Of course this does not imply that the head teacher is looking over the teachers' shoulder all the time, but he or she is

'involved' in important decisions on objectives and methods, and visibly cares about overall achievement levels and individual pupils' progress. From the set of components that were listed in Table 5.1 it is evident that the meta-control of the school leader is exercised in a non-authoritarian way, expressing concern about pupils, individual staff members, and team work.

Some authors, who define educational leadership, say more about structural conditions surrounding the instructional process, whereas others are more focused on cultural aspects. Irwin (1986, p. 126) belongs to the former category in mentioning the following aspects of educational leadership: the school leader:

- functions as an initiator and coordinator of the improvement of the instructional programme;
- states a clear mission of the school;
- has a task-oriented attitude;
- establishes clear objectives;
- supports innovation strategies;
- stimulates effective instruction;
- is quite visible in the organization;
- sees to it that pupils' progress is monitored regularly;
- delegates routine tasks to others;
- regularly observes both the work of teachers and pupils.

Leithwood and Montgomery (1982, p. 334) mention the following more cultural aspects of educational leadership:

- stimulation of an achievement-oriented school policy;
- commitment to all types of educational decisions in the school;
- stimulating cooperative relationship between teachers, in order to realize a joint commitment to the achievement-oriented school mission;
- advertising the central mission of the school and obtaining of support of external stakeholders.

In more recent views on educational leadership, inspired by the concept of the learning organization, motivating staff by providing incentives and creating consensus on goals are emphasized. Mitchell and Tucker's concepts of transactional leadership and transformational leadership (Mitchell & Tucker, 1992) form a case in point. Staff development and the 'human resource' factor are further underlined in these approaches. These newer perspectives do not create a sharp break with the longer existing conceptualizations of educational leadership, but emphasize the cultural and the staffing mode of schooling.

Scheerens (1992, p. 89) draws attention to the point that the rather heavy requirements of an educational leader do not necessarily rest on the shoulders of just one individual:

> *"At first glance the description of 'educational leadership' conjures up an image of a show of management strength: not only the routine work necessary for the smooth running of a school, but also active involvement with what is traditionally regarded as the work sphere of the routine assignments leave sufficient time for the more pedagogic tasks.*

Nevertheless, this leadership does not always have to come down to the efforts of one main leader. From the school effectiveness research of Mortimore et al (1988) it emerges that deputy heads in particular fulfil educational leadership duties. Delegation can go further than this level: it is desirable that, given the consensus of a basic mission for the school, there is as broad as possible a participation in the decision making. In the end certain effects of pedagogic leadership such as a homogeneous team, will fulfil a self-generating function and act as a substitute for school leadership (according to Kerr's (1977) idea of 'substitutes for leadership')."

5.3.4 Schools as learning organizations?

Before dealing more directly with the question about the relevance of the metaphor of the learning organization when applied to schools an attempt should be made to integrate the two conceptions of schools organizations and school management, presented in the above sections on the professional bureaucracy and educational leadership. At first sight the two perspectives provide considerable cognitive dissonance. How is the theoretically based image of the "professional bureaucracy", which also shows a lot of face-validity and common sense, to be reconciled with the empirically based concept of educational leadership?

In the first place, schools are nowadays not the exact copies of professional bureaucracies. Schools have been confronted with more demanding external requirements of both higher administrative levels and the consumers of education. In the "knowledge society", knowledge changes rapidly and there is a debate on whether to concentrate at teaching knowledge as such or rather strategies to acquire knowledge ("learning to learn"). As far as administration is concerned, in several countries schools are given more autonomy in the domains of management and finance whereas—sometimes—there is less autonomy in the domain of the curriculum.

All these external changes work as as many pressures on the school to reconsider its functioning and perhaps even to change and innovate. And the importance of the role of the school head is now widely recognized. Another important reason why matters may start to depart from the picture of the professional bureaucracy is the availability of technology. Not just teaching technology, like computer-assisted instruction, but also management and evaluation technology, in the form of school management information systems, pupil monitoring systems and school selfevaluation methods.

In the second place "educational leadership" is not completely contrary to certain requirements of the professional bureaucracy. Firstly, the educationally oriented school head can approach an individual teacher as a fellow-professional and colleague and, in this capacity discuss educational issues. Secondly, there can be a gradual implementation of creating meetings and work-sessions where teachers come together, and, in the presence of the head, discuss educational topics. The role of the school head as an educational leader does definitely not preclude a democratic attitude nor a collegial, supportive, coaching role. The point is that educational leadership by no means excludes a collegial, counseling-like approach, which would be more easily accepted by teachers. Thirdly, the educational leader can opt for a management strategy that leaves the core of

professional autonomy of teachers, namely the process of teaching, largely as it is. This approach comes down to "freeing process and monitoring output" and can be seen as a form of functional decentralization at the school level. In short, in many aspects the image of the school as a professional bureaucracy is still a valid image of the reality of school functioning in many educational systems. In the area of output standardization, an increased focus on results and outcomes and in the use of technology the image needs to be corrected and updated.

When turning to the question of the validity and usefulness of seeing schools as "learning organizations" the subsequent interpretations of organizational learning and structural characteristics of the learning organization will be examined one by one.

Organizational learning as the total of individual learning in the organization would seem to be a relevant way of looking at schools. The professional skills of the teachers form the core of the image of the professional bureaucracy. Although quantitative indicators are lacking, the impression is that re-training and in-service training of teachers is an important and wide-spread phenomenon in many countries. Coordination, however, can be seen as the Achilles-heel of schools. Individual learning can only result in organizational learning if the individual learning efforts are orchestrated, coordinated and brought under a common set of goals or organizational mission.

Organizational learning as "single-loop learning" faces the same problem of overcoming individualism. In addition it presupposes a substantive concern and some degree of analytic thinking about the goals and means of school functioning *as a whole*. The former is likely to be stimulated when external demands on the quality and the outcomes of schooling are becoming more pronounced. The latter is not easy to come by and would be quite dependent on the specific skills and sophistication of school heads. Approaches from educational support structures to help schools in this general field have been of three general types. Firstly in the sense of pro-active planning of school activities, e.g. by developing school working-plans, Secondly by introducing procedures and instruments for a more retroactive and diagnostic approach, by means of specific forms of school self-evaluation and thirdly by stimulating forms of professional consultation and cooperation between staff.

Organizational learning in the sense of double-loop-learning is the feature that in the fast moving business world is the key-piece of the learning organization. The degree to which this idea fits the reality of schools differs for the various educational levels, and will be higher for forms of tertiary education than for primary education. Particularly at the primary and secondary level there is likely to be a strong standardization of outcomes, for example in the form of examinations. To the degree that achievement outcomes of schools are standardized this is a "given" for the school, which means that there is no need for double-loop learning with regards to the key-area of the primary outcomes of the school. In other areas, like the pedagogical function of the school, some degree of reflection on norms might be considered relevant, however. Yet, the overall conclusion about the relative relevance of the three forms of organizational learning the first two: orchestrating individual learning and single-loop learning are considered most important for schools.

When considering the more structural features of the image of the learning organization, a few of them can be seen as features that are already present in more traditional images of the school, like the professional bureaucracy. "Minimal critical

specification" or subsidiarity has been present in traditional school organizations in an exaggerated form of a head who is in marginal control over the school's primary process. In the concept of educational leadership, particularly taken as "metacontrol" this principle obtains a clear application in schools. Perhaps there is also something of holographic structure *avant la lettre* inherent in the considerable professional autonomy of teachers and the traditionally marginal role of school management. For quite some time it has not been too hart a job for a teacher to act as head, when the need for this would arise.

The recommendation of organic structure as a pre-condition for organizational learning and learning organizations cannot be accepted uncritically when schools are being considered. The issue touches upon an old debate in education on whether teaching should be considered an art on the one hand or a "science" or technology on the other. According to the first view educational improvement is a matter of improving human resources and human relations in school organization and of improving cultural aspects.

According to the latter at least a core educational technology is feasible, and applicable to the extent that there is also consensus and standardization on key educational outcomes. This debate is not likely to be settled in an all or nothing way. It is, for example, quite feasible to differentiate between means and ends, on this issue. In quite a few educational systems there is simultaneously an enlargement of freedom and flexibility concerning means and processes, and a tightening of outcome requirements. Constructivist perspectives of learning and instruction stimulate experimentation and variation in the structuring of learning arrangements, and even to a loosening of the traditional structures for matching teachers and students. At the same time these developments challenge traditional priorities in the substance and methods of assessing student achievement. Challenges to traditional teaching and learning are also inherent in the direct and indirect impact of the information society. Indirect in the sense of changes in the cognitive orientation of students raised in an environment that is more and more dominated by mass media and ICT. Direct in the sense of an enormous range of still underutilized potential of ICT applications in the teaching and learning situation.

Despite of the fact that not all of the features of the construct of the learning organization appear to be directly relevant and applicable to schools, the conclusion is that it is still a stimulating metaphor for school improvement, in a context that is partly standardized but also very much in movement. The core of the matter is learning according to the cybernetic principle and negative feedback, which brings us back to the concepts of retroactive planning and ultimately the function of educational evaluation for educational planning an management.

5.4 Conclusion: The Centrality of External and Internal School Self-Evaluation in Learning and Adapting School Organizations

Improvement of schooling is a matter of optimizing adaptation to changing situational conditions and of optimizing instrumental effectiveness. To the degree that core outcomes are standardized, external monitoring of school performance can function as a stimulant and focus of internal school improvement. Internal school evaluation can follow in the wake of external evaluation, in the sense of specification and disaggregation of achievement outcomes and exploring potential sources and "causes" of the variation in

outcomes. In this sense school evaluation, both internally and externally, follows the pattern of traditional goal oriented, or in the terminology of Borich and Jemelka, “forward looking” evaluation.

To the degree that schools want to determine specific school goals in addition to the official, externally determined objectives, and to the extent that school want to identify intermediary goals or supportive processes, internal “backward looking” evaluation is required. Such backward looking evaluation supports a retroactive orientation to planning and is meant as an internal reflection on the goals and means of schooling.

From a theoretical point of view evaluation and monitoring processes are at the core of the model of “learning” organizations that seek to improve their external responsiveness and internal effectiveness. In a practical sense school evaluation and monitoring are considered as viable levers of school improvement and as a perhaps more effective innovation strategy than pro-active planning approaches. The arguments for this latter assertion can be summarized as follows:

- starting the improvement processes by means of an empirical scan of the existing functioning of the school and looking from the existing situation to a more idealized situation reflecting norms and goals avoids difficult processes of goal operationalization and provides an empirical basis to any discussion on goals and means;
- such an “evaluation centered” approach is likely to enhance result orientation in the development of goals and norms of school functioning, as any educational evaluation is likely to contain elements of outcome measurement;
- an important “side effect” of preparing and carrying out school self-evaluation activities is the fact that collaborative activity that it presupposes enhances taskrelated collaboration and coordination of work within schools;
- more or less implied is that such collaboration involves a participatory approach in which staff and school management work together;
- the results of the M&E activities are a practical basis for continued internal reflection and experimentation and a concrete basis for communicating with external stakeholders.

PART 3
Assessment of Student Achievement

6
Basic Elements of Educational Measurement

6.1 Introduction

In this chapter, an overview will be given of the basic elements of educational measurement. The main topic of this chapter is an inventory of the state of the art with respect to the specification and development of assessment instruments, both for knowledge and skills. So the emphasis will be on tests for the cognitive and psychomotor domain. (Tests for the affective domain, say the domain of attitudes, motivation and emotions, will not be treated here, for this topic one is referred to Anderson & Bourke, 2000).

This chapter is not meant as a manual for test construction. The objective of this chapter is to describe the relationships between the various activities related to educational measurement, to give an overview of the scientific and professional standards for educational measurement and to provide references where more detailed descriptions of the process of educational measurement can be found.

Criteria for educational measurement are well documented. According to the *Standards for Educational and Psychological Tests* (American Psychological Association [APA], American Educational Research Association & National Council on Measurement in Education, 1985), test developers are obliged "to anticipate how their tests will be used and misused,…and to design tests and accompanying materials in ways that promote proper use". Standards for proper test development and use are not confined to the United States. In the Netherlands, for instance, a permanent commission (the COTAN) of the Dutch Institute for Psychology (NIP) has defined a set of standards for psychological and educational tests and it uses these standards for continuous evaluation of all published tests. The results of these evaluations are published on a regular basis (Evers, van VlietMulder, Resing, Starren, van Alphen de Veer & van Boxtel, 2002). The standards issued by the COTAN consist of seven aspects:

- Purpose and scope;
- Quality of test material;
- Quality of the manual;
- Norms;
- Reliability;
- Content and construct validity;
- Criterion validity.

These aspects will be elaborated in this and the two following chapters. Besides from the perspective of quality criteria, educational measurement can also be viewed from a process perspective. Major steps in the process, that will be outlined below, are:

- Definition of a test purpose. This entails the target of the test (for instance, curriculum-based proficiencies, cognitive or psychomotor abilities) and the kind of decisions that have to be made (such as, mastery decisions, pass/fail decisions, selection, prediction).
- Definition of test specifications. Given the purpose of the test, the content domain of the test can be delineated and the level at which the content must be tested. Further, an appropriate test format must be chosen.
- Construction of test materials, such as construction of multiple-choice items and open-ended items, or the construction of performance assessments.
- Test administration. This includes the administration of traditional paper-andpencil tests, computerized adaptive testing, simulations and real-life assessments.
- Test scoring and test and item analysis, including the transformation of scores into grades and evaluation of the quality of the test as a measurement instrument.

Besides these “traditional” topics, several new topics have gained prominence with the growing influence of computer technology on educational assessment. This chapter will be concluded with some of these topics: assessment systems, item banking, optimal test construction, and computerized adaptive testing.

6.2 Test Purposes

The information provided by the test constructer should enable test users (instructors, schools, government agencies) and test takers (students) to judge whether the test is suited for the intended purpose. Therefore, information about the purpose and the scope of the test must be made available. That is, the ambition level of the inferences based on the outcome of the assessment should be clear. This entails explicit demarcation of the domain to be assessed, explicit operationalization of concepts and constructs, and explicit criteria for the selection of the content.

Many taxonomies of purposes for educational tests are available (Bloom, Hatings & Madaus, 1971; Gronlund, 1981; Mehrens & Lehmann, 1984). Here a taxonomy will be used which is an adaptation from a taxonomy for tests of ability and achievement by Millman and Green (1989), that is also suited for performance assessments and tests for affective domains. An overview of the taxonomy is given in Table 6.1. The rows of the table relate to the type of inference desired, which can either be a description of individuals, groups and systems. The columns of the table relate to the domain to which the inferences are made. A curriculum domain refers to skills, knowledge and attitudes acquired as a result of instruction on a curricular content. In the curricular domain the emphasis will generally be on educational achievement, but it also includes the acquisition of abilities, that is, knowledge and skills that transcendent the actual curriculum taught. The competency domain refers to knowledge, skills and attitudes that can be derived from theory or from practice that are not directly related to some curriculum.

Table 6.1Functional Description of Test Purposes (Adapted from Millman & Greene, 1989).

	Domain to which inferences will be made				
		Curricular domain		Competency domain	Future criterion setting
Type of inference desired	Before instruction	During instruction	After instruction		
Description of individual examinees' attainments	Placement	Diagnosis	Grading	Reporting	Guidance and counseling
Mastery decision	Selection	Instructional guidance, streaming	Promotion	Certification	Selection Admission Licensing
Description performance for a group or system	Preinstruction status for evaluation	Process and curriculum evaluation	Postinstruction status for evaluation; Reporting	Construct measurement for evaluation; Reporting	Research

There are, of course, explicit relations between curricula and competencies. Roughly speaking, the distinction between inferences to the curricular domain and to the competency domain is that the first inferences are about the educational achievement level, while the second are inferences about knowledge, skills and affective goals that transcendent the actual curriculum. Inferences to a future criterion setting are about knowledge, skills and affective goals that should persist in the period after the instruction ended. Then testing serves to predict performance in future settings.

6.3 Quality Criteria for Assessments

Irrespective of its purpose, a test must always conform to a number of quality criteria. As a measurement, a test must be valid and reliable. Besides the criteria for the test as a measurement instrument, there are criteria that relate to the testing procedure: appropriateness, feasibility and transparency.

Validity

Validity is related to the meaning, usefulness and correctness of the conclusions based on the test scores. Cronbach (1971) asserts that what needs to be valid is the meaning or interpretation of the scores as well as any implications for action that this meaning entails. This definition encompasses both an evidential aspect of validity and a

consequential aspect. The overview given here mainly endorses the first aspect. Messick (1989, 1995) views the concept of validity as "an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment". The interpretations of assessment scores are valid to the extent that these interpretations are supported by appropriate evidence. Mislevy, Steinberg and Almond (2003) view educational measurement as a form of "evidentiary reasoning", that covers the inferential procedure from observable student behavior in particular circumstances to their general level of knowledge and skills. This inference includes interrelated elements such as theories concerning the targeted domain, learning processes, the elements and processes involved in test construction, administration, scoring and reporting, statistical models and probability-based reasoning.

Depending on the inferences that must be made, two relevant forms of validity are content validity and criterion validity.

Content validity is the extent to which a test measures what it is intended to measure. In psychological testing, this is usually labeled construct validity, that is, the extent to which the test is a measure for some theoretical construct. Content and construct validity have two aspects: content relevance and representativeness. A judgment of content relevance and representativeness is based on a specification of the boundaries and structure of the domain to be tested. The domain can be explored using curriculum analysis, test analysis, job analysis and domain theory (Fleishman & Quaintance, 1984).

Criterion validity is the extent to which the test scores are empirically related to criterion measures. The relationships can be either predictive or concurrent. Criterion validity is directly related to the test purpose. If the test should predict performance in some future criterion setting, the test constructer should provide empirical evidence that supports this claim. Information on concurrent criterion validity can support the claims about the constructs being measured. This can take the form of convergent and divergent relations. The test scores, or sub-scores, should have substantial correlations with indicators or tests of the same or closely related constructs, and low correlations with indicators and tests of constructs that are theoretically distinct from the target construct. Finally, the measurement should be relatively independent of the measurement technique used. Convergence of indicators, divergence of constructs and consistency across methods can be concurrently studied using a correlation analysis technique called the multitraitmultimethod technique (Campbell & Fiske, 1959).

Reliability

Reliability is the extent to which a test measures consistently. Inconsistency stems from factors influencing the outcome of the measurement that are not part of the construct of interest. One might think of the properties of the test (for instance, test length) or the assessment procedure (for instance, rater effects). The aim of a reliability analysis is quantification of the consistency and inconsistency of the student performance on the test. The reliability of a test or assessment is affected by the objectivity of the scoring of the test, the specificity of the items or tasks, the difficulty level and the test length. Scoring is objective if there is a deterministic scoring rule that transforms the responses into scores (Ebel & Frisbie, 1986). In this definition, multiple-choice items can be

objectively scored, and open-ended questions cannot be objectively scored because this entails human judgments causing random fluctuations. However, also “fill-in-the blank” questions, short answer questions, arithmetic tasks with an unambiguous numerical outcome etc. can also be scored objectively. An item is called specific if only students that possess the relevant knowledge can give a correct response. Various errors in items provide students with clues to the correct response that are unrelated to the actual topics tested. Further, in the next chapter it will be shown that the possibility of guessing a correct response on a multiple-choice item decreases the reliability.

Appropriateness and Feasibility

The criteria related to appropriateness concern a number of practical considerations. The assessment should be efficient, in terms of time and costs required for test construction and administration. A test should be fair in the sense that the test is standardized and every student should be able to demonstrate his or her ability or proficiency level. Further, the allotted time should be sufficient, unless the test is a speed test.

Transparency

Students should know in advance the content coverage and content sources of the test, the item formats that will be used, the test scoring methods and rules applied, and the level of achievement needed to pass the test. In most instances, students are offered the opportunity to review their work and an appeal procedure.

6.4 Test Specifications

The ultimate goal of testing is to make valid inferences regarding some domain of proficiency or ability. An important step in the creation of assessments is specification of the content that should be assessed and the level of cognitive behavior that should be targeted. The content and cognitive behavior dimension can be used to create a so-called table of specifications where the cell entries give the relative importance of a specific combination of content and cognitive behavioral level. Other important test specifications are the time a test may take, the item format and the number of items.

6.4.1 Specification of test content

A specification of test content concerns the identification of specific areas of subject matter that should be included in the test. Millman and Green (1989) consider five substantive characteristics of test content that should provide a clear demarcation of the domain to be assessed in relation to the test purpose. The characteristics are as follows.

Sources of test content

The specification of the sources of test content depends on the domain to which the inferences will be made: the curricular domain, the competency domain or the domain of

a future criterion. In the curricular domain, the content can be derived from explicit curricular objectives, curricular outlines and blueprints, textbooks or other instructional material. In the competency domain, the content cannot be derived from a specific curriculum, but it is derived from theoretical conceptualizations of knowledge, skills, mental abilities and achievements. Tests developed to predict performance in a future criterion setting must be based on an analysis of the requirements of that setting. Millman and Green (1989) distinguish three steps in this analysis. First, the specific cognitive requirements of the criterion setting are identified through job analysis or task analysis, or through research or research synthesis with respect to future academic settings. Second, the content specification is developed using the criterion directly or indicators known or hypothesized to be related to the criterion. Third, the relationship between the performance on the predictive test and the performance in the criterion setting must be established.

Dimensionality of test content

Dimensionality refers to the conceptual or theoretical homogeneity or heterogeneity of the content domain. In principle, assessing a heterogeneous domain implies using separate test scores for different dimensions, which may affect the overall reliability of the test. The overall reliability is positively related with the correlation between the dimensions. As an alternative to using subscores, one may combine the subscores on the dimensions into a composite score. In that case, the weighting of the subscores should reflect the relative importance of each dimension in the conceptualization of the domain.

Domain- versus norm-referenced interpretation

A domain-referenced interpretation refers to an absolute performance level, whereas a norm-referenced interpretation refers to the performance level relative to a population. The content specification of a domain-referenced test requires a detailed description of the entire domain and subscores are usually attributed to all aspects of the performance. Norm-referenced tests, on the other hand, require a summary score, and the selected content should support the meaningfulness of this score.

Bandwidth versus fidelity

Closely related to the two previous points is the tradeoff between bandwidth and fidelity. Choosing test content involves a tradeoff between the breath of content coverage and the reliability of the subscores. Tests with a very narrow scope can be very reliable, but their relevance to the content domain may be negligible.

Content distribution

The distribution of items across the content domain should reflect the relevant domain conceptualization.

6.4.2 Specification of cognitive behavior level

Besides a detailed content specification, making valid inferences with respect to some domain of proficiency or ability requires an analysis of the cognitive level at which the target behavior should be performed. The most used taxonomy of levels cognitive behavior is the well-known taxonomy by Bloom (Bloom, Engelhart, Furst, Hill & Krathwohl, 1956). Bloom distinguishes six hierarchical categories: knowledge, comprehension (translation and interpretation), application, analysis, synthesis and evaluation. In many practical testing situations, this taxonomy is somewhat simplified and directly linked with item types. Here, a distinction is made between items assessing knowledge, understanding, application and problem solving. In this breakdown, items assess knowledge if they require reproduction without any substantial extension. Items assessing comprehension require production of new information based on the supplied information. Application requires the use of the information in some outside setting, where only one possible solution is valid. Finally, problem solving involves productive and creative thinking, in some outside setting, where there is usually more than one feasible solution.

An alternative taxonomy that has lately gained attention is based on the distinction between declarative and procedural knowledge (Snow & Lohman, 1989). Declarative knowledge involves facts, concepts, principles and procedures, that are stored in an organizational framework and retrieved when needed. Procedural knowledge, on the other hand, involves many related semantic networks. Snow and Lohman (1989) distinguish between context bound organizational frameworks that are more difficult to construct but easily to retrieve, and semantic memory which is easy to construct, for instance by rote learning, but only offers short-term benefits. Procedural knowledge is conceptualized as developed from declarative knowledge in a number of stages, at the end of which the behavior is automated. One step further in this hierarchy is so-called strategic knowledge (Greeno, 1980). This involves the development of goals and strategies for attaining certain objectives. Several authors (Crooks, 1988; Green, Halpin & Halpin, 1990; Stiggins, Griswold & Stikelund, 1989) argue that traditional achievement testing does not properly assess higher level educational outcomes, such as procedural and strategic knowledge. On the other hand, Roid & Haladyna (1982) attribute this apparent shortcoming to lack of adequate conceptualization of higher-level outcomes. Haladyna (1992, 1994) developed several advanced item-formats, such as the context-dependent item set, that appear quite suitable for measuring higher level outcomes.

Table of Specifications

The relation between the content and cognitive behavior dimension can be defined in a so-called table of specifications. An alternative name is content-by-processmatrix. The table serves as a blueprint for the test and can be used as an item-writing scheme. The principle objective of the table of specifications is to assure that the test is a valid reflection of the domain and test purpose. An example is given in Table 6.2. The artificial example pertains to a test for a course on research in the social sciences.

For the cognitive behavior dimension, a distinction is made between four item types: items assessing knowledge, items assessing understanding, items assessing application

and items assessing problem solving. The content dimension is hierarchically ordered in topics and subtopics. The cell entries of the table give the relative importance of a specific combination of content and cognitive behavioral level in the test. The percentages in the example of Table 6.2 define the way that the items are distributed over the content-by-level grid. When the total number of items has been fixed, the percentages translate to numbers of items.

Table 6.2 Table of Specifications.

		Cognitive level addressed by items				
Content		Knowledge	Understanding	Application	Problem Solving	
Statistics	Discrete Distributions	3	3	4	0	10
	Continuous Distributions	3	3	4	0	10
	Estimation	2	2	2	2	8
	Hypothesis Testing	2	4	4	0	10
Methodology	Experiments	3	4	4	4	15
	Observational Studies	3	3	4	5	15
	Case Studies	3	3	3	1	10
Measurement	Reliability	3	3	3	3	12
	Validity	1	2	3	4	10
		23	27	31	19	100

6.5 Test Formats

Also the choice of the type or format of test is governed by the quality criteria given above. This entails a tradeoff between validity, reliability, appropriateness, feasibility and transparency. For instance, some test format may be highly valid, in the sense that it is an almost perfect match to the criterion that should be measured; yet considerations of efficiency, comparability and economy may make the format completely infeasible. For choosing a test format, several taxonomies are available.

Haladyna (1992, 1994) distinguishes between selected-response formats, where the correct answer is selected among several choices, and constructed-response formats, where the response is constructed. The former format may also be labeled the multiple-choice format. The latter format encompasses open-ended questions (further divided into short- and long-answer questions) completion, essays, and performance assessments (which itself spans a universe running from small highly structures assignments, to

simulations and to portfolios containing work samples). On one hand, questions may be administered as an oral test, a paper-and-pencil test, or a computerized test, while a completion format only applies to paper-and-pencil and computerized tests. As a pragmatic solution, formats will be organized into selected response formats, constructed response formats (including oral tests and essays), and performance assessments (including simulation and portfolios).

6.5.1 Selected response formats

Selected response items are items where the student has to choose an alternative from some pre-specified list of alternatives. There are three basic versions.

a) True-false items. The item consists of a statement and the student has to determine whether this statement is true or false.
b) Multiple-choice items. The item consists of a stem and a number of choicealternatives. One of the response alternatives is the correct answer, the others are so-called distractors. The student has to pick the correct alternative or the best alternative.
c) Matching items. The item consists of two lists of alternatives, and the student has to match alternatives from each list.

These three formats will be discussed further below. In addition to the three basic formats, a number of more complex selected response formats have been developed that are aimed at testing complex thinking and problem solving.

d) Context-dependent item sets. In this format a number of selected response items with a basic format are organized in some larger framework. For instance, a test of language comprehension may consist of a number of texts, and the comprehension of each text is assessed using a number of multiple-choice items nested under the text. Item sets are described using labels as interpretative exercises, scenarios, vignettes, item bundles, problem sets, super-items and testlets.

Finally, the widespread use of computers has opened up a whole new range of possibilities in testing. Both the three basic formats and context-dependent item sets are straightforwardly adapted for presentation on computer. However, the use of computers has also facilitated a number of new possibilities that will be discussed under the heading “Innovations”.

True-false items

Compared to multiple-choice items, true-false items are relatively easy to construct, because there is no need to construct response alternatives. It is essential that the statement in the stem can only be classified as true or false if the student really has knowledge or understanding of the content. That is, a student without the proper knowledge and understanding should not be able to infer the truth of the statement via intelligent use of unintended clues. One of the main faults made in this format is that the wording of the statement closely matches the wording used in the instructional materials. In that case the item no longer measures knowledge or understanding but memory and rote recall.

The main advantage of true-false items is that a good content coverage can be achieved, because many items can be administered in a relatively short time. A point often made against using true-false items is that they are only suited for testing knowledge. However, Ebel and Frisbie (1991) and Frisbie and Becker (1991) give a number of suggestions on how to use true-false items to measure higher thought processes.

Another point of criticism on true-false items is related to the fact that the probability of guessing the correct response is 50%. This high guessing probability should, of course, be taken into account when determining a cut-off score. On the other hand, the guessing probability is the same for all students so it does not systematically distort the ordering of the students' performances. In the next chapter it is explained that the guessing probability is negatively related to the test reliability, so this has the consequence that the number of items that must be administered goes up as the guessing probability increases. The relation between the number of items that has to be administered and the guessing probability will also be returned to in the next chapter.

A last important point concerns the context in which the true-false items function. When students have the opportunity to peruse the items in advance, they can attain remarkably high scores using a strategy where they only look at the items in the "true"-category in the test preparation phase and merely recognize these items in the test phase. So the rationale that mastering all items in a large (public) item bank is analogous to mastering the target domain cannot be used in this context.

Multiple-choice items

Multiple-choice items are so common that many novices in education are unaware of the specific difficulties attached to this format. However, experience in largescale high-stakes testing programs shows that it takes substantial training before items are constructed that can function problem-free in a high-stakes setting. The quality criteria for multiple-choice items and item writing rules can be found in Haladyna (1994, 1997), important older contributions are Ebel (1951), Wesman (1971), and Woods (1977). In general, the items should be accurate and valid, there should be one and only one correct response, the alternatives should be mutually exclusive and the wording of the stem and the response alternatives should be unambiguous. Further, all options should be plausible and attractive to students who lack the specific knowledge or understanding addressed by the item. Intelligent and test-wise students should not be given clues about the correct alternative that are not based on the actual domain measured. Common clues are that the correct response is longer and differently worded than the incorrect alternatives. Finally, in most situations the possibility of decreasing the guessing probability by increasing the number of response alternatives is very limited. When the test constructer runs out of proper alternatives, highly illogical or even corny alternatives are added that students can eliminate on sight.

Properly constructed multiple-choice items have many advantages: the test is standardized and can be objectively scored, and the possibility of administering many items supports content coverage. The two main disadvantages are that the construction of proper items is time consuming, and the test format is unsuitable for testing proficiencies

requiring writing skills, the presentation of arguments, mathematical reasoning and for testing performance in real-life situations.

Matching items

Matching items are often recommended for testing associations, definitions, or characteristics or examples of concepts (Haladyna, 1994). Further, matching items are efficient because several questions are implicitly integrated into one item, and since the format does not entail construction of distracters, matching items are easier to construct than common multiple-choice items.

In practice, there is a tendency to make both lists of choices equally long. This has the disadvantage that making one wrong combination automatically induces additional errors. A solution is to have the students match elements of a shortlist to the elements of a much longer list. The constructing of two lists of matching options, which are homogeneous logically and homogeneous with respect to difficulty level is no minor task, so successful practical examples of this format are not numerous.

Context-dependent item set

A context-dependent item set consists of context-dependent material followed by a set of selected response items, usually multiple-choice items. The context-material may be in textual form, such as a problem, scenario or case study, or in pictorial form, such as a photo, figure, table or chart. As an example, Haladyna (1992) considers a so-called vignette, which is a problem-solving context in which examinees respond with decisions or actions via multiple-choice items. The main motivation for using context-dependent item sets is that they can be used for evaluating higher-order thinking, such as problem solving and critical thinking. The analysis of results from tests consisting of context-dependent item sets requires specific psychometric models to account for the variability of response behavior within and between the item sets. In psychometric literature, these psychometric models are referred to as testlet models or item bundle models (Wainer & Kiely, 1987). These models will be further explored in the next chapter.

Innovations

Roughly speaking, the innovations with respect to item types that are supported by computers fall into three categories: the mode of presentation, the response mode and item generation.

Item Presentation

One of the major advantages of administering tests via the computer is the possibility of using non-text media in items. This may increase the authenticity of the test. For instance, audio presentations can be integrated in tests of listening skills in language and music. Examples are tests of English proficiency of non-native speakers (ACT, Inc., 1999; ETS, 1998; Godwin, 1999; Nissan, 1999), or tests of listening skills for employees and professionals. An example is a listening comprehension test being investigated by the

Law School Admissions Council for possible inclusion in their exam program (ACT, Inc., 1998). Using video can also enhance task authenticity in tests. An interesting application is the video-based test of conflict resolution skills by Olson-Buchanan, Drasgow, Moberg, Mead, Keenan, and Donovan (1998).

Response mode

Traditional selected-response item types require students to mark a correct response alternative. Computer presentations can broaden this basic principle in various ways. One may ask students to click on and select the proper sentence from a reading passage, to select a part of a graphic, or to make selections in a data base (Parshall, Stewart, & Ritter, 1996). In a placement tests for adult basic education in the Netherlands, Verschoor and Straetmans (2000) use mathematics items where students have to select points in a figural histogram, on a scale, or on a dial.

For an extensive overview of innovative item types in computerized testing one is referred to Parshall, Davey, and Pashley (2000).

Item shells and item cloning

In item-cloning techniques (see, for instance, Bejar, 1993, or Roid & Haladyna, 1982) operational items are derived from "parent items" via one or more transformation rules. These parent items have been known as "item forms", "item templates", or "item shells", whereas the items generated from them are know now widely known as "item clones". Closely related to this approach are so-called "replacement set procedures" (Millman & Westman, 1989) where test items are generated from a parent item by the computer. In this approach, the computer puts the answers to multiple-choice items in random order, picks distractors from a list of possible wrong answers, and, in numerical problems, substitutes random numbers in a specific spot in the item stem and adjusts the alternatives accordingly. In this approach, items are generated "on-the-fly", that is, the computer generates a new version of the item for every student.

An important question is whether clones and items generated-on-the fly from the same parent item have comparable statistical characteristics. Empirical studies addressing this question are reported in, for example, Hively, Patterson and Page (1968), Macready (1983), Macready and Merwin (1973) and Meisner, Luecht and Reckase (1993). The general impression from these studies is that the variability between clones from the same parent is much smaller than between parents, but not small enough to justify the assumption of identical values. Of course, the size of the remaining variability depends on various factors, such as the type of knowledge or skill tested and the implementation of the item cloning technique. Psychometric models for analyzing results of tests based on item shells and clones will be discussed in the next chapter.

6.5.2 Constructed response formats

Constructed response formats can be ordered with respect to the objectivity of their scoring. The following two categories are distinguished.

a) Completion items and close tests. The student is presented with a statement or some larger piece of text and is required to fill in the blanks. In the first case, one speaks of a completion item, in the second case of a close test. The response is scored using a list of acceptable alternatives.
b) Short-answer items and short essay items. Short-answer items consist of a question requiring a limited answer. The answer can be restricted by limiting the available number of lines that can be used, or by limiting the number of elements asked for in the question (give three reasons, give four places, etc.). Essay items require longer answers. Essay tests consisting of just one task, say writing an essay about a certain topic, should rather be classified under performance assessments.

Completion items and close tests

In this format, it is essential that the blanks are placed in such positions in the text, that it is obvious to the student what class of response is required, say a name, a date, vehicle, a verb, etc. The scoring list should contain synonyms and acceptable misspellings, but not words that are related to completely different interpretations of the task.

Completion items resemble multiple-choice items in the sense that they can (in principle) be objectively scored and can provide a good content coverage as a result of the number that can be administered in a certain time span. However, they can elicit unwanted study habits focused on learning keywords without any real understanding. Therefore, completion items rarely evaluated favorably.

Short-answer items and short essay items

An important difference between selected and constructed response formats is that in the latter case, the student is far less certain of the kind of response that is required. Therefore, the test constructer should phrase the question in such a way that it is explicit about the scope of the response expected. This includes the length of the response, the number of arguments, causes or other elements required, which aspects need to be detailed, etc.

To enhance objectivity, rating open-ended questions such as short-answer items and essays should be based on a so-called rating model. The rating model includes a model response, a listing of elements that should be presents in the response and a scheme for the attribution of score points to these elements. Interrater reliability can be assessed via independent ratings of the responses.

Nowadays, scoring open-ended questions and even essays need not depend on human raters. Several software programs for scoring essays are available (Shermis and Burnstein, 2003; Burnstein, 2003). These programs have been developed from various theoretical orientations, such as linguistics, statistics and neural network theory. They learn to copy the decisions by one ore more human raters to such a level that the interrater reliability between the human example and the software package becomes very high.

To a large extent, the disadvantages of multiple-choice items are the advantages of open-ended questions, and, vise versa, the advantages of multiple-choice items are the disadvantages of open-ended questions. The number of open-ended questions that can be

presented in a specific time span is smaller that the number of multiplechoice items, so content coverage is more limited. Human scoring involves an element of subjectivity, so the reliability will generally be affected to some degree. Further, if writing skills and clarity of expression are not the focus of the assessment, differences in these abilities may confound the assessment. Finally, the items proper may be easy to construct, the construction of a consistent rating model is no minor task.

6.5.3 Performance assessments

Performance assessments are assessments where the behavior that must be displayed in the test situations has a close resemblance to the behavior required in the real-life domain of interest. That is, the performances included in the assessment are samples of the kind of performance emphasized in the generalization to a domain, or they are high-fidelity simulations of this kind of performance (Kane, 1992; Wiggins, 1989). The assessment may involve completely authentic hands-on tasks, but also simulations, such as solving in-basket-problems, role-playing, fact-finding, conducting interviews, or performing experiments.

Performance assessments are sometimes labeled “authentic measurements”, but this term is somewhat value-laden, since it is often used in a tradition that opposes so-called “traditional testing” (Wiggins, 1998). In general, the performances assessed are generally complex, and the assessment is characterized by a high fidelity, in the sense that the involved activities are directly related to activities outside the educational setting.

Kane, Crooks and Cohen (1999) have addressed the question how the validity of performance assessments can be established. Their analysis identifies three major inferences involved in the interpretation of performance assessments: scoring of observed performances, generalization to a domain of assessment performances like those included in the assessment, and extrapolation to the larger performance domain of interest. This chain of inference starts with posing the existence of a target domain and defining the student’s target score for the domain as the expected score over all possible performances in the domain. However, for practical reasons (time constraints, logistics, safety), the target domain is limited to a sub-domain called the universe of generalization. The student’s score on this domain is called the universe score. Observed performances should be a random or representative sample from the universe of generalization. The process of validation is grounding the inferences from the performances to the observed scores, from the observed scores to the universe scores, and from the universe scores to the target scores. When building performance assessments, several decisions must be made. The first is defining the level of standardization. Standardization supports interrater reliability and comparability across students, but it may threaten authenticity and generalizability to the target domain. Another decision concerns the length and complexity of the performance task. However, having one very specific and lengthy task rather than diversifying the assessment may lead a low reliability of the assessment. Evaluating the reliability of a performance assessment will be returned to in section 6.6.

6.5.4 Choosing a format

Selecting the most appropriate, effective, efficient item types to measure the intended outcomes requires balancing the practical constraints of a testing situation against the requirements of content coverage and cognitive level defined in the table of test specifications.

An important constraint is the number of items that can be administered in a certain time span. Table 6.3 gives an indication of the time it takes to administer various item types by Mehrens and Lehmann (1975). Of course, the figures are indications that may vary over students, item constructors, and the topic tested.

Table 6.3 Indication of Response Time per Item Type.

Item type	Response Time
True-false	50 seconds
Multiple-choice, 2 alternatives	50 seconds
Multiple-choice, 3 alternatives	60 seconds
Multiple-choice, 4–5 alternatives	75 seconds
Open-ended, response one word or sentence	1 minute
Open-ended, response quarter page	5 minutes
Open-ended, response half page	10 minutes
Open-ended, response one page	25 minutes
Open-ended, response two pages	60 minutes

Berk (1999) presents a rating of the advantages and disadvantages of four test item and assessment methods as given in Table 6.4. The tradeoff is between what items can measure best (characteristic 1) against their practical and technical strengths (the other characteristics). Berk (1999) concludes that the multiple-choice item is often to be preferred “because of the types of cognitive outcomes it can measure, and its marked advantages in content coverage, administration, scoring, and reliability over other item formats. However, it is not the preferred choice at the highest levels of cognition”.

Table 6.4 Ratings of Advantages and Disadvantages of Four Test Item and Assessment Methods (Adapted from Berk, 1999).

Characteristic	**Multiplechoice**	**Completion**	**Essay**	**Performance Assessment**[1]
1 **Cognitive Outcomes**				
Knowledge	++	++	++	++

	Comprehension	++	– –	++	++
	Application	++	– –	++	++
	Analysis	+	– –	++	++
	Synthesis	– –	– –	++	++
	Evaluation	+	– –	++	++
2	**Construction**				
	Difficulty	– –	–	+	+
	Efficiency	– –	–	+	–
	Cost	– –	–	+	–
3	**Content Coverage**				
	Scope	++	+	– –	– –
	Depth	++	– –	++	++
4	**Administration**				
	Difficulty	++	+	+	– –
	Efficiency	++	+	– –	– –
	Cost	++	+	–	– –
5	**Scoring**				
	Difficulty	++	–	– –	– –
	Efficiency	++	–	– –	– –
	Cost	++	–	– –	– –
	Guessing	–	+	++	++
	Accuracy	++	+	+	+
	Consistency	++	+	+	+

[1]Includes constructed-response formats other than completion and essay, such as direct observation, simulations, oral discourse and assessment center.

6.6 Test and Item Analysis

In the next chapter, an introduction to educational measurement theory will be given that is completely based on item response theory (IRT). IRT provides the theoretical underpinning for the construction of measurement instruments, linking and equating measurements, the evaluation of test bias, item banking, optimal test construction and computerized adaptive testing. However, all these applications require fitting an appropriate IRT model. If that model is found, it provides a framework for evaluation of item and test reliability. However, fitting an IRT model can be quite a tedious process,

and one can never be absolutely sure that the model fit is good enough. Therefore, in practical situations one needs indices of item and test reliability that can be effortlessly computed and do not depend on a model. These indices are provided by classical test theory (Gulliksen, 1950, Lord & Novick, 1968). Before the principles and major consequences of classical test theory are outlined, first an example of its application will be given. Consider the item and test analysis given in Table 6.5.

Table 6.5 Example of a Test and Item Analysis (Number of observations: 2290).

Mean=8.192

S.D.=1.623

Alpha=0.432

Item	p-value	ρ_{kc}	ρ_{kd}	ρ_{kd}
1	.792	.445	.014	.145
2	.965	.419	.035	.019
3	.810	.354	.147	.054
4	.917	.052	.155	.052
5	.678	.378	.388	.072
6	.756	.338	.048	.032
7	.651	−.091	.061	.291
8	.770	.126	.100	.121
9	.948	.472	.182	.172
10	.905	.537	.193	.037

The example concerns a test consisting of ten multiple-choice items with three response alternatives. The mean and the standard deviation of the frequency distribution of the number-correct scores are given in the heading of the table. Also given is the value of Cronbach's Alpha coefficient, which serves as an indication of test reliability. Alpha takes values between zero and one. As a rule of thumb, a test is usually considered sufficiently reliable if the value of Alpha is above 0.80. In the present example, this is not the case. Probably, the test length is too short, and further, there may be items that do not function properly. Inspecting the information on the items given below the heading can further assess this. The column labeled pvalue gives the proportions of correct responses. The columns labeled ρ_{kc} and ρ_{kd} give the correlation between the total score and the item score for the correct response and the two distractors, respectively. (The index *k* stands for the item, *c* and *d* for correct and incorrect, respectively). If the total score were a perfect proficiency measure, the item-test correlation for the correct response, ρ_{kc}, should be high, because giving a correct response is an indication of proficiency also. By an analogous reasoning, the item-test correlation for a distractor, ρ_{kd}, should be close to zero, or even negative, because keying a wrong alternative indicates a low proficiency level.

Now the number correct score is no perfect measure of ability because of the unreliability of test scores and because of the presence of poorly functioning items which detriment the number-correct score. However, the essence of the rationale for expecting an item-test correlation ρ_{kc} that is higher than ρ_{kd} remains valid. For the item statistics in Table 6.5, it can be inferred that the items 4 and 7 did not function properly: ρ_{kc} is close to zero for item 4 and negative for item 7. Further, for these two items, it can be seen that $\rho_{kd}>\rho_{kc}$, for the first distractor of item 4 and the second distractor of item 7. So these response alternatives were more appealing to proficient students as the supposedly correct alternatives. Practice shows that items with these patterns of item-test correlations are usually faulty. There might have been made an administrative error in identifying the correct response alternative, or there might be some error in the content of the correct alternative.

Reliability theory

The inferences made in the example about the reliability of the test and the quality of the items are based on classical test theory (CTT). CTT starts with considering two random experiments. In the first experiment a student, which will be labeled *i (i=1,…, N),* draws a test score from some distribution. So the test score of a student is a random variable. It will be denoted by X_i. The fact that the student's test score is a random variable reflects the fact that the measurement is unreliable, and subject to random fluctuations. The expected value of the test score is equal to the true score, that is,

$$E(X_i) = T_i$$

The difference between the test score and the true score is the error component *ei*, that is, $e_i=X_i-T_i$. The second random experiment concerns drawing students from some population. Now not only the observed scores Xi but also the true scores T_i are random variables. As a result, for a given population, it holds that the observed score of a randomly drawn student, say X, is equal to a true score, say T, and an error term e. That is

$$X = T + e \tag{1}$$

Since the test is only administered to each student once, the error terms within students and the error terms between students cannot be assessed separately; they are confounded. It is assumed that the expectation of the error term, both within students and between students is equal to zero, that is,

$$E(e) = 0$$

Further, the true score and the error term are independent, so their covariance is zero:

$$Cov(T, e) = 0$$

These assumptions do not imply a model; all that was done is making a decomposition of observed scores into two components: true scores and errors. This decomposition provides the basis for defining reliability. Suppose that there are two tests that are strictly parallel, that is, students have the same true score T on the test and the error components,

say *e* and *e'* have exactly the same variance. The observed scores on the tests will be denoted by *X* and *X'*. Then reliability, denoted by $\rho XX'$ can be defined as the correlation between two strictly parallel tests, so

$$\rho_{XX'} = \frac{Cov(X, X')}{\sqrt{Var(X)Var(X')}}$$

From this definition, a second definition of reliability can be derived. Since it holds that $Cov(X,X')=Cov(T+e, T+e')=Cov(T,T)+Cov(T,e')+Cov(T',e)+Cov(e,e')=Var(T)+0+0+0= Var(T)$, reliability can also be defined as

$$\rho_{XX'} = \frac{Var(T)}{Var(X)} \qquad (2)$$

So reliability is also the ratio of the true and observed score variance. In other words, it can be viewed as the proportion of systematic variance in the total variance. The reliability $\rho_{xx'}$ can be estimated in several ways. The first is trough the administration of two strictly parallel tests. In practice, this is often hampered by the fact that strictly parallel tests are difficult to find, and the administration of the two tests may induce learning effects. Another way is based on viewing all items in the test as a parallel measures. Under this assumption, reliability can be estimated by Cronbach's Alpha coefficient (see, Lord & Novick, 1968). If the items are dichotomous (that is, either correct or incorrect), Cronbach's Alpha becomes the well-known KR-20 coefficient.

Two important consequences of CCT can be delivered from the following considerations. Suppose that there is a test with test scores *X*, and, as in formula (1), $X=T+e$. Further, there is a criterion, say some other test, where the scores C can also be written as $C=T'+ e'$. The interest is in the correlation between the test scores and the scores on the criterion. So the observed criterion scores are the sum of true scores *T'* and error terms *e'*. In principle, the interest is not so much in the correlation between the observed scores, but in the correlation between the true scores *T* and *T'*. This correlation can be written as

$$\rho_{TT'} = \frac{\sigma_{TT'}}{\sigma_T \sigma_{T'}} = \frac{\rho_{XC}}{\sqrt{\rho_{XX'}\rho_{CC'}}}$$

So the correlation between the true scores is equal to the observed correlation divided by the square root of the reliabilities of the test scores and the criterion scores, respectively. From this identity, two important conclusions can be drawn. First, it follows that

$$\rho_{TT'} \geq \rho_{XC}$$

that is, the observed correlation is lower than the correlation between the true scores, say, the true correlation. This is the so-called attenuation effect. It entails that the unreliability of the scores suppresses the correlation between the observed scores. And if the test and the criterion were perfectly reliable, the observed correlation would equal the true correlation. The number of items in a test is an important antecedent of reliability, so the bottom line is that it is useless to compute correlations between very short, unreliable instruments. This could, in principle, be solved by correcting the observed correlation by

dividing it by the square root of the reliabilities of the test scores and the criterion scores. However, the bias in the estimates of these reliabilities has unpredictable consequences. In the next chapter an alternative approach using IRT will be sketched.

The second consequence follows from the fact that it can be derived that

$$\sqrt{\rho_{XX'}} \geq \rho_{XC}$$

Now if the criterion *C* is a strictly parallel test, this inequality implies that the square root of the reliability is an upper bound for validity. So validity and reliability are related and one cannot ignore reliability when pursuing validity.

To assess the reliability of a measurement the variance of the observed scores *X*, say σ^2_X, has been split up into two components, the variance of the true scores *T*, say σ^2_i, and the error variance, say σ^2_X. So that is,

$$\sigma^2_X = \sigma^2_i + \sigma^2_e$$

In many instances, however, specific sources of error variance can be identified, such as tasks, raters, occasions, and so forth. Estimation of these variance components can be done using an analysis of variance technique, known as generalizability theory (Brennan, 1992; Cronbach, Glaser, Nanda & Rajaratnam, 1972). As an example, consider an application to performance assessments (Brennan en Johnson, 1995), where there is interest in the effects of two so-called facets: raters *(r)* and tasks *(t)*. The total observed score variance of the assessments of raters of the responses of students to tasks can now be decomposed as

$$\sigma^2_X = \sigma^2_i + \sigma^2_t + \sigma^2_r + \sigma^2_{it} + \sigma^2_{ir} + \sigma^2_{tr} + \sigma^2_e$$

where σ^2_X is the true score variance of the students, σ^2_t is the variance attributable to the tasks, σ^2_r is the variance attributable to the raters, and σ^2_{it}, σ^2_{ir} and σ^2_{tr} are the interaction between persons and tasks, persons and raters, and raters and tasks, respectively. All variances on the left-hand side, except σ^2_i, can be considered as error variances. Estimation of the variance components is a so-called G-study. With the components estimated, a D-study can then be done to assess the number of tasks n_t and raters n_r that must be invoked to attain a certain level of reliability. These numbers can be computed in two ways, depending on the definition of reliability that is used. If the variance of the tasks, σ^2_t, the variance of the raters σ^2_r, and all variances attributed to interactions, σ^2_{it}, σ^2_{ir} and σ^2_{tr}, are considered as errors, the reliability coefficient becomes

$$\rho^2 = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_t/n_t + \sigma^2_r/n_r + \sigma^2_{it}/n_t + \sigma^2_{ir}/n_r + \sigma^2_{tr}/n_t n_r + \sigma^2_e/n_t n_r} \qquad (3)$$

This coefficient is relevant if absolute judgments are to be made about the true score level of the student. However, if relative judgments are to be made, that is, if differentiations are made between the students, irrespective the difficulty of the tasks or the strictness of the raters, the variance of raters and tasks, and the variance of the interaction between

these two are no longer relevant. Therefore, they are removed from the denominator, which leads to the coefficient

$$\rho^2 = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_{it}^2/n_t + \sigma_{ir}^2/n_r + \sigma_e^2/n_t n_r} \quad (4)$$

The final remark of this section pertains to the relation between reliability as it is defined in here using CTT, and the alternative definition defined using IRT that will be discussed in the next chapter. Reliability in CTT pertains to the extent to which students can be distinguished based on their test scores. For this, it is essential that the true scores of the students vary. The reliability coefficients defined above by formulas (2), (3) and (4) are all equal zero if $\sigma_i^2 = 0$. The dependence of the reliability on the variance of the true scores can be misused. Consider a test for visual ability administered to a sample of children of 8 years old enrolled in regular schools. It turns out that the reliability is too low to make an impression on test publishers. Therefore, the test constructor adds a sample of children of 8 years old enrolled in schools for visually handicapped to the sample. These children score very low, which blows up the score variance, and leads to a much higher reliability. As such, this circumstance does not invalidate the concept of reliability. What the test constructor has done is changing the definition of the population of interest. The first coefficient related to distinguishing between non-handicapped children, while the second coefficient related to distinguishing visually impaired children from children who are not visually impaired. The point is that CTT test and item indices must always be interpreted relative to some population. In the next chapter, it will be shown that the main motivation for the development of IRT is separating the effects of tests and items on one hand, and the population of students on the other hand. It will be shown that this leads to another definition of reliability.

6.7 Assessment Systems

Computerized assessment systems play an increasingly important role in educational testing. They support activities as processing answer sheets, performing statistical analyses of test results and printing reports. In computer based testing, items are presented and responded to via the computer. The data storage capacities of computers have created the possibility of developing large item banks from which tests can be assembled which are customized to specific applications. Tests can be tailor made to suit the ability level of a certain candidate or group of candidates. And examination results, results of national assessments and the results of pupil monitoring systems form the input of data warehouses, from which new tests can be assembled.

Among the newer developments where computers play an important role is testing on demand, that is, assembly and administration of a test whenever the educational process calls for it. The need for testing on demand is a result of the trend toward further flexibilization and modularization in education. Testing on demand has emphasized the importance of data communication and multi-user environments, where computer networks have supplanted the stand-alone personal computer. In the Netherlands, this movement toward large-scale test service systems is evident in the systems developed by

the Dutch Open University (the SYS system) and at the National Institute for Educational Measurement (Cito, the Cito-TSS system and the pupil monitoring system).

Another important new influence of information technology derives from its power to control multi-media environments which can be used to integrate auditive and visual media in the testing process. Multi-media environments and virtual reality are already widely used in education for simulation purposes and educational testing.

The purpose of this section is to present an overview of the elements of educational assessment systems or computerized testing service systems (TSS). Today's practice shows many different applications and the needs greatly vary from one user to another. On one hand, there are the network versions, run by professionals ranging for item writers to psychometricians and processing hundreds of students in test centers daily; on the other hand there are systems implemented on a personal computer used by individual teachers for supporting their educational work. This means that there is not just one TSS that suits everybody, in fact, for some users some elements of a TSS may not be interesting at all. For instance, the individual teacher who processes the data of classes may not have much use for advanced psychometry, if only, because the sizes of the samples obtained in classes are too small for computing reliable test and item statistics. In this section, the TSS will be sketched in a broad perspective, meaning that an overview of the aspects and relations of a TSS will be presented here and details will be mostly left alone. The aim is to present potential users a framework for deciding which elements play a role in the specific situation of interest. This overview is given by considering a theoretical model of a TSS, developed both at the Dutch Open University (Breukers, et al., 1992) and at Cito (Glas, 1997). Many of the features discussed here are implemented in the Cito-TSS, which is a large-scale, highly professional-oriented network system. Further, it will be indicated which commercial and non-commercial systems developed by others may play a role in computerizing one's testing process.

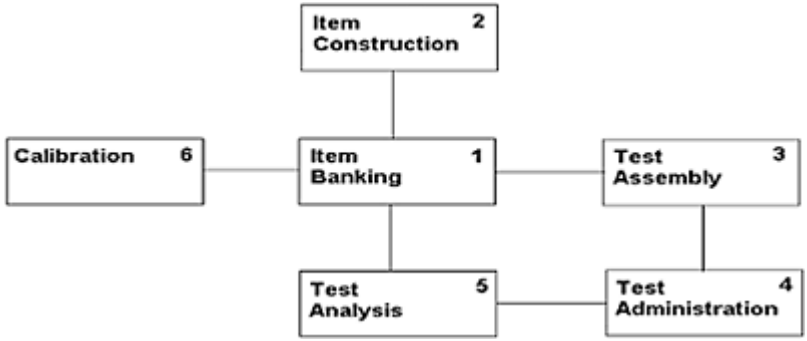


Figure 6.1 Overview of an assessment system.

Though the focus of this section is mainly on educational measurement, most of the material presented here also applies to other fields of assessment. An overview of the application of computerized testing in the field of psychological assessment can be found in Butcher (1987).

An educational assessment system, or testing service system is an integrated and computerized system for the construction and storage of items, assembly and deliverance of tests, and the analysis of the test results. The structure of the TSS follows from the various activities that can be distinguished in the functional flow of the testing process. These activities are item banking, item construction, test assembly, test administration, test analysis, and calibration. A distinct module supports every activity in the system; the complete system is depicted in Figure 6.1. The modules are given numbers related to the order in which they come into play.

6.7.1 Item banking

Item banking supports storage, maintenance and retrieval of items. An item is viewed in the broadest sense, so it may be any task or series of tasks presented to the students for the purpose of measurement. This may include items for paper-andpencil administration or computer-based test administration and tasks for performance assessments. Apart from the actual test material, the stored information may also include scoring instructions and feedback, both for the student and the test administrator.

Before the items can be entered into the system, a so-called item bank structure must be defined. The variables in the item bank structure will generally reflect the variables used in test specifications, so items may be categorized with respect to content matter and level of cognitive behavior tested. Since items and tasks may be used in different situations, the item bank structure will probably be broader than the table of specifications of a specific test. Further, since an item bank usually covers a whole range of tests, possible relations between items must be stored. For instance, a cluster of items might belong to the same case and must always be presented together, or items should not be simultaneously present in one test because they are too much alike.

Finally, the item bank may also store empirical information, such as the frequency of use of items, or the groups of students administered certain items. Empirical information also includes psychometric information such as statistics on item difficulty and the ability level of the populations, which responded to the item. Psychometric information can be used for norming new assessments assembled from the item bank.

6.7.2 Item construction

Item writing for paper-and-pencil test administration is nowadays generally done using a word processor. This has several advantages, such as the availability of layout facilities, easy integration of text and graphics and availability of a spelling checker and a thesaurus. Linking a word processor with the item bank creates the possibility of improving the quality of the items to be written by accessing existing items with a comparable item classification and the empirical information gathered from administration of these items.

Millman and Westman (1989) give an overview of the further possibilities of computerizing the process of item writing. They distinguish five levels of automation of item writing, labeled author supplied approach, replacement-set procedures, computer-supplied prototype items, subject-matter mapping and discourse analysis.

The *Author supplied approach* is item writing by an author using the word processor as a tool for checking the spelling, for integrating text and graphics, for importing layout macros, etc. In its ultimate form, the author supplied approach entails an expert- and management system to support defining item structures and tables of specifications for tests, writing parallel items, item reviewing, formal acceptation of the items, etc.

In *Replacement-set procedures*, test items are generated from a blue-print by the computer. As already outlined in a previous section, this entails that the computer puts the answers to multiple-choice items in random order, picks distractors from a list of possible wrong answers, and, in numerical problems, substitutes random numbers in a specific spot in the item stem.

In the *Computer-supplied prototype items* procedure, proposed by Millman and Westman (1989), the author and the computer interact to write the text of the item. Item writers specify the mental operation they wish to measure, in response to which the computer generates a set of prototypes, which is then further refined in a dialogue to produce an empty item, which is then supplied with content by accessing a database.

In *Subject-matter mapping*, the author and the computer work interactively to build a concept taxonomy (Merril & Tennyson, 1977, also see Minsky, 1975; Novak & Gowin, 1984) where the key concepts of the achievement domain and their relations are modeled. Based on this structure and a lexicon provided by the author, the computer generates crude test items for the item writer to review.

In the *Discourse analysis* approach, it was attempted to construct items directly from text. Wolfe (1976) developed algorithms to transform sentences into test questions, but the resulting items were either very simple or did not make sense, so this line of development seems to have been abandoned.

6.7.3 Item bank calibration

In the following two chapters, the role of measurement models, in particular IRT models, will be outlined in detail. Here a concise introduction and review of two important applications of IRT related of item banking will be given: optimal test construction and computerized adaptive testing. IRT models (Rasch, 1960; Birnbaum, 1968; Lord, 1980; Fischer & Molenaar, 1995) are characterized by three features: (1) they relate to responses of persons to items, (2) parameter separation, meaning that the influence of items and persons on the responses are modeled by disjunctive sets of parameters, say item difficulty parameters and person ability parameters, and (3) the stochastic nature of the responses of persons to items. Item and person parameters need not be scalars, it might well be the case that ability is multidimensional and must be represented by a vector of scalars. Parameter separation gives the possibility to store item characteristics, that is, item parameters in an item bank, which are independent of the characteristics of the sample of the students who responded to the items.

An important aspect of IRT models is that they are models and, as a consequence, their legitimacy must be tested. In other words, statistical proof must be presented that the representation of person ability and item difficulty of a specific IRT model sufficiently describes the observed responses of the persons to the items. If the specific IRT model does not fit, another IRT model should be sought. Fortunately, for a large class of IRT models, statistical testing procedures have been developed that will not only evaluate

model fit, but also give information with respect to specific model violations and with respect to the direction in which the IRT model should be altered to obtain a fitting model.

As already outlined above, for sound interpretation and use of educational and psychological measures evidence of construct validity is essential (see, for instance, Messick, 1975, 1984, 1989, AERA, APA & NCME, 1985). IRT models can be used to describe the relation between the responses on test items on the level of latent variables. Fit to an IRT model is empirical evidence that the observed responses can be explained by some underlying structure. The latent variables of the IRT model should, of course, be an appropriate representation of the hypothesis of the test constructor with regard to the construct to be measured. For instance, the hypothesis that a unidimensional construct is measured does not comply with a multidimensional IRT model. Having a fitting IRT model may corroborate construct validity, its does not imply reliability of the test. However, in the next chapter it will be shown that, given a fitting IRT model, the reliability of a test can be computed. Further, it will be shown that in the definition of reliability, also the test objective can be taken into account.

6.7.4 Optimal test assembly

Above it was outlined that test assembly is based on the definition of a table of specifications. Also in optimal test assembly the table of specifications plays an important role. The extension here is that the items are selected in such a way that the test is optimal in some psychometric sense. Optimal test assembly can only be carried out if data from previous item administration are available. Usually, the items are calibrated via an IRT model, and this IRT model is also used for specifying the criteria of optimality. However, also procedures have been proposed where the criteria are defined in terms of classical test theory.

One of the most important features of IRT models is the fact that the characteristics of persons and items are separately parameterized, which makes it possible to describe the characteristics of a test in terms of the item parameters only. Once the item parameters are considered known trough pre-testing, the characteristics of any test or examination constructed from the item bank can be predicted. Another important feature of IRT models is that test information consists of additive and independent contributions of the items (see, for instance, Hambleton, Swaminatan & Rogers, 1991).

The fact that test information is locally evaluated as a function of ability makes it possible to construct a test, which has optimal measurement properties at a certain ability level. The choice of the ability level of interest depends on the test objective. For a test where the objective is making pass/fail decisions, it is plausible to require that the test has maximal information at the cut-off point. If, on the other hand, the interest is in selecting high ability students or low ability students, maximal information should be in the high or low region of the latent continuum, respectively.

6.7.5 Computer based testing

In its simplest form, a computer based test need not be more than a paper-and-pencil test with multiple-choice items delivered on a computer system. The items are presented one

at a time, the student keys one of the response alternatives, the computer counts the number-correct score and produces the test result. However, in computer based tests it is possible to offer a much wider variety of item formats and test contents.

Test material may include graphics, even three-dimensional and rotational graphics, split screens, for instance, for showing reading material and questions at the same time, and simulations of real-life situations. The student may be offered support in the form of an on-screen calculator, mathematical tables and reference material. By linking the computer with CD-ROM or CD-I equipment, both the quality of simulations and the quantity of reference material may be increased to a large extent.

Also the way in which the student must respond to the material can be widely varied using computer based testing. For multiple-choice items, instead of entering the character of one of the response alternatives, the student can point and click with a mouse for choosing an option. Pointing and clicking can also be used in connection with graphic displays, for instance, for items that require the student to point out parts of the body, some machine, etc. This can be extended to having a student shading, blocking and replacing areas of a graphic display. The computer mouse can also be used for having a student draw something, which is then evaluated and scored by the computer. Notice that with this last example the realm of constructed-response formats is entered. A very advanced example of this kind of constructed-response format is the test for licensing architects developed by ETS, where a complete design environment is incorporated in the test and the student has to produce an architectural design that is evaluated by the computer.

Another salient feature of computer based testing is that it offers the possibility of response-driven branching. Here the response history of the student determines the next item to be delivered. The criteria for branching may be content-based. For instance, in a diagnostic multiple-choice test, the distractors may be constructed in such a way that the wrong answers reflect specific erroneous lines of reasoning, which can be analyzed further by presenting the proper questions. However, branching need not necessarily be content-based, as will be outlined in the following section, it can also involve psychometric objectives.

6.7.6 Adaptive testing

Tests assembled using the methodology of the section on test assembly do not depend on the responses given by the student. Although they may be optimal in the sense that they meet all the specifications imposed on the test, those tests do not necessarily provide maximum information for each individual student in the population. Tailoring a test using the student's responses can be motivated by two reasons: minimizing the length of the test and maximizing information with respect to ability. In the following chapter, it will be shown that the information obtained from an item response has a maximum if the item difficulty parameter (in some sense) matches the ability parameter. So if the ability of a respondent would be known, the optimal item can be chosen from the item bank based on the relation between the ability parameter and the item parameters. This suggests the following procedure. First a (small) number of items is given to obtain an initial estimate of ability. One might chose some items, which sufficiently cover the difficulty spectrum of the content matter to be tested. Then the next item administered is the item with

maximal information at the current ability estimate. Following the response to this item, a new estimate of ability is computed using the response pattern thus far. The next item is selected as above, but with the new ability estimate, and this is repeated until the estimation error is smaller than a pre-specified tolerance.

Delivery of computer-based tests can be done using standard modules in generalpurpose item banking systems as the Examiner, the CAT-System and MicroCAT. The last two packages also support adaptive testing. Besides by general-purpose packages, these facilities are also available in packages that are purposefully designed to deliver specific tests. For instance, the Accuplacer, developed by ETS for the College Board (ETS, 1990), is developed for administering computerized placement tests. These tests cover reading comprehension, sentence skills, arithmetic skills, elementary algebra skills and college level mathematics. Besides test administration, the package also includes a placement management system, where placement advice is given using the test results, and a placement research service system for updating the placement procedure. ACT developed a comparable system called Compass (ACT, 1993).

One step further from general-purpose software are the systems developed to support complete examination organizations. A good example is the system for the Graduate Record Examinations (GRE) in America organized by ETS in combination with Sylvan Learning Systems. The GRE is a computerized and adaptive test that is administered in test centers; in 1994 ETS had 245 test centers in operation. Some of the arguments of ETS (ETS, 1994) to switch to computerized test administration are:

- It makes it possible for students to schedule tests at their convenience, rather then limiting testing to a few unmovable dates;
- Tests can be taken in a more comfortable setting and with fewer people than in large, paper-and-pencil administrations;
- Faster score reporting to the student and electronic processing of results;
- Wider range of questions and test content.

Operating this system, of course, involves more than just software for computer adaptive testing, it involves an organization and systems for scheduling, administration, accounting, identification of students, troubleshooting, reporting and handling of complaints. Summing up, computer-based test administration is implemented in various models: as general-purpose software, as medium for specific tests, and as part of an extended examination system.

7
Measurement Models in Assessment and Evaluation

7.1 Introduction

Educational assessment and evaluation can be targeted at many levels. It can be targeted at the achievements of individual students, at schools, at school districts, at school systems, and even at countries. It can serve the purpose of certification and accreditation, the purpose of diagnosis and improvement, or it can support accountability. It can take the form of an examination system, a pupil monitoring system, a school evaluation system, or a national assessment. Educational assessment and evaluation are based on several data sources, such as the data from achievement and ability tests, students' and parents' background variables, such as socio-economic status, intelligence or cultural capital, school variables, and features of the schooling system.

In this chapter, it will be shown how these various measurements can be combined and related to each other. It will be shown that item response theory (IRT) provides a useful and theoretically well-founded framework for educational measurement. It supports such activities as the construction of measurement instruments, linking and equating measurements, and evaluation of test bias and differential item functioning. Further, IRT has provides the underpinnings for item banking, optimal test construction and various flexible test administration designs, such as multiple matrix sampling, flexi-level testing and computerized adaptive testing.

In the present chapter, a number of IRT models will be introduced. The models pertain to dichotomous items (items that are either scored as correct or incorrect) and polytomous items (items with partial credit scoring, such as most types of openended questions). They can both be used for scaling itemized tests and performance assessments. It will be shown how the models are estimated and tested, how tests are scored using IRT, how items are selected given specific test purposes.

In the next chapter, a number of applications are presented, such as the use of incomplete assessment designs, equating and linking of assessments, evaluation of differences between groups, and applications to multilevel analyses as used in school effectiveness research.

7.2 Unidimensional Models for Dichotomous Items

7.2.1 Parameter separation

Item response theory (IRT) models are stochastic models for two-way data; say the responses of students to items. An essential feature of these models is parameter

separation, that is, the influences of the items and students on the responses are modeled by distinct sets of parameters. To illustrate parameter separation, consider the two-way data matrix in Table 7.1.

Table 7.1 Data Matrix with Observed Scores.

Item	1	2	3	4	5	6	
Respondent							Score
1	2	3	1				6
2	4	5	3				12
3	3	4	2				9
4				4	5	3	12
5				3	4	2	9
6				2	3	1	6
7	3	4				1	8
8	2	3				0	5

The first 3 students responded to the first 3 items, students 4, 5 and 6 responded to items 4, 5, and 6, and the last two students responded to items 1, 2 and 6. Since different respondents took different tests, their total scores cannot be compared without additional assumptions. For instance, it is unclear whether the score 9 obtained by student 3 represents the same ability level as the score 9 obtained by student 5, because they might have responded to items of a different difficulty level. However, in the present highly hypothetical case, the data were constructed according to a very simple deterministic linear model given by

$$y_{ik} = \theta_i + b_k \tag{1}$$

where y_{ik} stands for the response of student/to item k. The student parameter θ_i can be viewed as the ability of student i and the item parameter b_k can be seen as the easiness item k. The values of θ_i and b_k, and the way in which they account for the data, are shown in Table 7.2. It can now be seen that student 3 has an ability level $\theta_3=1$, while student 5 has an ability level $\theta_5=2$. We can say that the ability parameters of the students have now been calibrated on a common scale. Further, we are now in a position that we can make predictions about unobserved responses. For instance, the predicted response of student 8 on item 5 is 0+2=2.

Table 7.2 Effects of Items and Students Separated.

Item	1	2	3	4	5	6	
Respondent							θi
1	0+2	0+3	0+1				0
2	2+2	2+3	2+1				2
3	1+2	1+3	1+1				1
4				3+1	3+2	3+0	3
5				2+1	2+2	2+0	2
6				1+1	1+2	1+0	1
7	1+2	1+3				1+0	1
8	0+2	0+3				0+0	0
b_k	2	3	1	1	2	0	

Of course, in practice this kind of deterministic model never perfectly fits the data. If one were extremely strict, one could reject the model as soon as one observation was out of line, for instance, if the response of student one to item one was 3 instead of 2. However, in the social and behavioral sciences, models are always approximations of reality, and there is always an element of arbitrariness in judging the appropriateness of a model. A second problem is the estimation of the parameters when we allow some model violations. In that case, the choice of the parameters is not directly obvious and one must, for instance, resort to minimization of some loss function. However, the choice of a loss function also entails an element of arbitrariness. Adopting a stochastic model for response behavior and using a statistical framework for parameter estimation and evaluation of model fit can solve these problems. Below, it will become apparent that this does not eliminate the element of arbitrariness entirely, but statistical theory provides a well-founded framework for tackling these matters.

7.2.2 The Rasch model

In the previous section, it was shown that the principle of parameter separation could be used to calibrate the student abilities on a common scale. In this section, a stochastic model for responses of students to items will be introduced.

In this, and the following sections, the focus is on dichotomous data. A response of a student i to an item k will be coded by a stochastic variable Y_{ik}. In the sequel, upper-case characters will denote stochastic variables. The realizations will be lower case characters. In the present case, there are two possible realizations, defined by

$$y_{ik} = \begin{cases} 1 & \text{if person } i \text{ responded correctly to item } k \\ 0 & \text{if this is not the case.} \end{cases} \qquad (2)$$

Above, we considered the case where not all students responded to all items. To indicate whether a response is available, we define a variable

$$d_{ik} = \begin{cases} 1 & \text{if a response of person } i \text{ to item } k \text{ is available} \\ 0 & \text{if this is not the case.} \end{cases} \quad (3)$$

For the moment, it will be assumed that the values are a-priori fixed by some test administrator. Therefore, d_{ik} can be called a test administration variable. We will not consider d_{ik} as a stochastic variable, that is, the estimation and testing procedure will be explained conditionally on d_{ik}, that is, with d_{ik} fixed. Later, this assumption will be broadened.

In an incomplete design, the definition of the response variable Y_{ik} is generalized such that it assumes an arbitrary constant if no response is available.

An example of a data matrix is given in Table 7.3. The arbitrary constants for unobserved valued of Y_{ik} are omitted.

Table 7.3 Data Matrix with Observed Scores.

Item	1	2	3	4	5	6	7	8		
Respondent									Score	θ_i
1	0	1	1	1					3	1.65
2	0	1	1	1					3	1.65
3	0	1	1	0					2	0.37
4	1	0	0	0					1	−0.90
5					1	1	0	1	3	0.77
6					1	0	0	1	2	−0.37
7					0	0	1	1	2	−0.37
8					1	1	1	0	3	0.77
9	1	0			0	1			2	0.14
10	0	1			1	1			3	1.42
b_k	1.57	−0.09	−0.67	0.73	−0.38	−0.38	0.20	−0.98		

The simplest model, where every student is represented by one ability parameter and every item is represented by one difficulty parameter, is the 1-parameter logistic model, better known as the Rasch model (Rasch, 1960). It is abbreviated as 1PLM. It is a special case of the general logistic regression model. This also holds for the other IRT models discussed below. Therefore, it proves convenient to first define the logistic function:

$$\Psi(x) = \frac{\exp(x)}{1+\exp(x)}$$

The 1PLM is then defined as

$$p(Y_{ik} = 1 \mid \theta_i, b_k) = \Psi(\theta_i - b_k)$$

that is, the probability of a correct response is given by a logistic function with argument $\theta_i - b_k$. Note that the argument has the same linear form as in Formula (1). Using the abbreviation $P_k(\theta)=p(Y_i=1|\theta, bk)$, the two previous formulas can be combined to

$$P_k(\theta_i) = \frac{\exp(\theta_i - b_k)}{1+\exp(\theta_i - b_k)} \qquad (4)$$

The probability of a correct response as a function of ability, $P_k(\theta)$, is the so-called item response function of item k. Two examples of the associated item response curves are given in Figure 7.1. The x-axis is the ability continuum θ. For two items, with distinct values of b_k, the probability of a correct response $\Psi(\theta - b_k)$ is plotted for different values of θ. The item response curves increase with the value of θ, so this parameter can be interpreted as an ability parameter. Note that the order of the probabilities of a correct response for the two items is the same for all ability levels.

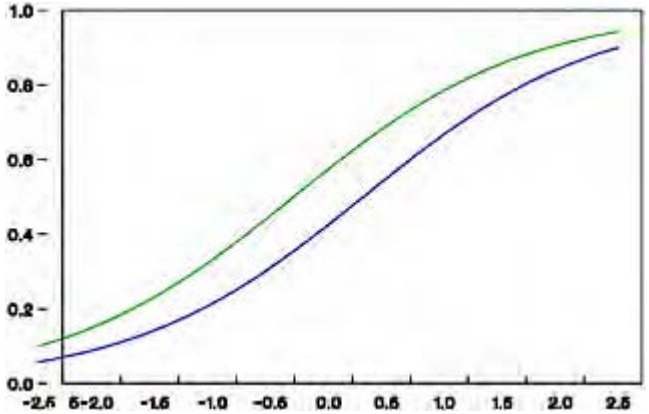


Figure 7.1 Response curves for two items in the Rasch model.

That is, the two item response curves are shifted. Further, the higher the value of b_k, the lower the probability of a correct response. So b_k can be interpreted as an item difficulty. This can also be inferred from the fact that in $\theta_i - b_k$ the item difficulty b_k is subtracted from the ability parameter θ. So the difficulty lowers the probability of a correct response.

The ability scale is a latent scale, that is, the values of θ cannot be directly observed, but must be estimated from the observed responses. The latent scale does not have a natural origin. The ensemble of curves in Figure 7.1 can be shifted across the x-axis. Or to put it differently, a constant value c can be subtracted from the ability and item parameters without consequences for the probabilities of correct responses, that is, $\Psi(\theta_i - b_k)=\Psi((\theta i - c)-(b_k - c))$. Imposing an identification restriction solves this indeterminacy of the latent scale. The scale is fixed by setting some ability or difficulty equal to some constant, say zero. One could also impose the restriction

$$\sum_{k=1}^{K} b_k = 0$$

Several estimation procedures for the ability and item parameters are available; they will be discussed below. Estimation boils down to finding values of the parameters such that the data are represented as good as possible by the model. In the example given here, a maximum likelihood estimation procedure was used. The last column of Table 7.3 gives the estimates of the ability parameters, the bottom line gives the values of the item difficulties. The estimation procedure will be outlined further below. The probabilities of correct responses can be estimated by inserting the parameter estimates in (5). The resulting values are displayed in Table 7.4. Note that also the probabilities of the unobserved responses can now be computed. With these estimates the expected scores on the test not administered can now be computed. These expectations are displayed in the last two columns of Table 7.4.

Model Fit

The distance between the responses and the expectations under the model are an indication of model fit. For instance, the response pattern of the first student, which was (0,1,1,1) can be compared with the expected response values (.52, .85, .91, .71). The closer the expectations to the observations, the better the model fit. As a practical test for model fit, this approach is far from optimal. Even for tests of moderate length and moderate sample sizes, the tables of observed and expected become quite big. Therefore, the information supplied this way is hardly informative about the nature of the model violations. This problem can, be solved by collapsing the table of frequency counts of response patterns into a smaller and more informative table. This will be returned to in sections on model fit.

Table 7.4 Data Matrix with Observed Scores.

Item	1	2	3	4	5	6	7	8	Expected Score	
Respondent									Test 1	Test 2
1	.52	.85	.91	.71	.88	.88	.81	.93	3.00	3.51
2	.52	.85	.91	.71	.88	.88	.81	.93	3.00	3.51
3	.23	.61	.74	.41	.68	.68	.54	.79	2.00	2.69
4	.08	.31	.44	.16	.37	.37	.25	.52	0.99	1.00
5	.31	.70	.81	.51	.76	.76	.64	.85	2.33	3.00
6	.13	.43	.57	.25	.50	.50	.36	.65	1.38	2.00
7	.13	.43	.57	.25	.50	.50	.36	.65	1.38	2.00
8	.31	.70	.81	.51	.76	.76	.64	.85	2.33	3.00
9	.19	.56	.69	.36	.63	.63	.49	.75		

10	.46	.82	.89	.67	.86	.86	.77	.92

7.2.3 Two- and three-parameter models

The Rasch model is derived from a number of assumptions (Fischer, 1974). One is that the number-correct scores of the students and the numbers of correct responses given to the items, defined

$$r_i = \sum_{k=1}^{K} d_{ik} y_{ik} \tag{5}$$

$$s_k = \sum_{i=1}^{N} d_{ik} y_{ik} \tag{6}$$

are sufficient statistics for unidimensional ability parameters θ_i and unidimensional item parameters b_k. That is, these statistics contain all the information necessary to estimate these parameters. With the assumption of independence between responses given the model parameters, and the assumption that the probabilities of a correct response as a function of θ_i are continuous, with the upper and lower limit going to zero and one, respectively, the Rasch model follows. One of the properties of the model is that the item response curves are shifted curves that don't intersect. This model property may not be appropriate. Firstly, the nonintersecting response curves impose a pattern on the expectations that may be insufficiently reflected in the observations, so that the model is empirically rejected because the observed responses and their expectations don't match. That is, it may be more probable that the response curves actually do cross. Secondly, on theoretical grounds, the zero lower asymptote (the fact that the probability of a correct response goes to zero for extremely low ability levels) may be a misspecification because the data are responses to multiple-choice items, so even at very low ability levels the probability of a correct response is still equal to the guessing probability.

To model these data, a more flexible response model with more parameters is needed. This is found in the 2-, and 3-parameter logistic models (2PLM and 3PLM, Birnbaum, 1968). In the 3PLM, the probability of a correct response, depends on three item parameters, a_k, b_k, and c_k, which are called the discrimination, difficulty and guessing parameter, respectively. The model is given by

$$\begin{aligned} P_k(\theta_i) &= c_k + (1-c_k) + \Psi(a_k(\theta_i - b_k)) \\ &= c_k + (1-c_k)\frac{\exp(a_k(\theta_i - b_k))}{1+\exp(a_k(\theta_i - b_k))} \end{aligned} \tag{7}$$

The 2PLM follows by setting the guessing parameter equal to zero, so upon introducing the constraint c_k=0 and the 1PLM follows upon introducing the additional constraint a_k=1.

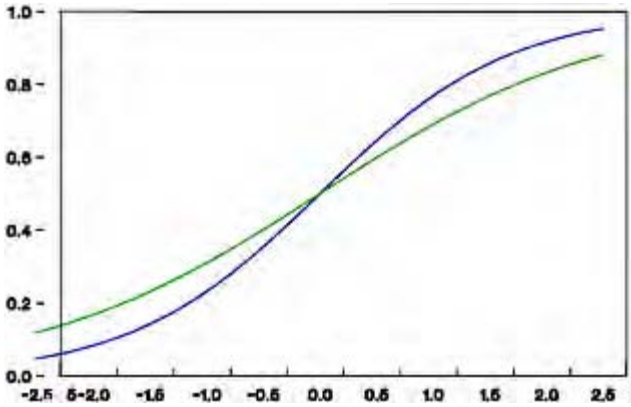


Figure 7.2 Response curves for two items in the 2PLM.

Two examples of response curves of the 2PLM are shown in the Figure 7.2. It can be seen that under the 2PLM the response curves can cross. The parameter a_k determines the steepness of the response curve: The higher a_k, the steeper the response curve. The parameter a_k is called the discrimination parameter because it indexes the dependence of the item response on the latent variable θ. This can be seen as follows. Suppose the 2PLM holds and a_k=0. Then the probability of a correct response is equal to

$$\Psi(0) = \frac{\exp(0)}{1+\exp(0)} = \frac{1}{2}$$

That is, the probability of a correct response is equal to a half for all values of the ability variable θ, so the response does not depend on θ. If, on the other hand, the discrimination parameter a_k goes to infinity, the item response curve becomes a step function: the probability of a correct response goes to zero if $\theta<b_k$ and it goes to one if $\theta>b_k$. So this item distinguishes between respondents with an ability value θ below or above the item difficulty parameter b_k. As in the 1PLM, the difficulty parameter b_k still determines the position of the response curve: if b_k increases, the response curve moves to the right and the probability of a correct response for a given ability level θ decreases, that is, the item becomes more difficult.

An item response curve for the 3PLM is given in Figure 7.3. The value of the guessing parameter was equal to 0.20, that is, c_k=0.20. As a result, the lower asymptote of the response curve goes to 0.20 instead of to zero, as in the 2PLM. So the probability of a correct response of students with a very low ability level is still equal to the guessing probability, in this case, to 0.20.

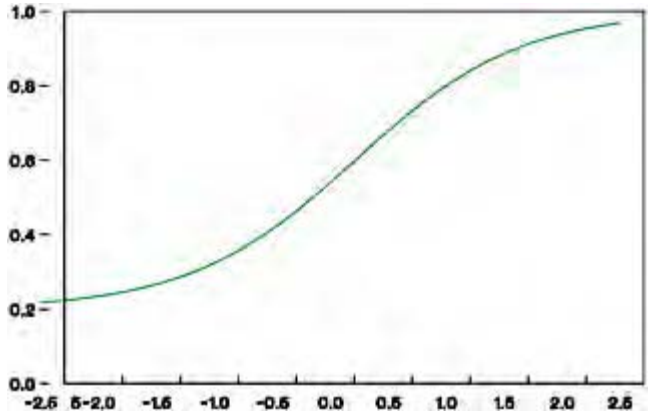


Figure 7.3 Response curve for an item in the 3PLM

Above it was mentioned that the 1PLM can be derived from a set of assumptions. On of these was the assumption that the number-correct scores given by Formula (5) are sufficient statistics for the ability parameters. Birnbaum (1968) has shown that the 2PLM can be derived from the same set of assumptions, with the difference that it is now assumed that the weighted sum score

$$r_i = \sum_{k=1}^{K} d_{ik} a_k y_{ik} \qquad (8)$$

is a sufficient statistic for ability. Note that the correct responses are now weighted with the discrimination parameters a_k. Since r_i is assumed to be a sufficient statistic, the weights a_k should be known constants. Usually, however, the weights a_k are treated as unknown parameters that must be estimated. The two approaches lead to different estimation procedures, which will be discussed in the next section.

It should be noted that the first formulations of IRT did not use the logistic function but the normal ogive function (Lawley, 1943, 1944; Lord, 1952, 1953a and 1953b). The normal ogive function $\Phi(x)$ is the probability mass under the standard normal density function left of x. With a proper transformation of the argument, $\Phi(x)=\Psi(1.7x)$, the logistic and normal ogive curves are very close, and indistinguishable for all practical work. Therefore, the 3PNO, given by

$$P_k(\theta_i) = c_k + (1-c_k) + \Phi(a_k(\theta_i - b_k)) \qquad (9)$$

is equivalent with the 3PLM for al practical purposes. The statistical framework used for parameter estimation often determines the choice between the two formulations.

The final remark of this section pertains to the choice between the 1PLM on one hand and the 2PLM and the 3PLM on the other. The 1PLM can be mathematically derived from a set of measurement desiderata. Its advocates (Rasch, 1960, Fischer, 1974, Wright & Stone, 1979) show that the model can be derived from the so-called requirement of

specific objectivity. Loosely speaking, this requirement entails invariant item ordering for all relevant subpopulations. The 2PLM and 3PLM, on the other hand, are an attempt to model the response process. Therefore, the 1PLM may play an important role in psychological research, where items can be selected to measure some theoretical construct. In educational research, however, the items and the data are given and items cannot be discarded for the sake of model fit. There, the role of the measurement expert is to find a model that is acceptable for making inferences about the students' proficiencies and to attach some measure of the reliability to these inferences. And though the 2PLM and the 3PLM are rather crude as response process models, they are flexible enough to fit most data emerging in educational testing adequately.

7.2.4 Estimation procedures

The parameters of an IRT model can be estimated by two methods: maximum likelihood estimation and Bayesian estimation.

Maximum likelihood estimation

First, we will consider the 1PLM in combination with a relatively simple estimation procedure labeled joint maximum likelihood (JML). To derive the estimation equations, we first consider the probability of a response pattern of a student *i*, denoted by the vector $y_i=(yil,\ldots, yik,\ldots y_{iK})$ given the student's ability parameter θ_i and a vector of item difficulties $b=(b_1,\ldots, bk,\ldots,b_K)$. If we assume that the responses given the parameters are independent, the product rule for independent observations can be used, and this probability is given by

$$p(\mathbf{y}_i \mid \theta_i, b) = \prod_{k=1}^{K} \left(P_k(\theta_i)^{y_{ik}} (1 - P_k(\theta_i))^{1-y_{ik}} \right)^{d_{ik}}$$

where $P_k(\theta_i)$ is the probability defined in Formula (4). The likelihood is the product of the probabilities of the students' response patterns:

$$L(\theta, b) = \prod_{i=1}^{N} p(\mathbf{y}_i \mid \theta_i, b) \qquad (10)$$

where θ stands for a vector of ability parameters $(\theta_1,\ldots, \theta_{i,\ldots}, \theta_N)$. The maximum of this likelihood is found upon solving the system of equations given by

$$r_i = \sum_{k=1}^{K} d_{ik} P_k(\theta_i), \text{ for } i = 1,\ldots,N \qquad (11)$$

$$s_k = \sum_{i=1}^{N} d_{ik} P_k(\theta_i), \text{ for } k = 1,\ldots K \qquad (12)$$

Note that this system consists of N+K equations. On the left-hand side, there are N+K observations. These observations are equated with their expected values on the right-hand side and these expected values are a function of N+K parameters. However, above it was already mentioned that the latent scale has to be identified using a restriction. So there are

only N+K-1 free parameters to estimate. On the other hand, not all equations in this system are independent, because both the summations $\Sigma_k\ s_k$ and $\Sigma_i\ r_i$ sum to the total number of correct responses given. The conclusion is that there are N+K-1 independent equations in N+K-1 free parameters. The system can be solved using a Newton-Raphson procedure (see, for instance, Molenaar, 1995).

An example of the outcome of the estimation procedure is given in the marginals of Table 7.3. The estimates of the item and student parameters given in the last row and last column of the table are the result of solving the JML equations given the data displayed in the table. Note that a score 3 on the first test reflects a higher proficiency level than the same score on the second test. Apparently, the first test was more difficult. Given these estimates, the probabilities of correct responses can now be computed using Formula (4). The results are given in Table 7.4. The last two columns give the expected number-correct scores under the model for test 1 and test 2, respectively. For the first four students, the number-correct score on the first test equals the observed number-correct score, because these expected scores emanate from solving the estimation equations given by the formulas (11) and (12). For these students the expected number-correct score on the second test is computed analogously, that is, by summing the estimated probabilities of correct scores on items on the second test. Also here it can be seen that the second test is easier, because the scores on the second test are higher than the scores on the first test.

It turns out that JML estimation is not entirely satisfactory. This is related to the fact that the number of student parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panchapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur in IRT models.

There are two maximum likelihood estimation procedures based on a likelihood function where the number of parameters does not depend on the sample size: the first one is conditional maximum likelihood (CML) estimation; the second one is marginal maximum likelihood (MML) estimation. They will be discussed in turn. CML estimation only applies to the 1PLM and some generalizations that will be shortly sketched later. The procedure is based on the fact that the likelihood on a response pattern given a value of the sufficient statistic for ability, so given ri, does no longer depend on θ_i. This is, in fact, the actual definition of sufficiency. The result will not be proved here, for a proof one is referred to Rasch (1960), Andersen (1977), Fischer (1974), or Molenaar (1995). The CML estimation equations are given by

$$s_k = \sum_{i=1}^{N} d_{ik}\, p(Y_{ik}=1 \mid r_i, b_k)\ ,\ \text{for } k = 1,\ldots,K \qquad (13)$$

where $p(Y_{ik}=1|r_1,b_k)$ is the probability of a correct response given the student's number-correct score r_i. This probability does not depend on θ. The system (13) consists of K-1 independent equations, and also here a restriction must be imposed. The estimation equations can be solved using a Newton-Raphson procedure. These equations have a structure that is analogous to the structure of the JML estimation equations in the sense that sufficient statistics for the item parameters are equated with their expected values, in this case, their expected values given the values of the sufficient statistics for the student parameters. Also in the present case the summation on the right-hand side is over the

items that were actually responded to. Of course, this procedure only produces estimates of the item parameters, and in many instances also estimates of the student parameters are needed. These parameters are estimated given the item parameter estimates; this will be returned to in the next section.

The 2PLM and 3PLM do not have sufficient statistics, and therefore, CML estimation is not feasible. An alternative and more general method for obtaining a likelihood function where the number of parameters does not depend on the sample size is by introducing the assumption that the ability parameters have some common distribution, and maximizing a likelihood that is marginalized with respect to the ability parameters. Usually, it is assumed that the ability distribution is normal with mean μ and standard deviation σ. For the 1PLM, the MML estimation equations are given by

$$s_k = \sum_{i=1}^{N} d_{ik} E\left[P_k(\theta_i)\middle|r_i\right] \qquad (14)$$

for $k=1,\ldots,K$, where $E[P_k(\theta_i)|r_i]$ is the expectation of a correct response with respect to the posterior distribution of θ_i given the number-correct score r_i. So here the observed total scores s_k are equated with their so-called posterior expectations. The item parameters are concurrently estimated with the mean and the standard deviation of the ability parameters. The estimation equations are given by

$$\mu = \frac{1}{N}\sum_{i=1}^{N} d_{ik} E\left[\theta\middle|r_i\right]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} d_{ik} E\left[\theta^2\middle|r_i\right] - \mu^2$$

so the mean and variance are equated with their respective posterior expectations. Kiefer and Wolfowitz (1956) have shown that MML estimates of structural parameters, say, the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models. The MML estimation equations for the 2PLM and 3PLM are slightly more complicated, but not essentially different, for details, refer to Bock and Aitkin (1981) and Mislevy (1984, 1986). The MML estimation procedure can also be used for estimating the parameters in the 2PNO and 3PNO.

For the 1PLM, the parameter estimates obtained using CML and MML are usually quite close. The MML estimates may be biased if the assumption about the ability distribution is grossly violated, but this rarely happens. Besides, tests of model fit are available to detect these violations, this point will be returned to later. Table 7.5 gives the CML and MML estimates for an artificial data set generated using the 1PLM. The sample size was 1000 respondents, with a standard normal ability distribution. The generating values for the item parameters are given in the fourth column. The fifth and seventh column give the CML and MML estimates, respectively. It can be seen that they are both close, and both close to the generating values. The standard errors of the estimates of the estimates are given in the sixth and eighth column. It can be verified that the generating values of the item parameters are well within the confidence regions of the estimates.

The table also gives some classical test theory indices. Cronbach's Alfa, which gives an indication of the overall reliability of the test, is given at the bottom of the table. The

columns labeled p-value and rit give the proportion correct item scores and the item-test correlation, respectively. Note that the item-test correlation, which gives an indication of the contribution of the item to the overall reliability, is highest for the items with a p-value closest to 0.50. Below it will be shown that this phenomenon is in line with conclusions from IRT. Finally, the table gives some indices of model fit, which will be discussed further below.

Table 7.5 CML and MML Estimates Compared.

			true	MML		CML		Tests of model fit		
item	p-value	rit	b	b	Se(b)	b	Se(b)	LM	df	p
1	.857	.369	−2.0	-2.084	.102	−2.002	.091	1.155	4	.885
2	.796	.418	−1.5	−1.600	.090	−1.519	.080	8.929	4	.063
3	.703	.465	−1.0	−1.022	.081	−.944	.072	2.164	5	.826
4	.614	.462	−0.5	−.554	.077	−.478	.068	5.138	5	.399
5	.512	.559	0.0	−.058	.075	.017	.067	5.322	5	.378
6	.530	.523	0.0	−.144	.076	−.069	.067	2.460	5	.783
7	.403	.522	0.5	.467	.077	.544	.068	1.985	5	.851
8	.287	.490	1.0	1.078	.082	1.157	.073	.978	5	.964
9	.252	.444	1.5	1.285	.085	1.365	.076	6.993	5	.221
10	.171	.402	2.0	1.847	.096	1.928	.086	6.051	5	.301
	Alpha=.605							LM =22.161, df=27, p=.729		

The choice between the two estimation procedures depends on various considerations. CML estimation has the great advantage that no assumptions are made about the distribution of ability. In the MML procedure, the assumption about the distribution of the ability is an integral part of the model, and a potential threat to model fit. In the section on model fit, a test for the appropriateness of the ability distribution will be described. On the other hand, MML estimation is useful when inferences about the ability distribution are the very purpose of the analyses. Examples of this will be given in the next chapter. Further, MML estimation is more general because it also applies to the 2PLM, the 3PLM, the 2PNO and the 3PNO. The major drawback of the Rasch model is that, in many instances, the model is too restrictive to fit the data, especially the assumption of identical discrimination indices for all items often finds little support. As compromise between the tractable mathematical properties of the Rasch model with and the flexibility of the 2PLM, Verhelst and Glas (1995) propose the so-called One Parameter Logistic Model (OPLM). In OPLM, difficulty parameters are estimated and discrimination indices are imputed as known constants. Therefore, the weighted sum score is a sufficient statistic for the ability parameter, and the CML estimation method as given in Formula (13) can still be used with the definition of the sufficient statistic r_i as in

Formula (8). In this way, the major advantage of CML estimation that no assumptions are made with respect to the ability distribution is preserved. In addition, Verhelst and Glas (1995) present well founded methods for formulating and testing hypotheses with respect to the magnitude of the discrimination indices. An example will be given below.

Bayesian estimation procedures

In the previous section, several maximum likelihood procedures for estimation of the parameters in an IRT model have been presented. These procedures belong to the realm of the classical so-called frequentist approach to statistical inference. This section considers an alternative approach, the so-called Bayesian approach. The motivations for the Bayesian approach are diverse. A rather mundane argument is that Bayesian confidence intervals are sometimes more realistic than frequentist confidence intervals. Another, more philosophical, argument has to do with the foundations of statistics. In frequentist approach a probability is the relative frequency of occurrence of some event in experiments repeated under exactly the same circumstances, while the Bayesian approach views probability also as a measure of subjective uncertainty. These philosophical matters, however, do not play a prominent role in the Bayesian approach to estimation in IRT, so they are beyond the scope of this chapter. There are two motives for the adoption of Bayesian approaches to IRT. The first motive has to do with the fact that item parameter estimates in the 2PLM and 3PLM are sometimes hard to obtain, because the parameters are poorly determined by the available data. This occurs because in the region of the ability scale where the respondents are located, the item response curves can be appropriately described by a large number of sets of item parameter values. To obtain "reasonable" and finite estimates, Mislevy (1986) considers a number of Bayesian approaches, entailing the introduction of prior distributions on the parameters. The approach is known under the general label of Bays modal estimation. The second motive has to do with the possibility of a Markov chain Monte Carlo (MCMC) algorithm for making Bayesian inferences. As will become clear in the sequel, more advanced IRT models give rise to complex dependency structures which require the evaluation of multiple integrals to solve the estimation equations in an MML or a Bayes modal framework. In the sequel, it will become clear that these problems are easily avoided in an MCMC framework.

In Bayesian inference, not only the data, but also the parameters are viewed as realizations of stochastic variables. This means that also the parameters have a distribution. Prior distributions can be used to express some prior belief about the distribution of parameters. So in the 1PLM, $p(b)$ and $p(\theta)$ may be the prior distributions of the item and student parameters, respectively. Bayesian inference focuses on the so-called posterior distribution, which is the distribution of the parameters given the data. So in the 1PLM, the posterior distribution of the parameters b and θ, given all response patterns, denoted by Y, is given by

$$p(b,\theta \mid Y) = \frac{p(Y \mid \theta,b)p(b)p(\theta)}{p(Y)}$$

In Bayes modal estimation (Mislevy, 1986) the main interest is in keeping the parameters from attaining extreme values by imposing priors. This can be done by two methods, in

the first method, the prior distribution is fixed, in the second approach, which is often labelled an empirical Bayes approach, the parameters of the prior distribution are estimated along with the other parameters. As in MML, the student parameters are integrated out of the likelihood. Further, point estimates are computed as the maximum of the posterior distribution (hence the name Bayes modal estimation, for more details, see, Mislevy, 1986).

The Bayes modal procedure essentially serves to keep the parameters from wandering off. However, the procedure still entails integrating out the ability parameters and for complex models this integration often becomes infeasible. To solve this problem, Albert (1992) proposed a Markov chain Monte Carlo (MCMC) procedure. In this procedure, a graphical representation of the posterior distribution of every parameter in the model is constructed by drawing from this distribution. This is done using the so-called Gibbs sampler (Gelfand & Smiths, 1990). To implement the Gibbs sampler, the parameter vector is divided in a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable posterior distributions.

If we apply this to the 1PLM we could divide the parameters into two components, say the item and ability parameters, and the Gibbs sampler would than imply that we first sample from $p(\theta|b,Y)$ and then from $p(b|\theta,Y)$ and repeat these iterations until the chain has converged, that is, when the drawn values are relatively stable and the number of draws is sufficient to produce an acceptable graph of the posterior distribution. Starting points for the MCMC procedure can be provided by the Bayes modal estimates produced described above, and the procedure first needs a number of burn-in iterations to stabilize. For more complicated models, MCMC procedures were developed by Johnson and Albert (1999) and Béguin and Glas (2001).

7.2.5 Local and global reliability

The CML and MML procedures only produce point estimates of the item parameters and, for MML, the population parameters. However, in many instances, such as in test scoring, the interest is in the student parameters, and an additional step for estimation of these parameters is needed. In this second step, the estimates of the item parameters are imputed as constants. There are two methods of estimating the students' θ-parameters: the first is a maximum likelihood (ML) estimation procedure; the second is a Bayesian procedure. They will be discussed in turn.

The ML procedure boils down to solving the ML-estimation equations given by Equation (11), imputing the item parameter estimates as constants. The standard error of these estimates can be used as an estimate of the local reliability of the test. This local reliability is derived as follows. The variance of the ML estimate of the ability parameter, denoted by $\hat{\theta}_i$, is the reciprocal of the so-called test information, that is,

$$Var(\hat{\theta}_i) = \frac{1}{I(\hat{\theta}_i)} \qquad (15)$$

where test information is defined as the sum of item information components $I_k(\hat{\theta}_i)$, so

$$I(\hat{\theta}_i) = \sum_{k=1}^{K} I_k(\hat{\theta}_i) \qquad (16)$$

Up to now, we have only swapped one definition for the next one, but these definitions have a background. For every student *i*, the solution of the estimation equation given in (11) is the maximum of the likelihood function $L(\theta, b)$ with respect to the variable θ_i. This maximum is found upon setting the first-order derivatives of the logarithm of this function with respect to θ_i equal to zero. So the equation is given by $dlogL(\theta, b)/d\theta_i=0$. The variance of the estimate is found by taking the second-order derivative, taking the opposite, and inserting the estimate $\hat{\theta}_i$. That is, the variance $Var(\hat{\theta}_i)$ is equal to $-d^2 \log L(\theta,b)/d\theta_i^2$ evaluated at $\hat{\theta}_i$. If we would work out this second-order derivative, we would see that the local independence between the item responses resulted in a summation of item information components, as is Formula (16), and, for the 1PLM, the item components are given by the item information function

$$I_k(\theta_i) = P_k(\theta_i)(1-P_k(\theta_i)) \qquad (17)$$

Note that the right-hand side has the same form as the variance of a Bernoullivariable. So due to the independence of student and item responses given the parameters, the IRT model is nothing else than a model for *N* times *K* Bernoullitrials, with the peculiarity that the probability of success differs from trail to trail. The fact that the test information is additive in item information is very convenient: every item has a unique and independent contribution to the total reliability of the test. Further, item information is always positive, so we can infer that the test information increases with the number of items in the test, and as a consequence, the standard error decreases if the number of items goes up. Further, we can now meaningfully define the optimal item for the measurement of a student's ability parameter θ. Suppose that the 1PLM holds, that is, the probability of a correct response $P_k(\theta)$ is given by Formula (4). If we enter this definition into the definition of item information given by Formula (17), it can be verified that item information is maximal if $\theta=b_k$, that is, if the item difficulty matches the ability parameter (This can be verified by taking the first-order derivative of the logarithm of item information with respect to θ and equating to zero). At the point $\theta=b_k$, the probability of a correct response is equal to a half, that is, $P_k(\theta)=0.5$. Intuitively, this result is quite plausible. If students are administered items which are far too difficult, hardly any correct response will be given and we hardly learn anything about their ability level. The same holds for the opposite: if the items are far too easy, all responses are will be correct. We have the maximum a-priori uncertainty if the odds of a correct response are even, and in that case we learn most when we observe the actual outcome.

For the 2PLM, item information is given by

$$I_k(\theta_i) = a_k^2 \; P_k(\theta_i)(1-P_k(\theta_i))$$

Note that item information in the 2PLM has the same form as in the 1PLM, with the exception of the presence of the square of the item discrimination parameter a_k. As a

result, item information is positively related to the magnitude of the discrimination parameter. For the 3PLM, item information is given by

$$I_k(\theta_i) = \frac{(1-c_k)^2 a_k^2 \ \psi_k^2(\theta_i)(1-\psi_k(\theta_i))^2}{P_i(\theta_n)(1-P_i(\theta_n))}$$

Note that if $c_k=0.0$ and $a_k=1.0$, item information is the same as for the 1PLM, as it should be. Further, item information is maximal if the guessing parameter is equal to zero, that is, if $c_k=0.0$, and item information decreases with the guessing parameter.

In a Bayesian framework, inferences about a student's ability parameter are based on the posterior distribution of θ. For the 1PLM, the number-correct score r_i is a sufficient statistic for the ability parameter, so it makes no difference whether we use the posterior distribution given r_i, given by $p(\theta| \ r_i,b,\mu,\sigma)$, or the posterior distribution given the complete response pattern y_i, say $p(\theta|y_i, b, \mu, \sigma)$. In the 2PLM and 3PLM, the number-correct score is not a sufficient statistic, so here we generally use the second posterior distribution. A point estimate of the student's ability can for instance be computed as the posterior expectation, the so-called EAP estimate, given by

$$E\left[\theta \middle| y_i, b, \mu, \sigma\right]$$

Of course, also the posterior mean or mode can be used as a point estimate for ability. The posterior variance given by

$$Var(\theta \mid y_i) \quad = \quad E\left[\theta^2 \middle| y_i, b, \mu, \sigma\right] - \left\{E\left[\theta^2 \middle| y_i, b, \mu, \sigma\right]\right\}^2$$

serves as indication of the local reliability. Computation of these indices needs estimates of the item parameters b and population parameters μ and σ. In a likelihood-based framework as MML, point estimates obtained in the MML estimation procedure can be imputed as constants. In a fully Bayesian framework posterior distributions are the very outcome of the estimation procedure, and the posterior indices of central tendency and the posterior variance can be easily computed from the draws generated in the MCMC procedure.

ML and EAP estimates of ability are generally not comparable. Consider the example of Table 7.6. The example pertains to the 1PLM, the same data and item parameter estimates as in Table 7.5 were used.

Table 7.6 Estimates of Latent Abilities.

Score	Freq	WML	SE	EAP	SD
0	9	−3.757	1.965	−1.809	.651
1	28	−2.416	1.056	−1.402	.626
2	70	−1.647	.860	−1.022	.608
3	130	−1.039	.779	−.661	.595
4	154	−.502	.741	−.311	.588
5	183	.007	.730	.033	.586

6	157	.515	.740	.378	.589
7	121	1.049	.775	.728	.596
8	97	1.648	.855	1.091	.609
9	39	2.403	1.049	1.472	.628
10	12	3.728	1.953	1.882	.654

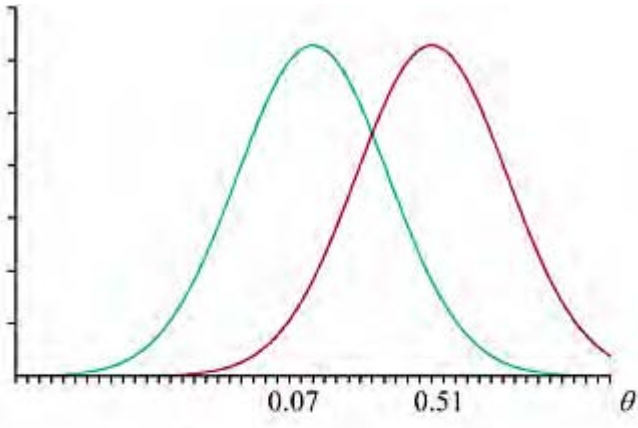
For all possible scores on a test of 10 items, the table gives the frequency, an ML estimate, its standard error, the EAP estimate and the posterior standard deviation. ML estimates for the zero and perfect score do not exist, and further, the ML estimates are slightly biased. Therefore, the estimates in Table 7.6 are computed using a weighted maximum likelihood (WML) procedure where a weighting function is added to the left-hand side of the estimation equations given by Formula (11), such that the estimators are unbiased to the order K^{-1} (Warm, 1989). WML also gives estimates for the zero and perfect score, though it can be seen in Table 7.2 that the standard errors of these estimates are quite large. Comparing the WML and EAP estimates, it can be seen that the EAP estimates are shrunken towards zero. From the frequency distribution, it can be inferred that the mean of the ability distribution might be zero, and this is in fact true, because the latent scale was identified by imposing the restriction $\mu=0$. In Bayesian statistics, this shrinking phenomenon is known as shrinking towards the mean. This phenomenon is caused by the fact the Bayesian estimates are often a compromise between the estimate imposed by the likelihood alone and the influence of the prior. If parameters have some common distribution, and the likelihood gives little information about the values of individual parameters, one might say that the estimates of these parameters borrow strength from each other via their common distribution. The result is that they are shrunken towards their mutual mean. Therefore, one might suggest that the WML estimates are preferable, unless information is sparse and a prior distribution must support the estimates.

Both the WML and EAP estimates have an asymptotic normal distribution. This can be used to determine confidence or credibility regions around the estimates. The assumption of normality is acceptable for most of the score range of a test longer than 5 items, except for the extreme scores, say the three highest and lowest scores. Consider the ability estimate associated with a number-correct score of 5. The ability estimate is 0.07 and its standard error is 0.730. For this short test of 10 items, this means that a 95% confidence region would range from −1.4301 to 1.4315. This means that the ability estimates associated with all scores from 3 to 7 are well within this confidence region. A 95% confidence region build using the posterior standard deviation ant the assumption of normality also includes the scores 2 and 8.

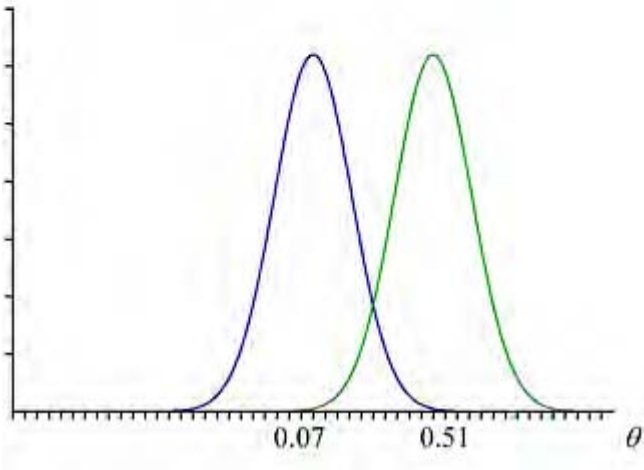
If this kind of reasoning is applied to assess to what extent scores on a test can be distinguished, it is reasonable to take the uncertainty of both scores into account. Consider the scores 5 and 6. Based on the WML estimates, they have normal distributions shown in the first panel of Figure 7.4. The means are 0.07 and 0.515, respectively, and the standard deviations are 0.730 and 0.740, respectively. The conclusion is that these two distributions greatly overlap. If we would assume that the two distributions represent distributions of ability values, the probability that a draw from

the first one would exceed a draw from the second on would be 0.31. This probability is on a scale ranging from a probability of 0.50 for complete unreliability (the distributions are identical) to a probability of 0.00 for complete reliability (the distributions are infinitely far apart). If three tests identical to the original test of 10 items could be added, that is, the test length would be 40, the variance would be divided by 4. This situation is depicted in the second panel of Figure 7.4. As a consequence, the probability that a random draw from the distribution associated with a score 5 would exceed a random draw from the distribution associated with a score 6.

The below calculus of local reliability was done for the 1PLM, where the numbercorrect score is a sufficient statistic for the ability parameter. In the 2PLM and 3PLM, the ability estimate depends on the weighted sum score or the complete response pattern, respectively. However, tests are usually scored using numbercorrect scores. As note above, number correct scores and weighted scores are usually highly correlated and the loss of precision is usually limited. To index the precision of a test following the 2PLM or 3PLM scores with number-correct scores, the same logic as above can be used in a combination with the Bayesian framework. That is, one can compute the posterior expectation $E(\theta|r,a,b,c)$ and variance $E(\theta|r,a,b,c)$ conditional on the number correct scores r, but with the 2PLM or 3PLM as a response model.



Panel 1: Test with 10 items



Panel 2: Test with 40 items

Figure 7.4 Confidence intervals for ability estimates.

Local reliability plays an important role in educational measurement in activities as standard setting. For instance, if a certain number-correct score serves as a cut-off score on some examination, the dispersions of ability estimates round the scores can serve as an indication of the number of misclassifications made. However, in many other instances, it is also useful the have an indication of global reliability. This index is based on the following identity.

$$Var(\theta) = E\left[Var(\theta|r)\right] + Var\left[E(\theta|r)\right] .$$

The identity entails that the total variance of the ability parameters is a sum of two components. The first component, $E[Var(\theta|r)]$, relates to the uncertainty about the ability parameter. Above, it was shown that the posterior variance of ability, $Var(\theta|r)$, gives an indication of the uncertainty with respect to the ability parameter, once we have observed the score r. By considering its expectation over the score distribution, we obtain an estimate of the average uncertainty over the respondents' ability parameters. The second term, $Var[E(\theta|r)]$, is related to the systematic measurement component. The expectation $E(\theta|r)$ serves as an estimate of ability, and by considering the variance of these expectations over the score distribution, we get an indication of the extent to which the respondents can be distinguished on the basis of the test scores. Therefore, a reliability index taking values between zero and one can be computed as the ratio of the systematic variance and the total variance, that is

$$\rho = \frac{Var\left[E(\theta|r)\right]}{Var(\theta)}$$

For the example of Table 7.6 the global reliability was equal to 0.63.

Optimal Test Assembly

Theunissen (1985) has pointed out that test assembly problems can be solved by binary programming, a special branch of linear programming. Binary programming problems have two ingredients: an objective function and one or more restrictions. This will first be illustrated by a very simple test assembly problem. Consider a test where the interest is primarily in a point θ_0 of the ability continuum and one wants to construct a test of L items that has maximal information at θ_0. Selection of items is represented by a selection variable d_k,$d_k=1$ if item k is selected for the test and $d_k=0$ if this is not the case. The objective function is given by

$$\sum_{k=1}^{K} d_k I_k(\theta_0)$$

which should be maximized as a function of the item selection variables d_k. This objective function must be maximized under the restriction that the test length is L, which translates to the restriction

$$\sum_{k=1}^{K} d_k = L$$

If the item is selected, $d_k=1$ so in that case the item contributes to the target number of items L and the total test information.

This basic optimization problem can be generalized in various directions, such as choosing more points on the ability scale, constructing more tests simultaneously, introducing time and cost constraints, and taking the constraints of the table of test specifications into account. More on optimal test assembly problems and the algorithms to solve these problems can be found in Boekkooi-Timminga (1987, 1989, 1990), van der Linden and Boekkooi-Timminga (1988, 1989) and Adema and van der Linden (1989). In

this section, some of the possibilities of optimal test assembly will be illustrated using an application by Glas (1997).

The problem is assembling tests for pass/fail decisions, where the tests have both the same observed cut-off score and the same cut-off point on the ability scale and are approximately equally reliable at the cut-off point. It is assumed that the Rasch model for binary items holds for all items in the item bank. For this problem the item selection variables are d_{ti} for $t=1,\ldots, T$ and $i=1,\ldots, K$, where again d_{ti} is equal to one if item i is selected for test t and zero otherwise. The binary programming that must be solved in the variables d_{tk} for $t=1,\ldots, T$ and is maximizing

$$\sum_{t=1}^{T}\sum_{k=1}^{K} d_{tk}\, I_k(\theta_0)$$

subject to

$$\sum_{t=1}^{T} d_{tk} \leq 1,\ k = 1,\ldots,K \tag{18}$$

$$-c_1 \leq \left\{\sum_{k=1}^{K} d_{tk} I_k(\theta_0)\right\} - \left\{\sum_{k=1}^{K} d_{tr} I_r(\theta_0)\right\} \leq c_1,\ t,r = 1,\ldots,T, t \neq r \tag{19}$$

$$-c_2 \leq \left\{s_0 - \sum_{i=1}^{K} d_{tk} P_k(\theta_0)\right\} \leq c_2,\ t = 1,\ldots,T \tag{20}$$

Inequality (18) imposes the restriction that every item should be present in one test only. Inequality (19) must assure that the difference in information of the two tests at the cut-off point is small. This is accomplished by choosing c_1 as a sufficiently small constant. The inequalities (20) must assure that the observed and latent cutoff scores of the two tests are sufficiently close. Also here this is accomplished by choosing c_2 as a sufficiently small constant. If the restrictions imposed by (18)–(20) are too tight, a solution to the optimization problem may, of course, not exist. For instance, it is not possible to construct two unique tests of 50 items from an item bank of 70 items. Besides the size of the item bank, also the distribution of item parameters, the distribution of items on the classification variables and the magnitude of c_1 and c_2 determine the feasibility of a solution.

As already noted, the test assembly problem discussed here is just one of the large variety of problems addressed in the literature. Test assembly procedures have been developed for minimizing test length under the restriction that test information is above a certain target for a number of specified points on the ability scale, minimization of administration time with a fixed number of items in a test, maximization of classical reliability, and optimal matching of a target for the observed score distribution of a test. Tables with objective functions and constraints covering most of the options available can be found in Van der Linden and Boekkooi-Timminga (1989).

7.2.6 Model fit

IRT models have specific properties that turn out to be useful in educational assessment. Examples are parameter separation, which is the basis for item banking, test equating and linking of educational assessments, and a local definition of reliability that supports optimal item selection and test construction. However, the utilization of these properties is only admissible if the model holds. Since IRT models are based on a number of explicit assumptions, methods have been developed that are focused on these assumptions. The first set of assumptions includes the form of the item response curves and (for the Rasch model) the sufficiency of the number-correct scores; the second set of assumptions includes local stochastic independence and unidimensionality. Though the tests are targeted at one specific assumption, it should be noted that the assumptions are related, and the specificity of the tests must not be exaggerated. Test statistics are computed using parameters under the null model, under the assumption that the IRT model holds, and violation of some assumption, say unidimensionality, will bias the statistic targeted at some other assumption, say a test statistic targeted at the form of the (unidimensional) item response curve.

The tests for the two sets of assumptions can be computed from two perspectives: the items and the respondents. In the first case, for every item an item fit statistic is computed to assess whether the item violates the model, in the second case person fit statistics are computed to assess whether the student responds according to the model. Both classes of tests play related, but slightly different roles. For instance, tests of item fit may be computed when calibrating a test battery, and a proper sample of the target population, giving proper responses, is available. Items that violate the IRT model can then, in principle, be removed from the test. In some instances, however, this may treated the validity in such a degree, that misfitting items cannot be removed. In that case, one may attempt to model response behavior with a more sophisticated IRT model, say a model with multidimensional ability parameters, or a model that allows for local dependence of responses. These models will be returned to in the sequel. If part of the sample does not give proper responses to the items, for instance because the respondents are not motivated and give a lot of guessed responses, then person fit tests may be used to identify these respondents. In that case, fitting the model becomes a process that iterates between estimation followed by evaluation of item fit and estimation followed by evaluation of person fit. The problem here is that person fit statistics are computed using item parameter estimates, and these will be biased in the presence of misfitting respondents. Fortunately, research shows that the bias in item parameter estimates remains limited with up to 20% misfitting students (Hendrawan, Glas & Meijer, 2001). Another problem is that removing respondents from the calibration sample may threaten the validity of the calibration results.

Testing the form of the item characteristic curve

Ideally, a test of the fit of the item response curve would be based on the assessment whether the responses given would match the response curve. However, firstly, the true

values of the respondents' ability values θ are not available. Secondly, if they would be available only a limited number of θ-values would be available. And thirdly, for every available θ-value, only one observed response is available. So we cannot accumulate responses to obtain sufficiently large values for the number of observed and expected responses to assess the (asymptotic) distribution of possible test statistics. The first point could be solved using estimates of θ, but in the CML and MML framework does not directly provide these estimates and item fit statistics in this framework using estimates outside this framework have theoretical problems. The fact that the 1PLM has sufficient statistics for the ability parameter, however, suggest an alternative approach. Respondents with the same number-correct score will probably be quite homogeneous with respect to their ability level (more precisely, the number-correct score has a monotone likelihood ratio in θ). In an MML framework, the 2PLM and 3PLM do not have sufficient statistics for θ, but usually, the number-correct score and the estimate of θ highly correlate. Therefore, a test of the item response curves can be based on the difference between the observed proportion of a correct response given the sum score and the analogous probability. This leads to a test statistic

$$Q_1 = \sum_{g=1}^{G} N_g \frac{(O_{gk} - E_{gk})^2}{E_{gk}(1 - E_{gk})} \qquad (21)$$

where O_{gk} and E_{gk} are the observed and expected proportion of respondents in subgroup g with a correct score on item k, respectively. Usually, the groups are groups with the same number-correct score, but there are alternatives, which will be returned to later. The way in which the expected value E_{gk} is computed differs depending on the estimation procedure. First, a likelihood-based framework will be treated, and then we will discuss the Bayesian approach.

Orlando and Thissen (2000) give the formulas needed to compute E_{gk} in an MML framework. The test statistic can also be computed in a CML framework (van den Wollenberg, 1982). In older versions of the test (Yen, 1981, 1984; Mislevy & Bock, 1990) subgroups were formed on the basis of their ability estimates rather than on the basis of their total score, and the expectation E_{gk} was computed as the mean predicted probability of a correct response in subgroup g. However, this approach (which is still prominent in some software packages) does not lead to a statistic with a tractable distribution under the null-hypothesis and to acceptable power characteristics (Glas & Suárez-Falćon, 2003). The reason is that in this case the grouping of respondents is not based on some directly observable statistic, such as the number-correct score, and, therefore, the observed frequencies were not solely a function of the data, but also of model-based trait estimates, which violates the assumptions of the traditional χ^2 - goodness-of-fit-test.

Also the approach used in the Q_1-test as defined by Orlando and Thissen (2000) has a disadvantage, because here a tractable distribution under the null-hypothesis and acceptable power characteristics can only be achieved when the test is computed with the sample of respondents partitioned on the basis of their number-correct scores (Glas & Suárez-Falćon, 2003). Especially for long tests, it would be practical when a number of adjacent scores could be combined, so that the total number of groups G would remain limited, say limited to 4 to 6 groups. The problem that score-groups cannot be combined

is caused by the fact that both the dependency between the observed proportions O_{gk} and their expectations E_{gk} caused by the parameter estimation are not taken into account. To obtain a test statistic where score-groups can be combined and where the asymptotical distribution can be derived analytically the complete covariance matrix of the differences O_{gk} -E_{gk} has to be taken into account. These problems are solved in the framework of the so-called Lagrange multiplier (LM) statistic (Rao, 1947; Aitchison & Silvey, 1958). The application to IRT is labeled the LM-Q1-test (Glas, 1988; Glas & Verhelst, 1995; Glas, 1998, 1999).

An example

In the following artificial example shows the some major steps in a fit analysis. In this example, responses of 1000 respondents to 10 items were generated according to the 2PLM. As in the previous example, the item difficulty parameters b_k were equal to −2.0(0.5)2.0, with the middle value 0.0 appearing twice. The item discrimination parameters *ak* were all equal to one, except for the items 5 and 6, which had a discrimination parameter equal to 0.5 and 2.0, respectively. Ability parameters were drawn from a standard normal distribution. First, the data were analyzed with the 1PLM. The item parameter estimates and their standard errors are given in Table 3, under the heading “Model 1”. It can be seen that the estimates of the difficulty parameters are sufficiently close to the true generating values, even though the items 5 and 6 violated the model. All results in Table 7.3 were obtained using the computer program OPLM (Verhelst, Glas & Verstralen, 1995). For every item, the LM-Q1-statistic was computed. The results are again shown under the heading “Model 1”. The values of the computed statistics are given under the label “LM-Q1”, the degrees of freedom and the significance probabilities are given under the labels “df” and “p”, respectively. It can be seen that the test was significant for the items 5 and 6. That is, the difference between the observed and expected proportions of correct responses in the score groups were such that the hypothesis that the observed proportions were properly described by the 1PLM had to be rejected.

Table 7.7 CML Estimates and Model fit.

			Model 1		
Item	b	SE(b)	LM-Q1	df	p
1	−1.97	.090	3.33	5	.649
2	−1.62	.082	4.18	5	.523
3	−1.02	.074	5.03	5	.412
4	−.51	.070	4.81	6	.567
5	−.05	.068	29.08	6	.000
6	−.09	.068	41.52	6	.000
7	.35	.069	4.06	6	.669
8	1.13	.074	5.38	5	.371

9	1.80	.084	5.12	5	.401
10	1.99	.088	7.88	5	.163
LM-Q1=71.73 df=27 p=.000					

			Model 2		
Item	b	SE(b)	LM-Q1	df	p
1	−1.99	.089	2.15	3	.541
2	−1.64	.082	.63	3	.889
3	−1.04	.074	3.58	4	.465
4	−.53	.070	3.28	4	.511
5	– –	– – –	– – –	–	– – –
6	– –	– – –	– – –	–	– – –
7	33	.069	.47	4	.976
8	1.11	.074	2.47	4	.649
9	1.78	.085	1.91	3	.589
10	1.98	.089	4.06	3	.255
LM-Q1=15.53 df=21 p= 796					

			Model 3			
Item	a	b	SE(b)	LM-Q1	df	p
1	4	−.50	.030	2.70	5	.746
2	4	−.41	.029	6.27	6	.394
3	4	−.25	.027	3.05	6	.802
4	3	−.21	.088	5.50	3	.138
5	9	.02	.014	.87	3	.831
6	1	.05	.187	2.87	5	.719
7	4	.08	.026	5.02	6	.541
8	5	.27	.026	.51	4	.972
9	4	.44	.029	1.55	5	.907
10	4	.49	.030	6.40	5	.269
LM-Q1=33.24 df=25 p=.125						

The presence of the two misfitting items did not interfere with the fit of the eight other items: all tests for these items were not significant. The bottom line gives the outcome of a version of the test were the item response curve of all items are evaluated simultaneously. It can be seen that this test of global model fit rejected the model.

Three routes are open to obtaining a fitting model for this data set: removing the misfitting items, or using the OPLM or 2PLM to model the data. The results of removing the two items are shown in Table 7.3 under the heading “Model 2”. The tests for the eight remaining items are not significant and also the global test statistic, shown at the bottom of the table, no longer rejects the model. An alternative to removing items is trying to model response behavior. Here two approaches were used: using the OPLM model or using the 2PLM. The first approach entails defining integer-valued discrimination indices that are imputed as constants in a CML estimation procedure. Initial estimates of these discrimination indices are found by estimating these indices using the 2PLM and then rounding to integer values. One might ask why the 2PLM is then not used straightaway. The reason is that both approaches have their advantages and disadvantages. The 2PLM is more flexible than the OPLM (though Verhelst & Glas, 1995, show that coverage of the parameter space using integer values for the discrimination parameters is quite compact) but it needs an assumption on the distribution of the ability for obtaining MML estimates. For the OPLM, CML estimation is feasible, and this estimation procedure needs no assumption on the distribution of ability. So both approaches have their merits and drawbacks. The estimates and the outcome of the fit tests are shown in Table 7.7 under the heading “Model 3”. The discrimination parameters are given under the label “a”. The generating values of the discrimination parameters of the item 1–4 and 7–10 were all equal to one. In the present analyses they are scaled to approximately 4. The original generating value for item 5 was half that of the discrimination parameters of the item 1-4 and 7-10, the value for items 6 was twice that value. Note that these ratios are still reflected in the values as they are given in Table 7.7. Further, the parameterization of the model is as in Formula (9), so the item difficulty parameters should be multiplied with the discrimination indices to make them comparable to the item parameters in the 1PLM.

Both the outcomes of the item oriented tests and the global test now support the OPLM model. In the present example, the data fit the OPLM right away. In practical situations, however, finding a set of discrimination parameters a_k to obtain model fit is usually an iterative process. Discrimination parameters a_k of misfitting items can be adjusted using the pattern of differences between observed en expected proportions of correct responses, increasing a_k when these differences suggest that the item response curve under the OPLM must be steeper, and decreasing a_k when these differences suggest the opposite. This iterative process is repeated until the proportion of misfitting items falls below the nominal significance level and the global test is no longer significant.

Table 7.8 MML Estimates and Model fit.

item	a	Se(a)	b	Se(b)	LMQ1	p	df
1	1.01	.135	−1.99	.127	0.67	.88	5
2	1.05	.130	−1.66	.113	5.97	.11	4
3	1.16	.123	−1.10	.097	2.88	.41	6
4	0.91	.104	−0.53	.079	4.57	.21	6
5	1.98	.229	−0.12	.106	2.59	.46	4

6	0.54	.079	−0.12	.069	1.18	.76	6
7	1.07	.116	0.30	.081	1.07	.78	6
8	1.27	.141	1.15	.104	2.33	.51	6
9	1.04	.128	1.73	.114	2.87	.41	5
10	1.09	.143	1.95	.129	10.15	.02	5

LM-Q1=33.24 df=25 p=.125

The other approach is applying the 2PLM in combination with MML estimation. Table 7.8 gives the MML estimates of the item parameters computed using the program Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). Note that the estimates reflect the true generating values very well. The fit statistics computed by Bilog-MG are not used here; they have the problems with their Type I error rate and power already mentioned above. Instead, the LM-Q1 statistics were computed using dedicated software. It can be seen that all test supported the 2PLM. The LM-Q1 item oriented tests were computed using four score groups. At the bottom of the table, the outcome of a global version of the LM-Q1-test is displayed. It can be seen that the model is supported, which is, of course, as expected.

Testing local independence and multidimensionality

The statistics of the previous section can be used for testing whether the data support the form of the item response functions. Another assumption underlying the IRT models presented above is unidimensionality. Suppose unidimensionality is violated. If the student's position on one latent trait is fixed, the assumption of local stochastic independence requires that the association between the items vanishes. In the case of more than one dimension, however, the student's position in the latent space is not sufficiently described by one unidimensional ability parameter and, as a consequence, the association between the responses to the items given this one ability parameter will not vanish. Therefore tests for unidimensionality are based on the association between the items.

Yen (1984, 1993) proposed a test statistic, which is based on the argument that the random error scores on the items k and l, defined by $d_k=y_k-P_k(\theta)$ and $d_l=y_l-P_l(\theta)$ are approximately bivariate normally distributed with a zero correlation. The test statistic was equal to the correlation (taken over respondents) of d_k and d_l, that is,

$$Q_{3kl} = \rho(d_k, d_l)$$

where $P_k(\theta)$ and $P_l(\theta)$ are evaluated using the EAP estimate of θ. If the model holds, the Fisher r-to-z transform of this statistic may have a normal distribution with a zero mean and a variance equal to 1/(N-3). Simulation studies reported by Yen (1984, 1993) showed that this approximation produces quite acceptable results. In the framework of the 1PLM and CML estimation, van den Wollenberg (1982) showed that violation of local independence can be tested using a test statistic based on evaluation of the association between items in a 2-by-2 table. Applying this idea to the 3PLM in an MML framework,

a statistic can be based on the difference between observed and expected frequencies given by

$$d_{kl} = n_{kl} - E(N_{kl})$$

$$= n_{kl} - \sum_{r=2}^{K-1} n_r P(Y_k = 1, Y_l = 1 \mid R = r)$$

where n_{kl} is the observed number of respondents making item *k* and item *l* correct, in the group of respondents obtaining a score between 2 and K-2, and $E(N_{kl})$ is its expectation. Only scores between 2 and K-2 are considered, because respondents with a score less than 2 cannot make both items correct, and respondents with a score greater than K-2 cannot make both items incorrect. So these respondents contribute no information to the 2-by-2 table. Using Pearson's X^2 statistic for association in a 2-by-2 table results in

$$S_{3kl} = \frac{d_{kl}^2}{E(N_{kl})} + \frac{d_{kl}^2}{E(N_{k\bar{l}})} + \frac{d_{kl}^2}{E(N_{\bar{k}l})} + \frac{d_{kl}^2}{E(N_{\bar{k}\bar{l}})}$$

where $E(N_{k\bar{l}})$ is the expectation of making item *k* correct and *l* wrong, and $E(N_{\bar{k}l})$ and $E(N_{\bar{k}\bar{l}})$ are defined analogously. Glas and Suárez-Falćon (2003) describe a version of the statistic for the MML framework, and Glas (1999) proposes a Lagrange-multiplier version of the test labeled LM-Q2.

An example

An artificial example was generated using the 2PLM with the same item parameters as in the previous example. However, local independence between the item 5 and 6 was violated: if a correct response was given to item 5, the difficulty parameter of item 6 was decreased by 0.50. Parameters were estimated by MML, and the LM-Q2-statistic was computed to verify whether the misfit could be detected. The results are shown in Table 7.9. Not that the model violation did not result in a substantial bias estimates of the item parameters. Further, only the LM-Q2-test for the pair item 5 and item 6 was significant, the other pairs were not affected. The observed and expected values on which the test was based are shown in the last two columns.

Table 7.9 Testing for Local Dependence.

Item	a	Se(a)	b	Se(b)	Item	LM-Q2	p	Obser	Expect
1	0.89	.14	−1.93	.12		– –	– –	– –	– –
2	1.00	.13	−1.43	.10	1	0.12	0.73	670	665.6
3	0.70	.11	−0.81	.07	2	0.89	0.35	555	541.7
4	0.78	.10	−0.43	.07	3	0.01	0.91	424	425.5
5	1.28	.14	−0.07	.08	4	0.01	0.92	346	347.2

6	1.36	.16	−0.49	.09	5	5.00	0.03	392	365.3
7	0.93	.11	0.44	.07	6	0.38	0.54	280	287.1
8	0.86	.12	0.86	.08	7	0.08	0.77	166	163.0
9	0.81	.12	1.36	.09	8	0.00	0.96	97	96.6
10	0.60	.13	1.76	.10	9	0.13	0.72	46	48.2

Testing differential item functioning

Differential item functioning (DIF) is a difference in item responses between equally proficient members of two or more groups. For instance, a dichotomous item is subject to DIF if, conditionally on ability level, the probability of a correct response differs between groups. One might think of a test of foreign language comprehension, where items referring to football might impede girls. The poor performance of the girls on the football-related items must not be attributed to their low ability level but to their lack of knowledge of football. Since DIF is highly undesirable in fair testing, methods for the detection of DIF are extensively studied. Overviews are provided by the books by Holland and Wainer (1993) and by Camilli and Shepard (1994). Several techniques for detection of DIF have been proposed. Most of them are based on evaluation of differences in response probabilities between groups conditional on some measure of ability. The most generally used technique is based on the Mantel-Haenszel statistic (Holland & Thayer, 1988), others are based on log-linear models (Kok, Mellenbergh & Van der Flier, 1985; Swaminathan & Rogers, 1990), and on IRT models (Hambleton & Rogers, 1989; Kelderman, 1989; Thissen, Steinberg & Wainer, 1993; Glas & Verhelst, 1995; Glas, 1998).

In the Mantel-Haenszel (MH) approach, the respondent's number-correct score is used as a proxy for ability, and DIF is evaluated by testing whether the response probabilities differ between the score groups. Though the MH test works quite well in practice, Fischer (1995) points out that its application is based on the assumption that the 1PLM holds. In application of the MH test in other cases, such as with data following the 2PLM or the 3PLM, the number-correct score is no longer the optimal ability measure. In an IRT model, ability is represented by a latent variable θ, and an obvious solution to the problem is to evaluate whether the same item parameters apply in subgroups that are homogeneous with respect to θ. As for the previous model violations, also DIF can be assessed with a Lagrange multiplier statistic (Glas, 1998).

Table 7.10 Parameter Generating Values, Estimates and the LM Statistic.

Item	b_k	$\hat{b}_k$	$se(\hat{b}_k)$	LM	Pr
1	−1.00	−0.91	0.12	1.33	0.25
2	0.00	0.13	0.11	0.27	0.61
3	1.00	1.13	0.12	1.14	0.29
4	−1.00	−0.93	0.11	1.14	0.29
5	0.0/0.5	0.41	0.11	18.03	0.00
6	1.00	1.04	0.12	0.02	0.90
7	−1.00	−0.77	0.12	0.05	0.83
8	0.00	0.11	0.11	0.01	0.92
9	1.00	1.03	0.11	0.11	0.74
Pop	μ	$\hat{\mu}$	$se(\hat{\mu})$		
1	1.00	1.00	0.11		
2	0.00	0.00	– – –		
Pop	σ_g	$\hat{\sigma}_g$	$se(\hat{\sigma}_g)$		
1	1.00	1.01	0.07		
2	1.50	1.41	0.08		

An example

Tables 7.10 and 7.11 give a small simulated example of the procedure. The data were generated as follows. Both groups consisted of 400 simulees, and responded to 9 items. The 1PLM was used to generate the responses. The item parameter values are shown in the second column of Table 7.10. To simulate DIF, for the first group the parameter of item 5 was changed from 0.00 to 0.50. Further, in DIF research it is usually implausible to assume that the populations of interest have the same ability distribution. Therefore, the mean and the standard deviation of the first group were chosen equal to 1.0 and 1.0, respectively, while the mean and the standard deviation of the second group were chosen equal to 0.0 and 1.5, respectively. Item and population parameters were estimated concurrently by MML. Table 7.10 gives the generating values of the parameters, the estimates and the standard errors.

The last two columns of Table 7.10 give the values of the LM statistic and the associated significance probabilities. In this case, the LM statistic has an asymptotic chi-square distribution with one degree of freedom. The test is highly significant for item 5. For the analysis of Table 7.11, item 5 has been splitted into two virtual items: item 5 was assumed to be administered to group 1, item 10 was assumed to be administered to group 2. So the data are now analyzed assuming an incomplete item administration design,

where group 1 responded to the items 1 to 9 and group 2 responded to the item 1 to 4, 10, and 6 to 9 (in that order). As a consequence, the virtual items 5 and 10 were only responded to by one group, and the LM test cannot be performed for these items. It can be seen in Table 7.11 that the values of the LM statistics for the other items are not significant, which gives an indication that the model fit now fits.

Table 7.11 Parameter Generating Values, Estimates and the LM Statistic after Splitting the DIF Item into two Virtual Items

Item	b_k	$\hat{b}_k$	$se(\hat{b}_k)$	LM	Pr
1	−1.00	−0.88	0.12	0.32	0.57
2	0.00	0.18	0.11	1.27	0.26
3	1.00	1.18	0.12	0.23	0.63
4	−1.00	−0.90	0.12	0.23	0.63
5	0.50	0.81	0.15		
6	1.00	1.10	0.12	0.22	0.63
7	-1.00	−0.73	0.12	0.11	0.74
8	0.00	0.16	0.11	0.23	0.63
9	1.00	1.09	0.11	0.90	0.34
10	0.00	0.08	0.15		
Pop	μ	μ	$se(\hat{\mu})$		
1	1.00	1.00	0.11		
2	0.00	0.00	– – –		
Pop	σ_g	$\hat{\sigma}_g$	$se(\hat{\sigma}_g)$		
1	1.00	1.01	0.07		
2	1.50	1.41	0.08		

Person fit

Applications of IRT models to the analysis of test items, tests, and item score patterns are only valid if the model holds. Fit of items can be investigated across students and fit of students can be investigated across items. Item fit is important because in psychological and educational measurement, instruments are developed that are used in a population of students; item-fit then can help the test constructor to develop an instrument that fits an IRT model in that particular population. Examples of the most important item-fit statistics were given in the previous sections. As a next step, the fit of an individual's item score pattern can be investigated. Although a test may fit an IRT model, students may produce patterns that are unlikely given the model, for example, because they are

unmotivated or unable to give proper responses that relate to the relevant ability variable, or because they have preknowledge of the correct answers, or because they are cheating.

As with item-fit statistics, also person-fit statistics are defined for the 1-, 2- and 3-parameter model, in a logistic or normal ogive formulation, and in a frequentist or Bayesian framework. Analogous to item-fit statistics, person-fit statistics are based on differences between observed and expected frequencies. A straightforward example is the W-statistic introduced by Wright and Stone (1979), which is defined as

$$W = \frac{\sum_{k=1}^{K}(y_k - P_k(\theta))^2}{\sum_{k=1}^{K} P_k(\theta)(1-P_k(\theta))}$$

Since the statistic is computed on for individual students, we drop the index *i*. Usually, the statistic is computed using maximum likelihood estimates of the item parameters (obtained using CML or MML), and maximum likelihood estimates of the ability parameter. A related statistic was proposed by Smith (1985, 1986). The set of test items is divided into *G* non-overlapping subtests denoted A_g $(g^{-1},\ldots, G)$ and the test is based on the discrepancies between the observed scores and the expected scores under the model summed within subsets of items. That is, the statistic is defined as

$$UB = \frac{1}{G-1}\sum_{g=1}^{G}\frac{\left[\sum_{k\in A_g}(y_k - P_k(\theta))\right]^2}{\sum_{k\in A_g} P_k(\theta)(1-P_k(\theta))}$$

One of the problems of these statistics is that the effects of the estimation of the parameters are not taken into account. However, as above, the test based on the UBstatistic can be defined as an LM-test (Glas & Dagohoy, 2003).

Person-fit statistics can also be computed in a fully Bayesian framework. Above it was outlined that in the Bayesian approach, the posterior distribution of the parameters of the 3PNO model can be simulated using a Markov chain Monte Carlo (MCMC) method. Person fit can then be evaluated using a posterior predictive check based on an index T(y,ξ), where y stands for the ensemble of the response patterns and ξ stands for the ensemble of the model parameters. When the Markov chain has converged, draws from the posterior distribution can be used to generate model-conform data y^{rep} and to compute a so-called Bayes p-value defined by

$$\Pr\left(T(y^{rep}, \xi) > T(y, \xi) \mid y\right)$$

So person-fit is evaluated by computing the relative proportion of replications, that is, draws of ξ from the posterior distribution $p(\xi \mid y)$, where the person-fit index computed using the data, $T(y, \xi)$, has a smaller value than the analogous index computed using data generated to conform to the IRT model, that is $T(y^{rep}, \xi)$. Posterior predictive checks are constructed by inserting the UB and UD statistics for $T(y, \xi)$. After the burn-in period, when the Markov Chain has converged, in every nth iteration, using the current draw of

the item and student parameters, a person-fit index is computed, a new model-conform response pattern is generated, and a value of the person-fit index for the replicated data is computed. Finally, a Bayesian pvalue is computed as the proportion of iterations where the statistic for the replicated data is greater than the statistic for the observed data.

Testing the assumption about the ability distribution

The previous sections discusses item-fit and person-fit. However, for MML estimation, it is assumed that the ability parameters have some common distribution, usually a normal distribution, and also this assumption needs to be tested. Unfortunately, the topic of testing this assumption has been underexposed. In fact, only for exponential family IRT models, that is, the 1PLM and the 2PLM with fixed item discrimination parameters (the so-called OPLM) a well-founded procedure based on statistics with a known asymptotic distribution is available. However, the parameters in these models can also be estimated using CML, so the relevance for these tests is limited to cases where interest is explicitly on the ability distribution.

Although the effects of misfitting items and incorrect assumptions about the ability distribution can hardly be separated, it is most practical to start evaluation of model fit by testing the appropriateness of the specified distribution: incorrect modeling of the ability distribution has an impact on evaluation of the model fit for all items, whereas a hopefully small number of misfitting items may have little effect on evaluation of the model for the ability distribution. Since the student's sum score is a sufficient statistic for the ability parameter, the statistic is based on evaluating the difference between the observed and MML expected score distribution (the score distribution given the MML estimates of the item and population parameters). For the OPLM with a normal ability distribution with a mean μ and a standard deviation σ, the test is based on the differences

$$n_r - E(N_r) \quad , \quad r = 0,\ldots,\max(r)$$

where the stochastic variable N_r stands for the number of students obtaining score r, n_r stands for its realization. The expectations are evaluated using MML estimates. If the discrimination indices are set equal to zero, the test is defined for the 1PLM. As above, the test statistic is a quadratic form, where the differences are weighted by their covariance matrix. The statistic has an asymptotic χ^2-distribution with *max(r)-2* degrees of freedom (Glas & Verhelst, 1989, 1995).

7.3 Models for Polytomous Items

7.3.1 Introduction

The present chapter started with an example of parameter separation where the responses to the items were polytomous, that is, in the example of Table 7.1 the responses to the items are scored between 0 and 5. Dichotomous scoring is a special case where the item scores are either 0 or 1. Open-ended questions and performance tasks are often scored polytomously. They are usually intended to be accessible to a wide range of abilities and to differentiate among test takers on the basis of their levels of response. Response

categories for each item capture this response diversity and thus provide the basis for the qualitative mapping of measurement variables and the consequent interpretation of ability estimates. For items with more than two response categories, however, the mapping of response categories on to measurement variables is a little less straightforward than for right/wrong scoring.

In the sequel, the response to an item k can be in one of the categories $m=0,\ldots, Mk$. So it will be assumed that every item has a unique number of response categories $1+M_k$. The response of a student i to an item k will be coded by stochastic variables Y_{ikm}. As above, upper-case characters will denote stochastic variables, the analogous lower-case characters the realizations. So

$$y_{ikm} = \begin{cases} 1 & \text{if person } i \text{ responded in category } m \text{ on item } k \\ 0 & \text{if this is not the case,} \end{cases}$$

for $m=0,\ldots, M_k$. A dichotomous item is the special case where $M_k=1$, and the number of response variables is then equal to two. However, the two response variables Y_{ik0} and Y_{ik1} are completely dependent, if one of them is equal to 1, the other must be equal to zero. For dichotomous items, a response function was defined as the probability of a correct response as a function of the ability parameter θ. In the present formulation, we define an item-category function as the probability of scoring in a certain category of the item as a function of the ability parameter θ.

For a dichotomous item, we have two response functions, one for the incorrect response and one for the correct response. However, as with the response variables also the response functions are dependent because the probabilities of the different possible responses must sum to one, both for the dichotomous case $(M_k=1)$ and for the polytomous case ($M_k>1$). The generalization of IRT models for dichotomous responses to IRT models for polytomous responses can be made from several perspectives, several of which will be discussed below. A very simple perspective is that the response functions should reflect a plausible relation with the ability variable. For assessment data, the response categories are generally ordered, that is, a response in a higher category reflects a higher ability level than a response in a lower category. However, items with nominal response categories may also play a role in evaluation; therefore they will be discussed later. Consider the response curves of a polytomous item with 5 ordered response categories given in Figure 7.5. The response curve of a response in the zero-category decreases as a function of ability. This is plausible, because as ability increases, the score of a respondent will probably be in a category $m>0$. Further, respondents of extremely low proficiency will attain the lowest score almost with a probability one. An analogous argument holds for the highest category: this curve increases in ability, and for very proficient respondents the probability of obtaining the highest possible score goes to one. These two curves are in accordance with the models for dichotomous items discussed in the previous sections. The response curves for the intermediate categories are motivated by the fact that they should have a lower zero asymptote because respondents of very low ability almost surely score in category zero, and respondents of very high ability almost surely score in the highest category. The fact that the curves of the intermediate categories are single-peaked has no special motivation but most models below have this property.

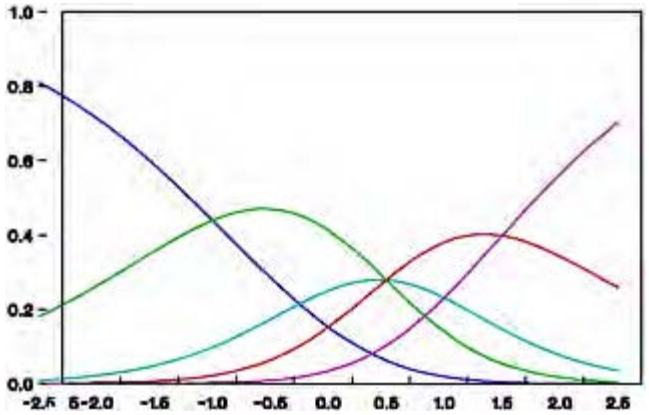


Figure 7.5 Response curves of a polytomously scored item.

Item response models giving rise to sets of item-category curves with the properties sketched here fall into three classes (Mellenbergh, 1995). Models in the first class are called adjacent-category models (Masters, 1982, Muraki, 1992), models in the second class are called continuation-ratio models (Tutz, 1990, Verhelst, Glas, & de Vries, 1997) and models in the third class are called cumulative probability models (Samejima, 1969). These models will be discussed in turn. It should, however, be stressed in advance, that though the rationales underlying the models are very different, the practical implications are often negligible, because their item-category response curves are so close that they can hardly be distinguished in the basis of empirical data (Verhelst, Glas, & de Vries, 1997). On one hand, this is unfortunate, because the models represent substantially different response processes; on the other hand, this is also convenient, because statisticians can choose a model formulation that supports the most practical estimation and testing procedure. In this sense, the situation is as in the case of models for dichotomous data where one can either choose a logistic or normal ogive formulation without much consequence for model fit, but with important consequences for the feasibility of the estimation and testing procedures. Finally, it should be remarked that logistic and normal ogive formulations also apply within the three classes of models for polytomous items, so one is left with a broad choice of possible approaches to modeling, estimation and testing.

7.3.2 Adjacent-category models

In Section 7.2.2, the Rasch model or 1PLM was defined by specifying the probability of a correct response. However, because only two response categories are present and the probabilities of responding in either one of the categories sum to one, Formula (4) could also be written as

$$\frac{p(Y_{ik}=1\,|\,\theta_i,b_k)}{p(Y_{ik}=0\,|\,\theta_i,b_k)+p(Y_{ik}=1\,|\,\theta_i,b_k)} = \frac{\exp(\theta_i-b_k)}{1+\exp(\theta_i-b_k)} \qquad (22)$$

that is, the logistic function $\Psi(\theta_i - b_k)$ describes the probability of scoring in the correct category rather than in the incorrect category. Formula (22) defines a conditional probability. The difficulty of item *k*, b_k, is now defined as the location on the latent θ scale at which a correct score is as likely as an incorrect score.

Masters (1982) extends this logic to items with more than two response categories. For an item with three ordered categories scored 0, 1 and 2, a score of 1 is not expected to be increasingly likely with increasing ability because, beyond some point, a score of 1 should become less likely because a score of 2 becomes a more probable result. It follows from the intended order $0<1<2,\ldots,<m_k$ of a set of categories that the conditional probability of scoring in *m* rather than in *m-1* should increase monotonically throughout the ability range. The probability of scoring in in *m* rather than in *m-1* is thus modeled as

$$\frac{p(Y_{ikm}=1\,|\,\theta_i,b_k)}{p(Y_{ik(m-1)}=1\,|\,\theta_i,b_k)+p(Y_{ikm}=1\,|\,\theta_i,b_k)} = \frac{\exp(\theta_i-b_{km})}{1+\exp(\theta_i-b_{km})} \qquad (23)$$

and b_{km} is the point on the latent θ scale where the odds of scoring in either category are equal. Because it is related to both the category *m* and category *m-1*, the item parameter b_{km} cannot be seen as the parameter of category *m* alone. Masters (1982) shows that these conditional probabilities can be rewritten to the unconditional probability of a student *i* scoring in category *m* on item *k* given by

$$p(Y_{ikm}=1\,|\,\theta_i,b_k) = \frac{\exp(m\theta_i-\sum_{g=1}^{m}b_{km})}{1+\sum_{h=1}^{M_k}\exp\left[h\theta_i-\sum_{g=1}^{m}b_{km}\right]} \qquad (24)$$

for $m=1,\ldots, M_k$. This model is known as the partial credit model (PCM). The important part in this formula is the nominator; the denominator is a sum over all nominators and it assures the response probabilities sum to one. Note that the probability of a response in the zero-category, denoted $Y_{ik0}=1$, has a nominator 1 and a denominator as in Formula (24).

The PCM can also be derived from a different perspective. As mentioned above, Fischer (1974) has shown that the Rasch model for dichotomous items can be derived from a set of assumptions, including sufficiency of the number correct score. In the PCM, the sufficient statistic for the ability parameter is the weighted sum score

$$R_i = \sum_{k=1}^{k} d_{ik} \sum_{m=1}^{M_k} mY_{ikm}$$

that is, the sum of the weights *m* of the categories in which the items were responded to (Andersen, 1977). However, this immediately suggests a generalization of the model. Authors as Kelderman (1984, 1989), Verhelst and Glas (1995) and Wilson and Masters (1993) have considered various more general sufficient statistics for ability. Among other models, they all consider the weighted-score statistic

$$R_i = \sum_{k=1}^{k} d_{ik} \sum_{m=1}^{M_k} a_{km}Y_{ikm}$$

where the weights are positive, integer-valued and ordered $a_{k1}<a_{k2}<,\ldots,<a_{kMk}$. This results in a model

$$p(Y_{ikm}=1\,|\,\theta_i,a_k,b_k) = \frac{\exp(a_{km}\theta_i-\sum_{g=1}^{m}b_{kg})}{1+\sum_{h=1}^{M_k}\exp\left[a_{kh}\theta_i-\sum_{g=1}^{h}b_{kg}\right]},$$

for $m=1,\ldots, M_k$. If the weights a_{km} satisfy conditions certain conditions (see Andersen, 1977, the conditions are mild and usually met), CML estimation is feasible. Further generalizations concern the status of the weights a_{km}. In the dichotomous case they can be treated as known constants or as unknown parameters that should be estimated. This, of course, also applies here. Several approaches are open. Muraki (1992) considers a model where the weights have the form $a_{km}=m\alpha_k$, where a_k is an unknown positive item discrimination parameter. Multiplying this item discrimination parameter with the category number m guarantees the ordering of the weights a_{km}. Muraki's formulation is known as the generalized partial credit model. Its parameters can be estimated using MML. Finally, Bock (1972) proposed the nominal categories model where the parameters a_{km} are free unknown estimands. In this very general formulation, the model specifies the probability of a student's response in one of several mutually exclusive and exhaustive categories as a function of stimulus characteristics and student attributes. It has the generalized partial credit model as a special case.

7.3.3 Continuation-ratio models

The partial credit model (PCM) by Masters (1982) is a unidimensional item response model for analyzing responses scored in two or more ordered categories. The model has some very desirable properties: it is an exponential family model, so minimal sufficient statistics for both the item and student parameters exist and CML estimation can be utilized. However, as shown, the relation between the response categories and the item parameters is rather complicated. As a consequence, the PCM may not always be the most appropriate model for analyzing data.

In the present section, an alternative to the PCM, called the Steps Model, is described, which is conceptually quite different. The development starts with considering a multistage testing design with dichotomous items, where the choice of a follow-up test is a function of the responses on the previous items. It is shown that it is possible to view polytomous response data as a special case of data emanating from a multistage testing design with dichotomous items, where every test consists of one dichotomous item only.

Verhelst, Glas and de Vries (1997) develop the model by assuming that a polytomous item consists of a sequence of item steps. Every item step corresponds with a so-called conceptual dichotomous Rasch item. Further, the student is only administered the next conceptual Rasch item if a correct response was given to the previous one. So it is assumed that the student keeps taking item steps until an incorrect response is given. It is assumed that if a conceptual item is administered, the Rasch model holds, so the probability of taking a step is given by

$$p(Y_{ikm}=1 \mid d_{ikm}=1,\theta_i,b_{km}) = \frac{\exp(\theta_i - b_{km})}{1+\exp(\theta_i - b_{km})}$$

where d_{km} is a design variable as defined for dichotomous items by Formula (3), b_{km} is the difficulty parameters of step m within item k. Let r_{ik} be the number of item steps taken within item k, that is,

$$r_k = \sum_{m=1}^{M_k} d_{km}\, y_{km}$$

In Table 7.12, for some item with $M_k=3$, all possible responses y_k, $y_k=(y_{k1}, y_{k2}, y_{k3})$ are enumerated, together with the associated probabilities $P(y_k \mid \theta, b_k)$.

Table 7.12 Response Probabilities in the Continuation-Ratio Model.

y_k	r_k	$P(y_k \mid \theta, b_k)$
0,c,c	0	$\frac{1}{1+\exp(\theta_i - b_{k1})}$
1,0,c	1	$\frac{\exp(\theta_i - b_{k1})}{[1+\exp(\theta_i - b_{k1})][1+\exp(\theta_i - b_{k2})]}$
1,1,0	2	$\frac{\exp(\theta_i - b_{k1})\exp(\theta_i - b_{k2})}{[1+\exp(\theta_i - b_{k1})][1+\exp(\theta_i - b_{k2})][1+\exp(\theta_i - b_{k3})]}$
1,1,1	3	$\frac{\exp(\theta_i - b_{k1})\exp(\theta_i - b_{k2})\exp(\theta_i - b_{k2})}{[1+\exp(\theta_i - b_{k1})][1+\exp(\theta_i - b_{k2})][1+\exp(\theta_i - b_{k3})]}$

From inspection of Table 7.12, it can be easily verified that in general

$$P(y_k \mid \theta, b_k) \frac{\exp\left[r_k\theta - \sum_{m=1}^{M_k} b_{km}\right]}{\prod_{h=1}^{\min(M_k,\, r_k+1)} [1+\exp(\theta - b_{km}]}$$

where $min(M_k, r_k+1)$ stands for the minimum of M_k and r_k+1. The model does not have sufficient statistics, so it cannot be estimated using CML (Glas, 1988b). The model is straightforwardly generalized to a model where the item steps are modeled by a 2PLM, or to a normal ogive formulation. With the definition of a normal ability distribution, any program for dichotomous data that can compute MML estimates in the presence of missing data can estimate the parameters. The same holds in a Bayesian framework, where any software package that can perform MCMC estimation with incomplete data can be used to estimate the model parameters.

7.3.4 Cumulative probability models

In adjacent-category models are generally based on a definition of the probability that the score, say R_k, is equal to m conditional on the event that it is either m or $m-1$, for instance,

$$P(R_k = m \mid R_k = m \text{ or } R_k = m-1) = \Psi(a_k(\theta - b_{km}))$$

Continuation-ratio models, on the other hand, are based on a definition of the probability of scoring equal to, or higher than m given that the score is at least $m-1$, that is

$$P(R_k \geq m \mid R_k \geq m-1) = \Psi(a_k(\theta - b_{km}))$$

An alternative, yet older, approach can be found in the model proposed by Samejima (1969). Here the probability of scoring equal to, or higher than m is not considered conditional on the event that the score is at least $m-1$, but this probability is defined by

$$P(R_k \geq m) = \Psi(a_k(\theta - b_{km}))$$

It follows that the probability of scoring in a response category m is given by

$$P(R_k = m) = P(Y_{ikm} = 1 \mid \theta, b_k) = \Psi(a_k(\theta - b_{km})) - \Psi(a_k(\theta - b_{k(m+1)})) \quad (25)$$

for $m=1,\ldots M_k-1$. Since the probability of obtaining a score M_k+1 is zero and since everyone can at least obtain a score 0, it is reasonable to set $P(R_k \geq M_k + 1)=0$ and $P(R_k \geq 0)=1$. As a result

$$P(R_k = 0) = P(Y_{ik0} = 1 \mid \theta, b_k) = 1 - \Psi(a_k(\theta - b_{k1}))$$

and

$$P(R_k = M_k) = P(Y_{ikM_k} = 1 \mid \theta, b_k) = \Psi(a_k(\theta - b_{kM_k}))$$

To assure that the differences in Formula (25) are positive, it must hold that $\Psi(a_k(\theta - b_{km})) > \Psi(a_k(\theta - b_{k(m+1)}))$, which implies that $b_1 < b_2 < \ldots < b_{Mk}$. Further, contrary to the case of continuation-ratio models, the discrimination parameter a_k must be the same for all item steps.

The model can both be estimated in a likelihood-based and Bayesian framework. The former is done using MML estimation; the procedure is implemented in the program Multilog (Thissen, 1991). Johnson and Albert (1999) worked out the latter approach in detail.

7.3.5 Estimation and testing procedures

Since continuation-ratio models can be viewed as models for dichotomous data obtained in a design with structural missing data, estimation and testing procedures directly follow from the procedures for the analogous IRT models. So the MML and MCMC procedures described above directly hold. An exception is CML estimation, which is not feasible for continuation-ratio models (Glas, 1998b). Both for cumulative-probabilities models and adjacent-category models MML and MCMC estimation procedures are feasible, but in practice, MCMC procedures are most practical for cumulative-probabilities models

(Johnson & Albert, 1999) and MML procedures are most practical for adjacent-category models (Glas & Verhelst, 1989). It is beyond the scope of this chapter to treat all estimation and testing procedures in detail, but to give a flavor of the methods, a likelihood-based estimation and testing procedures will be sketched for the partial credit model.

The theory for MML estimation presented above for dichotomous items can also be used for the PCM (Bock & Aitkin, 1981; Mislevy, 1984, 1986; Glas & Verhelst, 1995). The probability of a student / scoring in category *m* on item *k* is given by Formula (24). Using local independence, the probability of a student's response pattern is given by

$$P(\mathbf{y}_i \mid \theta_i, b) = \prod_{k=1}^{K}\left[\prod_{j=0}^{M_k} P(Y_{ikm}=1 \mid \theta_i)^{x_{ikj}}\right]^{d_{ik}}$$

where y_i is the response pattern of student *i*. Assuming independence between respondents, the likelihood given the entire data set is the product of these expressions, so the likelihood is analogous to the likelihood in the dichotomous case, given by Formula (10). As for the 1PLM, also here three maximum likelihood estimation procedures are available: joint maximum likelihood (JML), conditional maximum likelihood (CML) and marginal maximum likelihood (MML).

The JML estimation equations are straightforward generalization of the analogous equations for dichotomous items. They are given by

$$r_i = \sum_{k=1}^{K} d_{ik}\left[\sum_{m=1}^{M_k} mP(Y_{ikm}=1 \mid \theta_i, b_k)\right], \qquad \text{for } i=1,\ldots,N$$

and

$$s_{km} = \sum_{i=1}^{N} P(Y_{ikm}=1 \mid \theta_i, b_k), \qquad \text{for } k=1,\ldots,k,\ j=1,\ldots,m_k$$

So the respondents' sum scores r_i and the number of respondents scoring in category *m* of item *k* are equated with their respective expected values. However, as in the dichotomous case, JML does not result in consistent estimators, because the number of student parameters goes to infinity as the sample size goes to infinity. To obtain such estimates we can use either CML or MML.

In CML estimation, a likelihood function of the item parameters given the observed sufficient statistics is maximized. This leads to the estimation equations

$$s_{km} = \sum_{i=1}^{N} P(Y_{ikm}=1 \mid r_i, b)$$

for k=1,…, k and j=1,…, m_k. Here the numbers of respondents scoring in category *m* of item *k* are equated with their conditional expected values.

In MML estimation, the likelihood is marginalized with respect to the ability parameters under the assumption that they have a common normal distribution. This leads to the estimation equations

$$s_{km} = \sum_{i=1}^{N} E[P(Y_{ikm}=1 \mid \theta_i, b_k) \mid \mathbf{y}_i, b, \mu, \sigma]$$

for k=1,…, k and j=1,…, m_k, where the expectation is with respect to the posterior distribution of θ given the response pattern y_i.

For solving the estimation equations, Bock and Aitkin (1981) employ the EM algorithm [expectation-maximization algorithm], where the unobserved values of θ are considered to be missing data. The term EM algorithm was introduced in Dempster, Laird and Rubin (1977). It is a general iterative algorithm for ML estimation in incomplete data problems. It handles missing data, firstly, by replacing missing values by a distribution of estimated values, secondly, by estimating new parameters, thirdly, by re-estimating the distribution of missing values assuming the new parameter estimates are correct, and fourth, re-estimate parameters, and so forth, iterating until convergence.

Evaluation of the fit of response curves

Generalization of the Q_1 tests by Orlando and Thissen (2000) to polytomous data turns out to be infeasible. The reason is that these tests are evaluated using score groups. This leads to the problem that responses in high item categories are often unobserved for at low score levels and responses in low item categories are often unobserved for at high score levels. Therefore, the table on which the test is based usually has too many empty cells. The solution is to group score levels, which can be done using the framework of LM tests (Glas, 1999). As above, this LM test is denoted as LM-Q1. Also as above, the score range is partitioned into G subsets, and it is evaluated whether the observed and expected number of responses in the item categories conforms the model. An indicator function $w\left(\mathbf{y}_i^{(k)}, g\right)$ is defined that is equal to one if the sum score on the response pattern without item k falls in subrange g, and equal to zero if this is not the case. To simplify the notation, we will first reparameterize the PCM using a transformation of the item parameters $e_{km} = \sum_{h=1}^{m} b_{kh}$. Then the alternative model on which the LM test is based, is given by

$$P(Y_{imk} = 1 \mid \theta_i, d_k, \delta_{kg}, w(\mathbf{y}_i^{(k)}, g) = 1)$$

$$= \frac{\exp(j\theta_n - (e_{km} + \delta_{kmg}))}{1 + \sum_{h=1}^{Mk} \exp(h\theta_i - (e_{kh} + \delta_{khg}))}$$

for m=1,…, M_k. Under the null model, which is the PCM model, the additional parameter δ_{kmg} is set equal to zero. Notice that parameter δ_{kmg} is different for each category m. In the alternative model, the additional parameter is a free parameter, $\delta_{kmg} \neq 0$. For the LM-Q1 test, it can be shown that the test can be based on the differences

$$s_{kmg} - \sum_{n|g} E[P(Y_{ikm} = 1 \mid \theta_i, b_k) \mid \mathbf{y}_i, b, \mu, \sigma]$$

for k=1,…, K, j=1, M_k, and g=1,…,G. So the test is based on the difference between the observed number of responses in category m of item k of the respondents in subgroup g and its posterior expectation. This expected value is computed using the PCM without the additional parameters δ, so it is computed under the null model. If the difference between

the observed and expected values is large, this means that the PCM model did not fit the data and the additional parameters δ_{kmg} are necessary to obtain model fit.

Evaluation of local independence

Also local independence can be evaluated using the LM framework (Glas, 1999). As above, this LM test is denoted as LM-Q2. A possible dependency between the items *k* and item *l* is modeled as

$$P(y_{ikm}=1, y_{ilp}=1 \mid \theta_i, \mathbf{e}_k, \mathbf{e}_l, \delta_{kl}) =$$

$$\frac{\exp(m\theta_i - e_{km} + h\theta_i - e_{lp} + \delta_{kmlp})}{1+\sum_g \sum_h \exp(g\theta_i - e_{kg} + h\theta_i - e_{lh} + \delta_{kglh})}$$

Note the parameter δ_{igjh} models the association between the two items. The LM2 test is used to test the special model, $\delta_{igjh}=0$, against the alternative model, $\delta_{igjh}\neq 0$.

If the theory of the LM test is applied, it turns out that the test is based on the difference

$$s_{ikgjh} - \sum_{n=1}^{N} E(P_{kg}(\theta_i)P_{jh}(\theta_i) \mid \mathbf{y}_i, a, b),$$

for $g=1,\ldots, M_k$ and $h= 1,\ldots, M_j$. So, this is the difference between the number of students with an observed response in category g of item *j* and an observed response in category *h* of item *j* with its posterior expected value. The expected value is computed using the null model with local independence as assumption. If the LM2 test is significant, the additional parameter is necessary to obtain model fit. The pair of items is locally dependent meaning that an answer on one item influences the answer on the other item.

Person fit

For the UB test, the complete response pattern is split up into a number of parts, say the parts g–$0,\ldots, G$. Then it is evaluated whether the same ability parameter θ can account for the responses in all partial response patterns. Let A_g be the set of the indices of the items in part *g*. We pose the alternative model that this is not the case, that is, for g>0, we pose the model

$$P_{kmg}(\theta) = P(Y_{km}=1 \mid \theta, k \in A_g) = \frac{\exp[m(\theta_o+\theta_g)-b_{km}]}{1+\sum_{h=1}^{M_k}\exp[h(\theta_0+\theta_g)-b_{kh}]}$$

One group g should be used as a reference. As was already shown above, an *LM* statistic can be defined as a quadratic form in the first-order derivatives with respect to θ_0.

Analogously, an LM test for local independence can be based on a model where the response on item *i* depends on the response $r_k = \sum_{m=0}^{M} m y_{km}$ on item *k*. The model is given by

$$P_{km}(\theta) = P(Y_{km} = 1 \mid \theta, r_k,) = \frac{\exp[m(\theta + \delta r_k) - b_{km}]}{1 + \sum_{h=1}^{M} \exp[h(\theta + \delta r_k) - b_{kh}]}$$

Note that δy_k can be interpreted as a shift in ability that is proportional to the response level on item *k*. An *LM* statistic can be defined as a quadratic form in the first-order derivatives with respect to δ.

7.4 Multidimensional Models

In many instances, it suffices to assume that ability is unidimensional. However, in other instances, it may be a priori clear that multiple abilities are involved in producing the manifest responses, or the dimensionality of the ability structure might not be clear at all. In such cases, multidimensional IRT (MIRT) models can serve confirmatory and explorative purposes, respectively. As this terminology suggests, many MIRT models are closely related to factor analytic models; in fact, Takane and de Leeuw (1987) have identified a class of MIRT models that is equivalent to a factor analysis model for categorical data.

MIRT models for dichotomously scored items were first presented by McDonald (1967) and Lord and Novick (1968). These authors use a normal ogive to describe the probability of a correct response. The idea of this approach is that the dichotomous response of student *i* to item *k* is determined by an unobservable continuous random variable. This random variable has a standard normal distribution and the probability of a correct response is equal to the probability mass below some cut-off point η_{ik}. That is, the probability of a correct response is given by

$$p_k(\theta_i) = \Phi(\eta_{ik}) = \Phi(\sum_{q=1}^{Q} a_{kq}\theta_{iq} - b_k)$$

where $\Phi(.)$ is the cumulative standard normal distribution, θ_i is a vector with elements θ_{iq}, $q=1,\ldots,Q$, which are the Q ability parameters (or factor scores) of student *i*, b_k is the difficulty of item *k*, and a_{kq} $(q=1,\ldots,Q)$ are Q factor loadings expressing the relative importance of the Q ability dimensions for giving a correct response to item *j*. For the unidimensional IRT models discussed above, the probability of a correct response as function of ability could be represented by a socalled item response curve. For MIRT models, however, the probability of a correct response depends on a Q-dimensional vector of ability parameters θ_i so $P_k(\theta_i)$ is now a surface rather than a curve. An example of an item response surface by Reckase (1977) is given in Figure 7.6.

The item pertains to two ability dimensions. The respondents' ability vectors $(\theta_{il,}\ \theta_{i2})$ represent points in the ability space and for every point the probability of a correct response is given by the matching point on the surface. Note that if one dimension is held constant, the probability of a correct response increases in the other dimension. So both dimensions can be interpreted as ability dimensions.

Further, it is assumed that the ability parameters θ_{iq}, $q=1,\ldots,Q$, have a Qvariate normal distribution with a mean-vector μ with the elements μ_q, $q=1,\ldots, Q$, and a covariance matrix Σ. So it is assumed that Q ability dimensions play a role in test response behavior.

The relative importance of these ability dimensions in the responses to specific items is modeled by item-specific loadings a_{kq} and the relation between the ability dimensions in some population of respondents is modeled by the correlation between the ability dimensions.

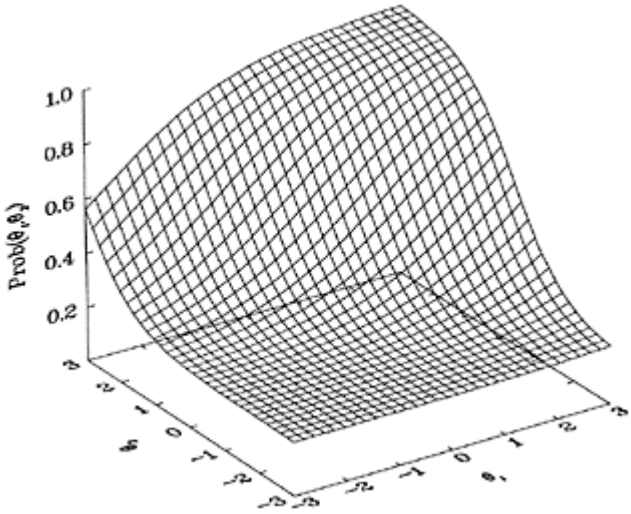


Figure 7.6 Item response surface for a multidimensional IRT model (Reckase, 1977).

In the example of Figure 7.6, the probability of a correct response does not go to zero if the abilities go to minus infinity. In that case, the model must be extended to

$$P_k(\theta_i) = c_k + (1 - c_k)\Phi(\eta_{ik})$$

by introducing a guessing parameter c_k. A comparable model using a logistic rather than a normal-ogive representation has been proposed by Reckase (1985, 1997) and Ackerman (1996a and 1996b).

As in the unidimensional case, restrictions have to be imposed on the parameters to identify the model. One approach to identify the model is setting the mean and the covariance matrix equal to zero and the identity matrix, respectively, and introducing the constraints $a_{jq}=0$, $j=1,\ldots,$ Q−1 and $q=j+1,\ldots,Q$. So here the latent ability dimensions are independent and it is assumed that the first item loads on the first dimension only, the second item loads on the first two dimensions only, and so on, until item $Q-1$, which loads on the first $Q-1$ dimensions. All other items load on all dimensions. An alternative approach to identifying the model is setting the mean equal to the zero, considering the covariance parameters of proficiency distribution as unknown estimands. The model is then further identified by imposing the restrictions, $\alpha_{jq}=1$, if $j=q$, and $\alpha_{jq}=0$, if $j\neq q$, for $j=1,\ldots, Q$ and $q=1,\ldots, Q$. So here the first item defines the first dimension, the second item defines the second dimension, and so forth, until item Q which defines the Q-th

dimension. Further, the covariance matrix Σ describes the relation between the thus defined latent dimensions.

In general however, these identification restrictions will be of little help to provide an interpretation of the ability dimensions. Therefore, as in an exploratory factor analysis, the factor solution is usually visually or analytically rotated. Often, the rotation scheme is devised to approximate Thurstone's simple-structure criterion (Thurstone, 1947), where the factor loadings are split into two groups, the elements of the one tending to zero and the elements of the other toward unity.

As an alternative, several authors (Glas, 1992; Adams & Wilson, 1996; Adams, Wilson & Wang, 1997; Béguin & Glas, 2001) suggest to identify the dimensions with subscales of items loading on one dimension only. The idea is to either identify these $S<Q$ subscales a priori in an confirmatory mode, or to identify them using an iterative search. The search starts with fitting a unidimensional IRT model by discarding non-fitting items. Then, in the set of discarded items, items that form a second unidimensional IRT scale are identified, and this process is repeated until S subscales are formed. Finally, the covariance matrix Σ between the latent dimensions is estimated either by imputing the item parameters found in the search for subscales, or concurrently with the item parameters leaving the subscales intact.

Several methods have been proposed to estimate the model. The first approach is to use a two-step procedure where the first step consists of estimating the covariance matrix of the latent variables using tetrachoric correlations and the second step consists of factor analyzing this matrix using standard software (LISREL, Jöreskog & Sörbom, 1996; EQS, Bentler, 1992; LISCOMP, Muthén, 1987). A second approach, developed by McDonald (1967, 1982), is based on an expression for the association between pairs of items derived from a polynomial expansion of the normal ogive. The procedure is implemented in NOHARM (Normal-Ogive harmonic Analysis Robust Method, Fraser, 1988). The third approach, using all information in the data, and therefore labeled "'Full Information Factor Analysis'", was developed by Bock, Gibbons and Muraki (1988). This approach is a generalization of the marginal maximum likelihood (MML) estimation procedure for unidimensional IRT models (see, Bock & Aitkin, 1981), and has been implemented in TESTFACT (Wilson, Wood & Gibbons, 1991). MML estimates for MIRT models with subscales can be obtained using CONQUEST (Wu, Adams & Wilson, 1997). Finally, fully Bayesian approaches with computational methods based on the Gibbs sampler were proposed by Shi and Lee (1998) and Béguin and Glas (2001). For an overview of the relative merits of the various procedures refer to the latter two articles.

For the generalization of the MIRT model to polytomous items, the same three approaches are possible as in the unidimensional case: adjacent-category models, continuation-ratio models and cumulative probability models. All three possibilities are feasible, but only the former and the latter will be discussed here to explicate some salient points.

In the framework of the cumulative probabilities approach, a model for polytomous items with M_k ordered response categories can be obtained by assuming M_k standard normal random variables, and M_k cut-off points η_{ikm} for $m=1,\ldots, M_k$. The probability that the response is in category m is given by

$$p_{km}(\theta_i) = \Phi(\eta_{ik(m-1)}) - \Phi(\eta_{ikm})$$

$$\text{where } \eta_{ikm} = \sum_{q=1}^{Q} \alpha_{kq}\theta_{iq} - b_{km}, \quad \eta_{ik(m-1)} > \eta_{ikm}, \quad \eta_{ik0} = \infty, \text{ and } \eta_{ikM_k} = -\infty$$

Takane and de Leeuw (1987) point out that also this model is both equivalent to a MIRT model for graded scores (Samejima, 1969) and a factor analysis model for ordered categorical data (Muthén, 1984). This model can be estimated using standard software for factor analysis (see above) or using a fully Bayesian approach in combination with a MCMC algorithm (Shi & Lee, 1998).

In the framework of adjacent categories models, the logistic versions of the probability of a response in category m can be written as

$$p_{km}(\theta_i) = \exp\left[m\sum_{q=1}^{Q} a_{kq}\theta_{iq} - \sum_{h=1}^{m} b_{kh} \right] / h(\theta_i, a_k, b_k)$$

where $h(\theta_i, a_k, e_k)$ is some normalizing factor that assures the sum over all possible responses on an item is equal to one. The probability $p_{km}(\theta_i)$ is determined by the compound $\sum_{q=1}^{Q} a_{kq}\theta_{iq}$ so every item addresses the abilities of a respondent in a unique way. Given this ability compound, the probabilities of responding a certain category are analogous to the unidimensional partial credit model by Masters (1982). Firstly, the factor *m* indicates that the response categories are ordered and that the expected item score increases as the ability compound $\sum_{q=1}^{Q} a_{kq}\theta_{iq}$ increases. And secondly, the item parameters b_{kh} are the points where the ability compound has such a value that the odds of scoring either in category *m-1* or *m* are equal.

7.5 Multilevel IRT Model

7.5.1 Models for item parameters

In the previous sections, variability of item parameters was treated as a fixed effect, that is, the item parameters were a finite number of unique entities. In the present section, the focus is on item parameters as random effects, that is, the item parameters are seen as exchangeable draws from a distribution. Interest in item sampling relates to the introduction of computer-generated items in educational measurement. Using item-cloning techniques (see, for instance, Bejar, 1993, or Roid & Haladyna, 1982), items can be generated by a computer from a smaller set of “parent items” through the use of transformation rules. An example is the “replacement set procedure” (Millman & Westman, 1989), where elements of the parent item (e.g., key terms, relations, numbers, and distractors) are randomly chosen from well-defined sets of alternatives. Because this

introduces (slight) random variation between items derived from the same parent, it becomes efficient to model the item parameters as random and shift the interest to the hyperparameters that describe the distributions of the item parameters within parents (Glas & van der Linden, 2001, 2003). Another example of a model with random item parameters is given in Janssen, Tuerlinckx, Meulders and de Boeck (2000). To support standard setting on a criterion-referenced test with sections of items in the test grouped under different criteria, an IRT model is developed with random item parameters drawn from different distributions for different sections. A Bayesian argument for this approach is that if the only thing known a priori about the items is that they are grouped under common criteria, they are exchangeable given the criterion and can be treated as if they are a random sample.

Glas and van der Linden (2001, 2003) define the model as follows. Consider a set of item populations $p=1,\ldots, P$ of size $K_1,\ldots, K_P$, respectively. The items in population p will be labeled $k_p=1,\ldots, K_p$. The first-level model is the 3PLM, which describes the probability of a correct response as $p(y_{ik_p} \mid \theta_i, a_{k_p}, b_{k_p}, c_{k_p})$ as in Formula (7), but with the subscript changed from k to k_p. In the Level 2 model, the values of the item parameters $b_{k_p} a_{k_p} c_{k_p}$ are considered as realizations of a random vector. It is assumed that the item parameters, say ξ_{k_p}, have a 3-variate normal distribution with mean μ_p and a covariance matrix Σ_p. To support the assumption of normality, the item parameters are transformed as $\xi_{k_p} = (a_{k_p}, b_{k_p}, \text{logit}\, c_{k_p})$ or as $\xi_{k_p} = (\log a_{k_p}, b_{k_p}, \text{logit}\, c_{k_p})$. The logit transformation is standard way to map a probability, such as c_{k_p}, to the real continuum, and taking the logarithm of a_{k_p} assures that a_{k_p} is positive.

In general, the model can be estimated by Bayesian methods based on the MCMC procedure (for the 1PLM, see, Janssen, Tuerlinckx, Meulders & de Boeck, 2000; for the 3PLM, see, Glas & van der Linden, 2001) or by MML (Glas & van der Linden, 2003). However, Glas and van der Linden (2001) point out that the application can interfere with the feasibility of certain estimation procedures. This arises in computer-based item generation were the computer generates a new item for each examinee ("item generation on the fly"). An example is some arithmetic task where new values of variables are drawn in every presentation of the item. In that case, the item parameters a_{k_p}, b_{k_p} and c_{k_p} are unique for each examinee and there is only one item response available to estimate these three parameters. Because of this under-determination, these parameters cannot play a role as auxiliary variables in the MCMC procedure. In the MML estimation, however, they can be treated as nuisance parameters and integrated out of the likelihood function.

7.5.2 Testlet models

A testlet is a subset of items related to some common context. Haladyna (1994) refers to context-dependent item sets. Usually, these sets take the form of a number of multiple choice items organized under or within some text. Haladyna (1994) gives examples of comprehension type items sets and problem solving type item sets. When a test consists of a number of testlets, both the within and between dependence between the items play a role. One approach is to ignore this hierarchical dependence structure and analyze the test as a set of atomistic items. This generally leads to an overestimate of measurement

precision and bias in the item parameter estimates (Sireci, Wainer & Thissen, 1991; Yen, 1993; Wainer & Thissen, 1996). Another approach is to aggregate the item scores within the testlet to a testlet score and analyze the testlet scores using an IRT model for polytomously scored items. This approach discards part of the information in the item responses, which will lead to some loss of measurement precision. However, this effect seems to be small (Wainer, 1995). The rigorous way to solve the problem is to model the within and between dependence explicitly. Bradlow, Wainer & Wang (1999, also see, Wainer, Bradlow, & Du, 2000) introduce a generalization of the 3PLM given by

$$p(y_{ik} \mid \theta_i, a_k, b_k, c_k, \gamma_{it(k)}) = c_k + (1 - c_k) + \Psi(a_k(\theta_i - b_k + \gamma_{it(k)}))$$

where $t(k)$ is the testlet to which item k belongs and $\gamma_{it(k)}$ a student-specific testlet effect. It is assumed that $\gamma_{it(k)}$ has a normal distribution with a mean equal to zero and a variance that gauges the importance of the testlet effect. Further, it is assumed that θ has a standard normal distribution.

The parameters in the model can be estimated in a Bayesian framework using MCMC (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow & Du, 2000) or in a frequentist framework using MML (Glas, Wainer & Bradlow, 2000).

7.5.3 Models for ratings

Closely related to the testlet model is the IRT model for analyzing ratings by Patz and Junker (1999a, 1999b). Suppose that a student i performs tasks labeled $k=1,\ldots, K$, and the tasks are rated by raters labeled $t=1,\ldots, T$. For simplicity, it is assumed that the response variables y_{ikt} are dichotomous. Then the problem is that the ratings pertaining to the same item cannot be viewed as independent, because they relate to the same response by student /. First a model will be presented and then it will be shown that this model provides an acceptable specification of the dependence of the responses of different raters pertaining to the same task. Consider a model where the students' ability parameters have a standard normal distribution with density g(θi). For every task k, the student gives a response ξ_{ik}. The raters base their ratings on this response, but in the model it is an unobserved latent response that depends on the ability level of the respondent θ_i. This dependence is modeled by introducing a distribution for ξ_{ik} that depends on θ_i. It is assumed that the distribution is normal with a density denoted by $h(\xi_{ik}| \theta_i, \sigma)$. Further, the model contains parameters for constant effects: bk models the item difficulty and δ_t models the leniency of the rater. With these assumptions, the likelihood is given by

$$L = \prod_{i=1}^{N}\prod_{k=1}^{K}\prod_{t=1}^{T} P_{kt}(\xi_{ik} - b_k + \delta_t)^{y_{ikt}} (1 - P_{kt}(\xi_{ik} - b_k + \delta_t))^{1-y_{ikt}} h(\xi_{ik} \mid \theta_i, \sigma) g(\theta_i)$$

where $P_{kt}(\xi_{ik} - b_k + \delta_t)$ is the probability of a correct response $P(Y_{ikr}=1|\xi_{ik}, b_k, \delta_t)$. This probability can, for instance, be modeled by a logistic function, say $P_{kt}(\xi_{ik} - b_k + \delta_t) = \Psi(\xi_{ik} - b_k + \delta_t)$. This model could be further enhanced with item discrimination and guessing parameters. The model can be estimated by MML after integrating out the unobserved variables ξ_{ik} and θ_i, or in a Bayesian framework using the MCMC algorithm (see, Patz & Junker, 1999a, 1999b).

8
Applications of Measurement Models

8.1 Test Equating and Linking of Assessments

8.1.1 Data collection designs

In the introduction of the previous chapter, it was shown that one of the important features of IRT is the possibility of analyzing so-called incomplete designs. In incomplete designs the administration of items to persons is such, that different groups of persons have responded to different sets of items. In the present section, a number of possible data collection designs will be discussed.

A design can be represented in the form of a persons-by-items matrix. As an example, consider the design represented in Figure 8.1. This figure is a graphical representation of a design matrix with as entries the item administration variables d_{ik} *(I =1,…, N* and *k=1,…, K)* defined by Formula (3) in the previous chapter. The item administration variable d_{ik} was equal to 1 if person *i* responded to item *k*, and 0 otherwise. At this moment, it is not yet specified what caused the missing data. There may be no response because the item was not presented, or because the item was skipped, or because the item was not reached. In the sequel it will be discussed under which circumstances the design will interfere with the inferences. For the time being assume that the design was fixed by the test administrator and that the design does not depend on an a-priori estimate of the ability level of the respondents.

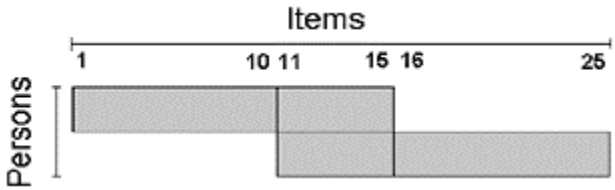


Figure 8.1 Design linked by common items.

In the example, the total number of items is *K=25*. The design consists of two groups of students, the first group responded to the items 1 to 15, and the second group responded to items 11 to 25. In general, assume that *B* different subsets of the total of *K* items have been administered, each to an exclusive subset of the total sample of respondents. These subsets of items will be indicated by the term ‘booklets’. Let *I* be the set of the indices of

the items, so $I=\{1,\ldots, K\}$. Then the booklets are formally defined as non-empty subsets of I_b of I, for $b=1,\ldots, B$. Let K_b denote the number of elements of I_b, that is, K_b is the number of items in booklet b. Next, let V denote the set of the indices of the respondents, so $V=\{1,\ldots, N\}$, where N is the total number of respondents in the sample. The sub-sample of respondents getting booklet b is denoted by V_b and the number of respondents administered booklet b is denoted N_b. The subsets V_b are mutually exclusive, so $N=\Sigma_b N_b$.

To obtain parameters estimates on a common scale, the design has to be linked. For instance, the design of Figure 8.1 is linked because the two booklets are linked by the items 11 to 15, which are common to both booklets. A formal definition of a linked design entails that for any two booklets a and b in the design, there must exist a sequence of booklets with item index sets $I_a, I_{b1}, I_{b2},\ldots, I_b$ such that any two adjacent booklets in the sequence have common items or are administered to samples from the same ability distribution. The sequence may just consist of I_a and I_b. Assumptions with respect to ability distributions do not play a part in CML estimation. So CML estimation is only possible if the design is linked by sequence $I_a, I_{b1}, I_{b2}\ldots, I_b$ where adjacent booklets have common items.

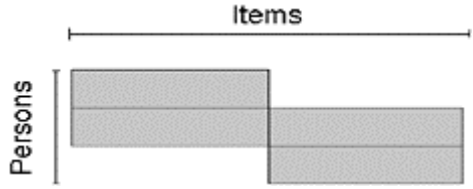


Figure 8.2 Linking by common persons.

This definition may lead to some confusion because it interferes with the more commonly used terms “linking by common items” and “linking by common persons”. Figure 8.1 gives an example of a design linked by common items because the two booklets have common items. Figure 8.2 gives an example of a design that is commonly labeled “linked by common persons”. The definition of a linked design applies here because the first and second booklet have common items and the second and last booklet have common items. Further, the first and last booklet are linked via the second booklet.

An example of linking via common ability distributions is given in Figure 8.3. Again, common items link the middle two booklets. The respondents of the first two booklets are assumed to be drawn from the first ability distribution and the respondents of the last two booklets are assumed to be drawn from a second ability distribution. It must be emphasized that, in general, designs linked by common items are far preferable to designs that are only linked by common distributions, since the assumptions concerning these distributions add to the danger that the model as a whole does not fit the data. Assumptions on ability distributions should be used to support answering specific research questions, not as a ploy for mending poor data collection strategies.

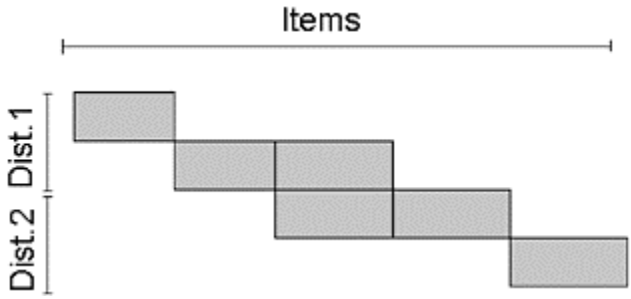


Figure 8.3 Linking by a common distribution.

8.1.2 Multi-stage testing

In the previous section, it was mentioned that missing responses could be the result of many possible causes: presenting items in an incomplete design, skipping items, or not reaching items. Whenever data are collected, however carefully, the possibility, origin and treatment of "missing responses" should be considered. Even if care is taken to ensure that all appropriate respondents are contacted and provide some data, responses on individual variables may be missing, uncodeable or in a category such as 'don't know' or 'not applicable'. If missing observations are present, then the mechanism causing the incompleteness in the data can be characterized according to its degree of randomness. Rubin (1976) described and named a number of types of mechanism. Let D be the missing data indicators, in the present case, D can be viewed as a matrix with as entries the missing data indicators d_{ik} defined by Formula (3). Further, a distinction is made between the observed data y_{obs}, say the observed response patterns of the students, and the unobserved or missing data y_{miss}, say the parts of the person-by-item-matrix where the related design variable d_{ik} equals zero. Following Rubin, data are missing at random (MAR) if the distribution of the design does not depend on the missing data, that is,

$$p(D \mid y_{obs}, y_{mis}, \varphi, x) \quad = \quad p(D \mid y_{obs}, \varphi, x)$$

where φ is a vector of the parameters of the missing data process, and x are covariates that might also determine the missing data process. So the data are MAR if the variables determining the missingness are all observed. In a likelihood-based framework, there is an additional requirement that the space of the parameters of interest (say the item, person and population parameters) and the parameters of the missing data process should be distinct. If MAR and distinctness hold, then maximizing the likelihood of the actually observed data is equivalent with a maximization taking the missing data process into account. That is, we can use the actually observed data alone to obtain estimates of the parameters of interest. In a Bayesian framework, besides MAR, it should also hold that the prior of the parameters of interest and the parameters of the missing data process φ should be independent, and in that case, inferences based on the posterior given the actually observed data suffice.

This has various implications. To mention a few situations where MAR does not hold:

1) If difficult items are differentially skipped by high and low ability students;
2) If a time limit is imposed and speed is correlated with ability;
3) If the test administration design is based on a-priori estimates of ability, or on other covariates that correlate with ability, and these estimates or covariates are not part of the model.

However, there are situations where MAR does hold that are very useful. We will discuss the case of response-contingent designs, such as multi-stage testing and computerized adaptive testing.

Consider the design of Figure 8.4. In this design, all respondents are administered a so-called routing test, say a test of 10 items. If a respondent's score is less than or equal to 5, an easy follow-up test is administered; if the score is more than 5, a difficult test is administered. The procedure is motivated by the fact that matching the ability level of the respondents with the difficulty level of the items results in optimization of the precision of both the item and ability parameter estimates, as was shown in Section 5.2.5.

In this case, MML estimates of the item and population parameters are consistent because the data are MAR, that is, the design is completely determined by the sum scores on the routing test. A small simulated example may illustrate this further. Consider the item parameter estimates in Table 8.1. The design was as in Figure 8.4, the routing test consisted of 10 items, the two follow-up tests consisted of 5 items each. The 1PLM was used to generate the data of 2000 respondents. The ability parameters had a standard normal distribution. Form the true item parameters in the second column, it can be seen that the first follow-up test was easy, while the second was difficult. The MML estimation procedure was used to obtain the item parameter estimates. Note that the response-contingent design did not bias the estimates.

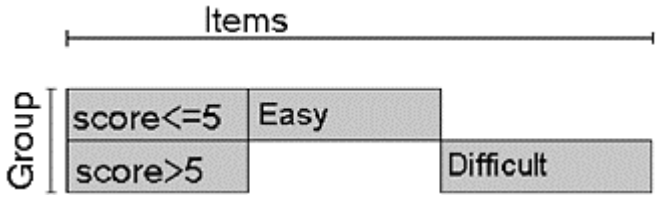


Figure 8.4 Two-stage testing design.

Table 8.1 MML Item Parameter Estimates Obtained in a Multi-Stage Testing Design.

Item	b	b	Se(b)
1	−1.0	−.901	.039
2	−.5	−.460	.037
3	.0	.026	.034
4	.5	.479	.037
5	1.0	1.038	.042

6	−1.0	−1.012	.043
7	−.5	−.542	.041
8	.0	.030	.033
9	.5	.467	.039
10	1.0	.968	.043
11	−1.0	−1.089	.076
12	−.5	−.536	.071
13	.0	−.093	.069
14	−.5	−.436	.065
15	−1.0	−1.066	.075
16	1.0	1.054	.073
17	.5	.593	.070
18	.0	.077	.069
19	.5	.490	.065
20	1.0	1.099	.075

The estimates of the ability parameters and their standard errors are given in Table 8.2. The estimates were obtained by weighted maximum likelihood with the MML item parameter estimates imputed as constants. Note that a certain observed score on the second booklet represents a higher ability level than the same score on the first booklet. This is as expected, because the second booklet was more difficult. In Table 8.1, it can be seen that the mean difficulty of the first booklet is −.75, while the mean difficulty of the second booklet is 0.75. In Table 8.2, it can be seen that −0.75 and 0.75 are indeed the locations on the latent scale where the respondents administered the first and second booklet attain the smallest standard errors.

Table 8.2 Ability Estimates Obtained in a Multi-Stage Testing Design.

	Booklet 1			Booklet 2		
Score	Freq	θ	Se(θ)	Freq	θ	Se(θ)
0	13	−3.99	1.83	0	−3.47	1.85
1	28	−2.80	.96	0	−2.26	.97
2	57	−2.19	.75	0	−1.63	.76
3	81	−1.75	.65	0	−1.16	.67
4	112	−1.38	.60	0	−.78	.61
5	171	−1.06	.57	0	−.44	.58

6	186	−.76	.55	20	−.13	.56
7	193	−.47	.55	96	.15	.55
8	162	−.18	.55	123	.45	.55
9	105	.10	.56	158	.74	.56
10	38	.41	.58	132	1.04	.57
11	0	.75	.61	130	1.37	.60
12	0	1.12	.66	86	1.74	.66
13	0	1.587	.764	53	2.187	.755
14	0	2.215	.972	46	2.800	.963
15	0	3.426	1.851	10	3.995	1.837

The example shown here is a two-stage testing design. Of course, the design can be branched further, for instance, with four tests in the third stage, eight tests in the fourth stage, etc. A limiting case of multistage is computerized adaptive testing (CAT). Here, every test administered consists of one item, and every item administered is selected from an item bank in such a way that the item parameters and the running estimate of ability are matched to obtain maximum precision. A good introduction to CAT can be found in the introductory volume edited by Wainer (1990), for a more advanced overview refer to van der Linden and Glas (2000). With the advent of powerful computers, application of CAT in large-scale high-stakes testing programs has taken a high flight. Well-known examples in the United States are the Nursing-licensing exam (NCLEX/CAT) by the National Council of State Boards of Nursing and the Graduate Record Examination (GRE). Ever since many other large-scale testing programs have followed. It seems safe to state that at the moment the majority of large-scale testing programs either has already been computerized or are in the process of becoming so. The main motivations for CAT are: (1) CAT makes it possible for students to schedule tests at their convenience; (2) tests are taken in a more comfortable setting and with fewer people around than in large-scale paper-and-pencil administrations; (3) electronic processing of test data and reporting of scores is faster; and (4) wider ranges of questions and test content can be put to use (Educational Testing Service, 1996). In the current CAT programs, these advantages have certainly been realized and appreciated by the examinees. When offered the choice between a paper-and-pencil and a CAT version of the same test, typically most examinees choose the CAT version.

8.1.3 Test equating

Above, several examples were given where the scores on different tests or booklets were equated directly via the latent scale. However, in most instances, tests are not scored using the latent scale, or using statistics derived from fitted IRT models. Most tests are scored using the number-correct score, or some weighted score. In the latter case, the weights are usually chosen according to content-based considerations, rather than, for instance, as the 2PLM the discrimination parameters. In the following sections, it will be

shown how tests can be equated in this kind of situations. Equating the cut-off scores on examinations will be used as an example. In the next section, a number of possible equating designs will be discussed. Next, computation of equivalent scores on subsequent examinations using observed number-correct score equating based on IRT (IRT-OS-NC equating) will be discussed. Then, the results obtained from equating the 1995 language comprehension examinations to their respective reference examinations will be presented. One of the problems studied will be the extent to which the 1—and 2PLM produce comparable results. Finally, two methods for the computation of confidence intervals for the equating functions will be described and compared.

At the end of secondary education in the Netherlands, students participate in central examinations. The grade level they achieve is an important component of the grade level of their certificate. Although much attention is given to producing examinations of equivalent substantive content and difficulty, research has shown that the difficulty of examinations can still fluctuate significantly over the years. (see the Inspection of Secondary Education in the Netherlands, 1992). Further, this research has shown that also the proficiency level of the examinees fluctuates significantly over time. Therefore, a test equating procedure (Angoff, 1971, Holland & Rubin, 1982, Kolen & Brennan, 1995) has been developed for setting the cut-off scores of examinations in such a way that differences in difficulty of examinations are taken into account. The cut-off scores of new examinations are equated to the cut-off score of a reference examination. The reference exanimation (selected by the Committee for the Examinations in Secondary Education) was such that its quality and difficulty presented a suitable reference point.

Designs

The examination data as such are insufficient for equating, because the examinations are not linked by common items or persons. Three designs for creating this link will be discussed. The designs are shown in the Figures 8.5, 8.6 and 8.7.

Before discussing these designs, three important points about the representations used here should be made. Firstly, the sample sizes are not proportional to the sizes of the shaded areas of the in the figures. Secondly, in the second and third design all the items of the two examinations figure in the linking tests. This, of course, need not be the case, in fact in the previous section a design was considered where the link consisted of only three items. A procedure for evaluating standard errors for the equating function will be presented that is very helpful in studying these matters in future research. Thirdly, since the items are in the order in which they appear in the examinations, the order in which the items are presented to the respondents in the linking groups cannot be inferred from the figures.

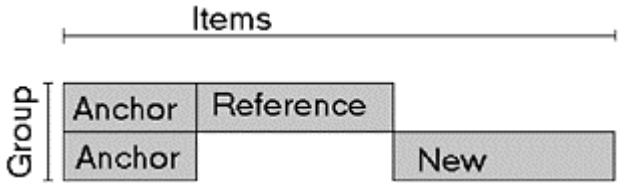


Figure 8.5 Anchor-item design.

In the first design of Figure 8.5, every year, some weeks before the examination takes place, the students are given an anchor test covering material comparable to the content matter of the examinations. In this design, the problem of secrecy is taken care of by keeping the anchor test secret. A problem of this design is that there might be differences in response behavior between the administration of the anchor test and the actual examination. Firstly, the level of proficiency of the students might change during the weeks between taking the anchor test and the examination. This

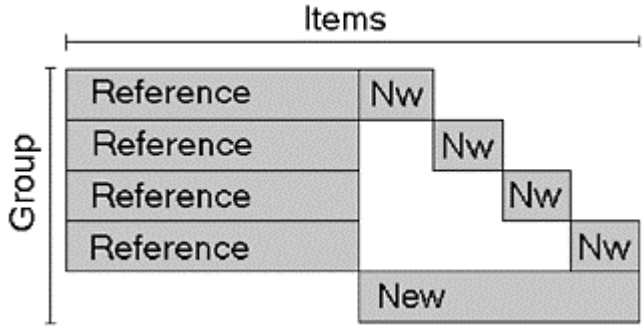


Figure 8.6 Pretest design.

might create a model violation, because, as was shown above, person parameters are supposed to be constant in the rows of the data matrix. If all students increase in proficiency by the same amount, this does not necessarily create a model violation. However, if there are differences in motivation between the two administrations, the shift in estimated proficiency might not be uniform across students, in the sense that some of the students are equally motivated when taking the anchor test and the examination, while others might only be motivated on the actual examination. Further, differences in ability might be accompanied by differences in item parameters, which creates an additional model violation. For instance, if there is a lot of guessing on the anchor test, the same set item parameters will not properly describe response behavior on the two occasions. This, of course, does not disqualify the anchor test approach in general, in many situations the gain in proficiency will be negligible or uniform and there will be no change in the item parameters. However, for the present application these considerations led to the decision to choose another approach.

The second design, depicted in Figure 8.6, is a so-called pretest design. The design shown in Figure 8.6 pertains to the standard-setting procedure used for the Swedish Scholastic Aptitude Test (SweSAT, Emons, 1998). In this design, the students taking the reference examination also respond to items of a future examination. The additional items have no relevance for the final mark obtained by these students and the students are not told in advance which items do not belong to the actual examination. Another strong point of the SweSAT pretest design is that the motivation of the students used in the pretest is guaranteed.

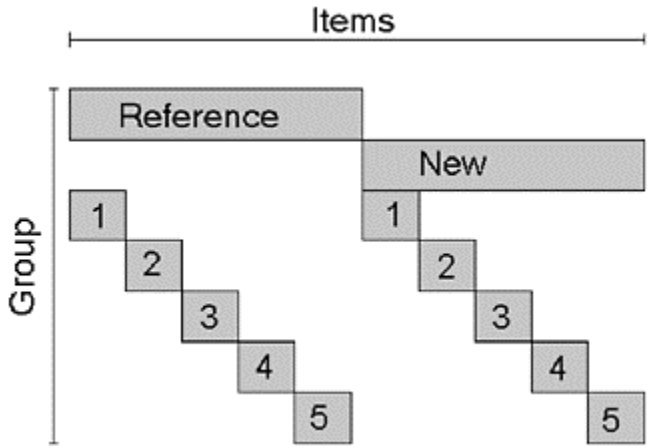


Figure 8.7 Post-test design.

The third design depicted in Figure 8.7 is the design that was actually chosen for equating the examinations discussed here. In this design, linking groups consisting of students not participating in the actual examinations respond to items of the old and the new examination. In the application of this design to the equating problem discussed here, the linking groups were presented their tests directly after the new examination was administered. Five linking groups were sampled from another stream of secondary education and the design was such that the linking groups covered all items of the two examinations. The items were related to reading passages and every linking group was administered two passages, an old one and a new one. One of the concerns when planning the design was to avoid order effects. If, for instance, items from the new examination had always been last, declining concentration and fatigue may result in lowering performance, so that the items of the new examination would appear to be more difficult. Therefore, the order in which old and new reading passages were administered to the linking groups alternated, in the sense that some groups were administered an old passage followed by a new one, while others obtained the alternative sequence.

An important advantage of IRT equating is that the proficiency level of the linking groups and the examination populations need not be the same, in fact, below a multiple group MML estimation procedure will be proposed where every group in the design has its own ability distribution. However, there are also restrictions to the freedom of recruiting linking groups, because their responses must fit the same IRT model as the responses of the examinees. If, for instance, the linking groups do not seriously respond to the items, equating the two examinations via these linking groups would be seriously threatened. Therefore, much attention must be given to the procedure for collecting the data of the linking groups. In the present application, the tests for the linking groups were presented as school tests with consequences for grades.

The equating function

Below, only examinations with dichotomous items will be considered. In the actual project also polytomously scored examinations were equated, but the procedure is essentially the same as the procedure that will be presented here; for results one is referred to Glas and Béguin (1996). Item responses are modeled by the 1PLM, 2PLM, PCM, and GPCM. It will be assumed that every group in the design is sampled from a specific ability distribution. So, for instance, the data in the design depicted Figure 7 are evaluated using seven ability distributions, that is, one distribution for the reference population, one for the examinees of the new examination, and five for the linking groups. Let the ability parameters of the respondents of population b, $b=1,\ldots, B$ have a normal distribution with density $g(\theta|\mu_b, \sigma_b)$. More specifically, the ability parameter of the respondent i has a normal distribution with density, $g(\theta|\mu_{b(i)}, \sigma_{b(i)})$ where $b(i)$ signifies the population to which person i belongs.

Once the parameters of the IRT model have been estimated, the next step is performing equi-percentile equating using estimates of the frequency distribution of the two examinations produced by some population, say, the reference population. That is, equi-percentile equating is carried out as if the reference population had made both examinations. Consider the example of Table 8.3. This example was computed using the data of the 1992 and 1995 examinations English language comprehension at HAVO level, which consisted of 50 dichotomously scores items. In the second and fourth column of Table 8.3, parts of the cumulative relative frequency distributions of the reference and new examination produced by the populations actually administered these two tests are displayed. The complete distributions from which these cumulative distributions were computed are displayed in Figure 8.8a and 8.8b. The figures also contain the estimated score distribution of the reference population on the new examination and of the new population on the reference examination. From these two estimates the cumulative distributions in the third and fifth column of Table 8.3 can be derived. These estimates are computed as follows.

Table 8.3 Cumulative Percentages of the Reference and new Population on the Reference and new Exam.

Population	Reference		New	
Exam	Ref.	New	New	Ref.
Score	Cum. Perc.	Cum. Perc.	Cum. Perc.	Cum. Perc
25	19.8	16.2	11.7	14.7
26	23.6	19.1	14.4	17.5
27	28.0	22.3	17.9	20.7
28	31.8	25.8	21.5	24.2
29	35.9	29.7	25.8	28.0
30	41.0	33.8	29.6	32.1

31	45.6	38.3	34.1	36.6
32	50.4	42.9	38.3	41.3
Mean.	32.3	33.2	34.5	33.6
Std.	7.5	7.5	7.2	7.4
Se (Mean)	.16	.38	.16	.36
Se (Std.)	.10	.11	.11	.12

In the design displayed in Figure 7, the populations can be labeled *b=1, 2,…, 7*. In this design, every population is associated with a specific design vector d_b, that is, the design vector indicating which items were administered to the sample of population *b*. Let d_{ref} and d_{new} be the design vector of the reference and the new examination, respectively. Further, let P_r be the proportion of respondents obtaining a score *r*. The proportion of respondents in population *b* obtaining a score *r* on some examination is estimated by its expected value, that is, as the expected proportion of respondents of a population characterized by population parameters μb and σ_b obtaining a score *r* on a test characterized by the design vector *d*. So the proportion obtaining score r on the reference examination is estimated by

$$E(P_r \mid \boldsymbol{d}_{ref}, a, b, \mu_b, \sigma_b) = \sum_{\{y \mid r, \boldsymbol{d}_{ref}\}} \int P(y_i \mid \boldsymbol{d}_{ref}, \theta, a, b) g(\theta \mid \mu_b, \sigma_b) d\theta$$

where $\{y| r, d_{ref}\}$ stands for the set of all possible response patterns on the reference examination resulting in a score *r*. Notice that this expectation only depends on the parameters of the items of the reference examination. In the same manner, one can also estimate the proportion of students in population *b* obtaining a score *r* on the new examination using

$$E(P_r \mid \boldsymbol{d}_{new}, a, b, \mu_b, \sigma_b) = \sum_{\{y \mid r, \boldsymbol{d}_{new}\}} \int P(y_i \mid \boldsymbol{d}_{new}, \theta, a, b) g(\theta \mid \mu_b, \sigma_b) d\theta$$

where $\{y|r, d_{ref}\}$ stands for the set of all possible response patterns on the new examination resulting in a score *r*.

Returning to the example of Table 8.3, the third column contains part of the cumulative distribution of respondents of the reference population on the new examination The complete cumulative distribution is displayed in Figure 8.8b, together with a confidence interval, and the observed cumulative distribution produced by the reference sample. Computation of confidence intervals will be returned to in the sequel. The cut-off score for the new examination is set in such a way that the expected percentage of respondents failing the new examination in the reference population is approximately equal to the percentage of examinees in the reference population failing the reference examination. In the example of Table 8.3, the cut-off score of the reference examination was 27; as a result 28.0% failed the exam. If this percentage is held constant for the reference population, the new cutoff score should be 29. Obviously, the new examination is easier. This is also reflected in the mean score of the two examination displayed at the bottom of the table. The old and the new cut-off scores are marked with a

straight line in the first column. It can be seen that the percentage of students in the new population failing the new examination is 25.8%. This suggests that the new population is more proficient than the reference population. Also this difference is reflected in the mean scores of the two populations if the examination is held constant.

Figure 8.8a

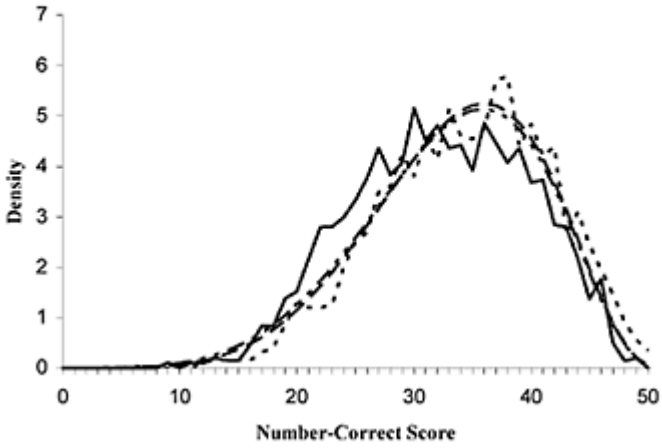


Figure 8.8b

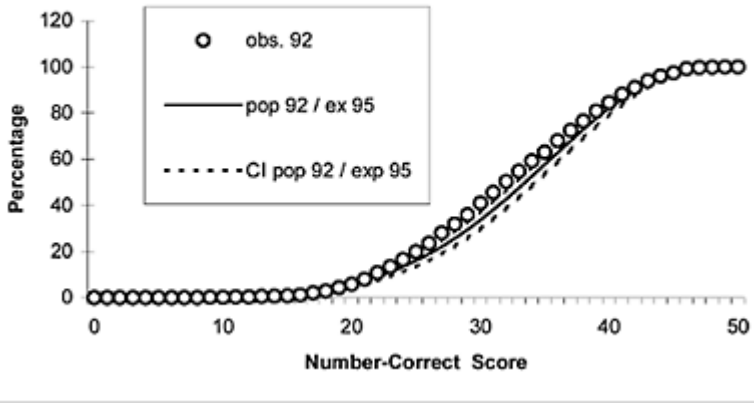


Table 8.4 Data Overview.

	Max	Reference Exam				New Exam				Link
Topic	Score	K_{ref}	N_{ref}	Mean	Std	K_{new}	N_{new}	Mean	Std	N_{link}
German D	50	50	2115	31.72	6.92	50	2021	34.00	6.28	1033
German H	50	50	2129	34.51	5.59	50	2015	32.08	6.27	607
English D	50	50	1693	35.14	6.91	50	2010	34.74	6.87	1137
English H	50	50	2039	32.32	7.45	50	2003	34.45	7.23	873
French D	50	50	1666	33.18	7.39	50	2097	32.28	7.23	1037
French H	50	50	2144	35.72	6.80	50	2138	34.02	7.21	428
Dutch	90	39	1572	56.17	12.05	44	2266	59.01	9.82	701

Results

In the examination campaign of 1995, the cut-off scores of a number of examinations where equated to the cut-off scores of older examinations. The results of seven examinations in language comprehension will be used as an example. The topics of the exams are listed under the heading “Topic” in Table 8.4. The examinations were administered at two levels, topics labeled “D” in Table 8.4 are at MAVO-D-level, topics labeled “H” are at HAVO level. The older examinations were originally administered between 1989 and 1993. All examinations consisted of 50 dichotomous selected response items, except for the examination on language proficiency in Dutch, which consisted of polytomously scored items.

The examination data consisted of samples of candidates from the complete examination populations, the sample sizes are shown in the columns labeled N_{ref} and N_{new} of Table 8.4. The means and standard deviations of the observed frequency distributions of the examinations are shown in the columns labeled Mean and Std. For each design there were 5 linking groups. The total numbers of respondents in the linking groups are shown in the last column of Table 8.4 under the label N_{link}. Each linking group had approximately the same number of respondents. Every linking group made approximately the same number of items and every item in the design was presented to one linking group only.

One of the problems addressed here is whether the 1PLM and the 2PLM produced similar results. In Table 8.5, the results of the equating procedure are given for the version of the procedure where all distributions are estimated by their expected values. For each topic, four score points, were evaluated, r=20, 25, 30, 35. These scores are listed in the column labeled “r”. Further, for all examinations the actual cut-off score was

evaluated, in Table 8.5, the results pertaining to these scores are printed in boldface characters. The results obtained via the reference population are listed in the columns 3 to 5, the results obtained via the new population are listed in columns the 6 to 8. The third column, labeled "ϕ_R^1", contains the scores on the new examination associated with the scores of the reference examination computed using the 1PLM. So, for instance, score 20 on the reference examination language comprehension German at HAVO-D level is equated to a score 24 on the new examination, score 25 on the reference examination is equated to a score 29 on the new examination, etcetera. In the next column, labeled "ϕ_R^2", the resulting scores are given as they were obtained using the 2-PLM, so in this case a score 20 on the reference examination is equated to a score 25 on the new one. Notice that for a score 20 the 1PLM and the 2PLM are one score point off. Column 5, labeled" $\phi_R^1 - \phi_R^2$", contains the difference between the new scores obtained via the 1PLM and the 2PLM. For convenience, the sum of the absolute values of these differences is given at the bottom line of the table. So for the 32 scores equated here, the absolute difference in equated score points computed using the 1PLM and the 2PLM is 12 and the absolute difference between equated scores is never more than 2.

Table 8.5 Results of the Equating Procedure.

		Via reference population			Via new population			Procedures compared	
		GPCM	NRM		GPCM	NRM		GPCM	NRM
Topic	r	ϕ_R^1	ϕ_R^2	$\phi_R^1 - \phi_R^2$	ϕ_N^1	ϕ_N^2	$\phi_N^1 - \phi_N^2$	$\phi_R^1 - \phi_N^1$	$\phi_R^2 - \phi_N^2$
GermanD	20	24	25	−1	24	24	0	0	1
	25	29	30	−1	29	29	0	0	1
	30	34	34	0	34	34	0	0	0
	31	35	35	0	35	35	0	0	0
	35	38	38	0	38	38	0	0	0
GermanH	20	18	19	-1	18	19	−1	0	0
	25	24	24	0	24	24	0	0	0
	30	**29**	**29**	**0**	**29**	**29**	**0**	**0**	**0**
	35	34	34	0	34	34	0	0	0
EnglishD	20	19	21	−2	19	21	−2	0	0
	25	24	26	−2	24	26	−2	0	0
	28	**28**	**28**	**0**	**28**	**28**	**0**	**0**	**0**
	30	30	30	0	30	30	0	0	0
	35	35	35	0	35	35	0	0	0

EnglishH	20	21	21	0	21	21	0	0	0
	25	26	26	0	26	26	0	0	0
	27	**29**	**29**	**0**	**29**	**29**	**0**	**0**	**0**
	30	31	31	0	31	31	0	0	0
	35	36	36	0	36	36	0	0	0
FrenchD	20	21	22	−1	21	22	−1	0	0
	25	**26**	**26**	**0**	**26**	**26**	**0**	**0**	**0**
	30	31	31	0	31	31	0	0	0
	35	36	37	−1	36	36	0	0	1
FrenchH	20	19	19	0	19	19	0	0	0
	25	24	24	0	24	24	0	0	0
	30	**28**	**29**	**−1**	**28**	**29**	**−1**	**0**	**0**
	35	34	34	0	34	34	0	0	0
Abs.sum				12			11	0	4

The three following columns contain information comparable to the three previous ones, only now the scores were computed via the new population. Notice that the results obtained using the reference and the new population are much alike.

This is corroborated in the two last columns. These contain the differences in results obtained using either the reference or new population; the column labeled "$\phi^1_R - \phi^1_N$" shows the differences for the 1PLM and column labeled "$\phi^2_R - \phi^2_N$" shows the differences for the 2PLM.

Two conclusions can be drawn from Table 8.5. First, the 1PLM and the 2PLM do produce different results, but these differences are not spectacular: the sum of the absolute values of the differences given at the bottom of the table are 12 and 11 score points over all examinations and equated scores, and the absolute difference is never more than two score points. The second conclusion is that using either the reference or new population for determining the difference between the examination makes little difference, at the bottom of the table it is shown that the sum of the absolute values of the differences are 0 and 5 score points.

Above, it was mentioned that the procedure could be carried out in two manners: one where all relevant distributions are estimated by their expected values, and one where observed distributions are used as far as they are at hand. The above results refer to the former approach. Application of the second approach produced results that are far less satisfactory with respect to the population used for setting equivalent scores. That is, for the 1PLM, the summed differences between using the reference and the new population rose from 0 to 20, for the 2PLM, this difference rose from 5 to 35. In other words, the requirement of equating that an equating function should be invariant over populations and symmetric (see Petersen, Kolen & Hoover, 1989) is better met using expected frequencies only.

Confidence intervals

When a practitioner must set a cut-off score for some examination, the first question that comes to mind is about the reliability of the equating function. In the example of Table 8.3 a cut-score of 27 on the reference examination is equated with a cut-off score 29 on the new examination upon observing that the percentage 28.0 in the second column is closest to the percentage 29.7 in the third column. But to what extent can these percentages be relied upon? In Table 8.5, 90% confidence intervals are given for the estimated percentages on which equating is based. Their computation will be treated below. Consider the information on the English reading comprehension exam, which was also used for producing Table 8.2. In the boldface row labeled "English H" information is given on the results of the reference population making the reference examination. In the column labeled "Obs.Perc." the percentage of students scoring 27 or less is repeated, in the column labeled "Exp.Perc." its expected frequency under the 1PLM is given. The columns labeled "Lower Bound" and "Upper Bound" contain the bounds of the confidence interval of the latter estimate. Finally, in the columns marked "Obs.-Exp." and "Z", the difference between the observed and expected percentages and their normalized difference are given. This normalized difference was computed by dividing this difference by its standard error. This normalized difference can be seen as a very crude measure of model fit. Together with the plots of the frequency distributions given in Figures 8.8a and 8.8b, these differences give some indication of how well the model applies.

Continuing the example labeled "English H" in Table 8.6, in the three rows under the boldface row, for three scores, the estimates of cumulative percentages for the reference population confronted with the new exam and their confidence intervals are given. For all topics, these three scores are chosen in such a way that the middle score is the new cut-off score. In the columns under the label "Obs.-Exp." the differences between the observed and expected cumulative percentages are given. The widths of the confidence bands give an indication of the precision with which the observed and estimated percentages can be compared. For instance, in the "English H" example of Table 8.6, 28% is located well within the range of the confidence band related to score 29, while it is near the upper confidence bound related to score 28. If the observed percentage of the cut-off score is replaced by an estimated percentage, the confidence band of this estimate, which is given in the boldface row, also comes into play. But the question essentially remains the same:

Table 8.6 Confidence Intervals for Cumulative Percentages.

Topic	Score	Obs. Perc.	Lower Bound	Exp. Perc.	Upper Bound	Obs.-Exp.	Z
German D	**31**	**27.3**	**25.3**	**26.9**	**28.4**	**−0.4**	**−0.4**
	30	17.7	20.2	22.7	−7.0	−4.6	
	31	22.3	25.0	27.8	−2.2	−1.3	
	32	26.6	29.6	32.6	2.4	1.3	

German H	**30**	**23.1**	**22.4**	**23.4**	**24.5**	**0.3**	**0.5**
	28	16.5	19.5	22.5	−3.6	−2.0	
	29	20.5	23.9	27.2	0.8	0.4	
	30	25.2	28.9	32.5	5.8	2.6	
English D	**28**	**18.1**	**16.4**	**17.6**	**18.8**	**−0.5**	**−0.7**
	27	14.0	16.0	18.1	−2.0	−1.6	
	28	16.8	19.1	21.3	1.0	0.7	
	29	20.0	22.5	24.9	4.4	3.0	
English H	**27**	**28.0**	**28.3**	**29.7**	**30.8**	**1.6**	**2.0**
	28	22.9	25.6	28.4	−2.3	−1.4	
	29	26.6	29.5	32.4	1.5	0.9	
	30	30.5	33.6	36.7	5.7	3.0	
French D	**25**	**21.0**	**19.0**	**20.1**	**21.2**	**−0.9**	**−1.3**
	27	16.0	18.0	20.1	−2.9	−2.4	
	28	18.9	21.2	23.4	0.2	0.1	
	29	22.2	24.6	27.0	3.7	2.5	
French H	**30**	**22.4**	**20.6**	**21.8**	**22.9**	**−0.6**	**−0.8**
	28	15.9	19.0	22.1	−3.4	−1.8	
	29	19.2	22.6	26.0	0.2	0.1	
	30	22.9	26.6	30.2	4.2	1.9	

are the estimates precise enough to justify equating an old cut-off score with a unique new one, or are the random fluctuations such, that several new cut-off scores are plausible. Taking into account the cost of sampling linking groups, in the present example the precision reflected in Table 8.5 was considered sufficient.

The bootstrap method was used for computing the confidence intervals in the above example. The bootstrap method (Efron, 1979; Efron & Gong, 1983) entails repeated re-sampling with replacement from the original data. The sample size of these re-samples is the same as the size of the original sample and the probability of an element being sampled is the same for all response patterns in the original sample. By estimating the model parameters on every re-sample the standard error of the estimator can be evaluated.

8.2 Multiple Populations in IRT

8.2.1 Differences between populations

Suppose respondents are sampled from two populations, say males and females, and the interest is in evaluating the difference in the mean ability level of the two populations. A background variable gender is introduced as

$$x_i = \begin{cases} 1 & \text{if } i \text{ is a male} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

and gender differences in ability are modeled as

$$\theta_i = \mu + \beta x_i + e_i \qquad (2)$$

where it is assumed that e_i has a normal distribution with mean zero and variance σ^2. Note that μ is the mean ability level of the females, while $\mu+\beta$ is the mean of the males. So β is the effect of being male. In IRT, the assumption that the variance σ^2 is equal over groups is easily generalized to the assumption that groups have unique variances, say σ_g^2. Maximum marginal likelihood (MML) estimation is probably the most used technique for parameter estimation in IRT models (Bock & Aitkin, 1981; Thissen, 1982; Mislevy 1984, 1986; Glas & Verhelst, 1989). In this approach, a distinction is made between structural parameters, which need to be consistently estimated and nuisance parameters, which are not of primary interest. MML estimation derives its name from maximizing the log-likelihood that is marginalized with respect to the nuisance parameters. In the present case, the likelihood is marginalized with respect to the ability parameters θ. This leads to the marginal likelihood

$$L(\delta,\beta,\mu,\sigma;y,x) = \prod_i^I \int_{-\infty}^{\infty} \prod_k^K p(y_{ik} \mid \theta_i,\delta_k)^{d_{ik}}\, g(\theta_i \mid \mu,\beta,\sigma,x_i)\, d\theta_i \qquad (3)$$

where $g(\theta_i \mid \mu,\beta,\sigma,x_i)$ stands for the normal density. The reason for maximizing the marginal rather than the joint likelihood of all parameters is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of person parameters grows proportional with the number of observations, and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panchapakesan (1969) show that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that MML estimates of structural parameters, say the item and population parameters of an IRT model, are consistent under fairly reasonable regularity conditions, which motivates the general use of MML in IRT models.

Table 8.7 gives a small simulated example of the procedure. The data were generated with the 1PLM. The design consisted of 9 items administered to two groups. Group 1 consisted of 100 simulees, who responded to the items 1 to 6. The second group consisted of 400 simulees, responding to the items 4 to 9. So the items in the so-called “anchor” were responded to by 500 simulees. The true item parameters are shown in the second

column of Table 8.7, the MML parameter estimates and their standard errors are shown in the third and fourth column, respectively.

Table 8.7 Parameter Values and Estimates.

Item	b_k	b_k	$se(b_k)$
1	−1.00	−0.71	0.33
2	0.00	−0.04	0.30
3	1.00	1.18	0.29
4	−1.00	−1.10	0.14
5	0.00	−0.17	0.13
6	1.00	1.09	0.14
7	−1.00	−1.09	0.15
8	0.00	0.00	0.14
9	1.00	0.94	0.15
Pop	β	β	$se(\beta)$
1	1.00	1.07	0.22
Pop	σ_g	σ_g	$se(\sigma_g)$
1	1.00	1.13	0.18
2	1.50	1.45	0.10

Note that the standard errors are inversely proportional to the number of simulees responding to the item. The bottom lines of the table give the generating values for β, σ_g (g=1,2), their estimates and their standard errors. In this example, μ has been set equal to zero to identify the scale of θ. The test whether the two groups have the same mean ability level, that is, the test of the null hypothesis β=0 against the alternative $\beta \neq 0$, can be based on the ratio of the estimate with its standard error: $\hat{\beta} / se(\hat{\beta})$. In the present case, the outcome is 1.07/0.22=4.864. Under the null-hypothesis the statistic has a standard normal distribution so the nullhypothesis is clearly rejected.

This approach can be generalized in various ways. One could introduce a second variable, say

$$x_{2i} = \begin{cases} 1 & \text{if } i \text{ lives in an urban area} \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

and consider the model

$$\theta_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + e_i \qquad (5)$$

If x_{1i} stand for gender, then β_{12} stands for the interaction of being male and living in an urban area, and, as above, a test of the hypothesis $\beta_{12}=0$ against the alternative $\beta_{12}\neq 0$ can be based on the parameter estimate relative to its standard error. The next section gives further generalizations of this approach.

8.2.2 Multilevel regression models on ability

In much social research, elementary units are clustered in higher-level units. A wellknown example is educational research, where pupils or students are nested within classrooms, classrooms within schools, schools within districts and so on. Multilevel models (ML models) have been developed to take the resulting hierarchical structure into account, mostly by using regression-type models with random coefficients (Aitkin, Anderson & Hinde, 1981; Goldstein, 1986; Raudenbush & Bryk, 1986; Longford, 1987). However, if, variables in these multilevel models contain large measurement errors, the resulting statistical inferences can be very misleading (Fuller, 1987). Measurement error can be modeled in the framework of classical test theory (see, for instance, Longford, 1993) and IRT (Mislevy & Bock, 1989; Adams, Wilson & Wu, 1997; Fox & Glas, 2001, 2002). In the classical framework, the variance component due to unreliability, can either be imputed in the model, or it can be estimated within the model, for instance by splitting test scores tests into subtest scores. The IRT framework is a generalization of the linear model described in the previous section. The approach entails the definition of a multilevel linear model, where latent variables from IRT measurement models are entered either as dependent or as independent variables. The resulting model is the so-called multilevel IRT model (MLIRT model, Fox & Glas, 2001, 2002). The general model is defined as follows. The dependent variables are observed item scores yijk, where the index *i (i=1,…, n_j)* signifies the respondents, the index *j (j=1,…, J)* signifies the level two clusters, say the schools, and the index *k (k=1,…, K)* signifies the items. The first level of the structural multilevel model is formulated as

$$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij},\ldots,+\beta_{q'j}x_{q'ij} + \beta_{(q'+1)j}\xi_{(q'+1)ij} +,\ldots,+\beta_{Qj}\xi_{Qij} + e_{ij} \qquad (6)$$

where the covariates x_{qij} *(q=1,…, q')* are manifest predictors and the covariates ξ_{qij} *(q=q'+1,…, Q)* are latent predictors. Finally, e_{ij} are independent and normally distributed error variables with mean zero and variance σ^2. In general, it is assumed that the regression coefficients β_{qj} are random over groups, but they can also be fixed parameters. In that case, $\beta_{qj}=\beta q$ for all *j*. The Level 2 model for the random coefficients is given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}z_{1qj} +,\ldots,\gamma_{qs'}z_{s'qj} + \gamma_{q(s'+1)}\zeta_{(s'+1)qj} +,\ldots,+\gamma_{qS}\zeta_{Sqj} + u_{qj} \qquad (7)$$

where z_{sqj} *(s=1,…, s')* and ζ_{sqj} *(s=s'+1,…, S)* are manifest and latent predictors, respectively. Further, u_{qj} are error variables which are assumed independent over *j* and have a Q-variate normal distribution with a mean equal to zero and a covariance matrix *T*.

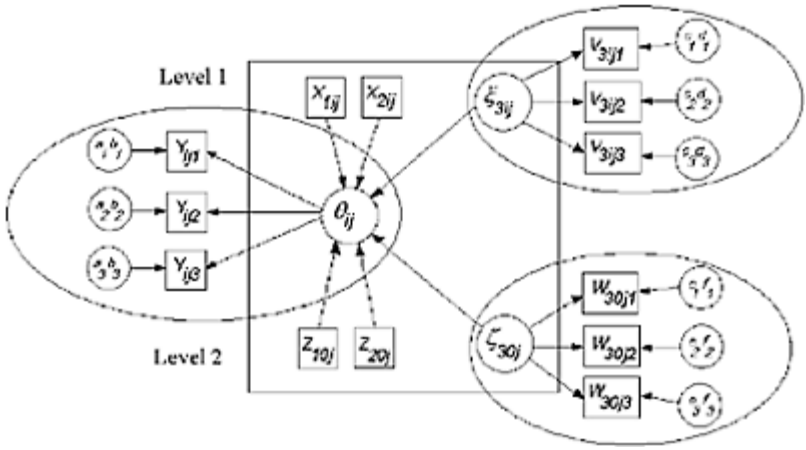


Figure 8.9 Path diagram of a multilevel IRT model.

An example of a MLIRT model is given in the path diagram in Figure 8.9. The structural multilevel part is presented in the big square box in the middle. The structural model has two levels: the upper part of the box gives the first level (a within-schools model), and the lower part of the box gives the second level (a between-schools model). The dependent variable θ_{ij}, say math ability, is measured by three items. The responses to these items are modeled by the 2PLM with item parameters a_k and b_k, $k=1,\ldots, 3$. Note that the measurement error models are presented by the ellipses. Both levels have three independent variables: two are observed directly, and one is a latent variable with three binary observed variables. For instance, on the first level, X_{1ij} could be gender, X_{2ij} could be age, and ξ_{3ii} could be intelligence as measured by a three item test. On the second level, Z_{10j} could be school size, Z_{20j} could be the school budget and ζ_{30j} could be a school's pedagogical climate, again measured by a three item test. In order not to complicate the model, it is assumed that only the intercept β_{0j} is random, so the Level 2 predictors are only related to this random intercept. So the slopes are fixed.

The parameters in the MLIRT model can be estimated in a Bayesian framework with a version of the Markov chain Monte Carlo (MCMC) estimation procedure: the Gibbs sampler (Fox & Glas, 2001, 2002). There are many considerations when choosing between a frequentist framework (such as MML) and Bayesian framework (such as MCMC), but the reason for adopting the Bayesian approach given by Fox and Glas (2001, 2002) is a practical one: MML involves integration over the nuisance parameters, and in the present case these integrals become quite complex. In the Bayesian approach, the interest is in the posterior distribution of the parameters, say $p(\theta, \delta, \beta, \mu, \sigma \mid y)$. In the MCMC approach samples are drawn from the posterior distribution of the parameters of interest, and in this process, nuisance parameters can play a role as auxiliary variables. So the problem of complex multiple integrals does not arise here.

To give some idea of the output of the procedure, consider an application reported Shalabi (2002). The data were a cluster sample of 3,384 grade 7 students in 119 schools.

At student level the variables were gender (0=male, 1=female), SES (with two indicators: the father's and mother's education, scores ranged from 0 to 8), and IQ (range from 0 to 80). At school level: leadership (measured by a scale consisting of 25 five-point Likert items, administered to the school teachers), school climate (measured by a scale consisting of 23 five-point Likert items) and mean-IQ (the IQ scores aggregated at school level). The items' scores for the leadership and climate variables were recoded to be dichotomous (0,1, and 2=0; 3, and 4=1). The dependent variable was a mathematics achievement test consisting of 50 multiplechoice items. The 2PLM was used to model the responses on the leadership and school climate questionnaire and the mathematics test. The parameters were estimated with the Gibbs sampler. For a complete description of all analyses, one is referred to Shalabi (2002); here only the estimates of the final model are given as an example. The model is given by

$$\theta_{ij} = \beta_{0j} + \beta_1 \mathrm{SES}_{ij} + \beta_2 \mathrm{Gender}_{ij} + \beta_3 \mathrm{IQ}_{ij} + e_{ij}$$

and

$$\beta_{0j} = \gamma_{00} + \gamma_{01} \mathrm{Mean\text{-}IQ}_j + \gamma_{02} \mathrm{Leadership}_j + \gamma_{03} \mathrm{Climate}_j + u_{0j}$$

The results are given in Table 8.8. The estimates of the MLIRT model are compared with a traditional multilevel analysis where all variables were manifest.

The observed mathematics, leadership and school climate scores were transformed in such a way that their scale was comparable to the scale used in the MLIRT model. Further, the parameters of the ML model were also estimated with a Bayesian approach using the Gibbs sampler. The columns labeled C.I. give the 90% credibility intervals of the point estimates; they were derived from the posterior standard deviation. Note that the credibility regions of the regression coefficients do not contain zero, so all coefficient can be considered significant at the 90% level. It can be seen that the magnitudes of the fixed effects in the MLIRT model were larger than the analogous estimates in the ML model. This finding is in line with the other findings (Fox & Glas, 2001, 2002; Shalabi, 2002), which indicates that the MLIRT model has more power to detect effects in hierarchical data where some variables are measured with error.

Table 8.8 Estimates of the Effects of Leadership, Climate and Mean IQ.

	MLIRT estimates		ML estimates	
	Estimates	C.I.	Estimates	C.I.
γ_{00}	−1.096	−2.080 − −.211	−0.873	−1.20 − −0.544
β_1	0.037	0.029–0.044	0.031	0.024–0.037
β_2	0.148	0.078 −0.217	0.124	0.061 −0.186
β_3	0.023	0.021–0.025	0.021	0.019–0.022
γ_{01}	0.017	0.009–0.043	0.014	0.004–0.023
γ_{02}	0.189	0.059− 0.432	0.115	0.019 −0.210

γ_{03}	−0.136	−0.383 − − 0.087	−0.116	−0.236–0.004
Variance components				
$\tau 0^2$	0.177	0.120–0.237	0.129	0.099–0.158
σ^2	0.189	0.164–0.214	0.199	0.190–0.210

PART 4
Monitoring the Effectiveness of Educational Systems

Introduction to Part 4

This part starts out with a chapter that provides definitions and conceptualization of education indicators and indicator systems. This chapter also contains examples of indicators at system level and at school level. The next two chapters refer to the research literature on school effectiveness as a basis for the further modeling of indicator systems. In Chapter 10 a basic conceptualization is provided, whereas Chapter 11 reviews the empirical evidence from studies carried out in industrialized and developing countries in more detail. Chapter 12 provides a more detailed list of variables and items that can be used to measure educational context, input and process indicators inspired by the school effectiveness research literature. Chapter 13 specifically looks at the issue of “value added” performance indicators.

9
Conceptualization of Education Indicators at System and at School Level

9.1 Introduction

As stated in Chapter 2, educational indicators are statistics that allow for value judgements to be made about key aspects of the functioning of educational systems. To emphasize their evaluative nature, the term "performance indicator" is frequently used.

Included in this definition of educational indicators are:

- the notion that we are dealing with measurable characteristics of educational systems;
- the aspiration to measure "key aspects", be it only to provide an "at a glance profile of current conditions" (Nuttall, 1989) rather than in-depth description;
- the requirement that indicators show something of the quality of schooling, which implies that indicators are statistics that have a reference point (or standard) against which value-judgements can be made.

Usually policymaking at national level is seen as the major source of application of indicators (indicator systems as policy-information systems). This view on the application of indicators should be enlarged, however, since consumers and "third parties" like private industry, are also seen as users of the information that indicator systems provide. Likewise, the education system at local administrative level and even individual schools could also use indicators to support policymaking (indicator systems as management information systems).

During the last decade various types of collections of indicators, usually referred to as indicator-systems, have been proposed and a part of these have also been developed and actually used. Van Herpen (1989) gives a comprehensive overview of what he calls "conceptual models of educational indicators". For our purpose it is sufficient to discern some major developments in these various approaches to conceptualizing education indicator systems.

Economic and social indicators are the origin of educational indicators. "Social indicators of education" describe educational aspects of the population, whereas educational indicators describe the performance of the educational system (Van Herpen, 1989, p. 10). The first trend in the development of educational indicators was the transition from descriptive statistics to measuring performance, or, more generally, a shift towards statistics of evaluative importance.

When looking at developments in educational indicators at the National Center for Statistics of the US Department of Education, a second trend can be discerned. At first the Center offered descriptive statistics on the state of the educational system, including

data on inputs and resources. Since 1982, "outcome" and "context" data were given a more prominent place, and in a proposal to redesign the education data system, "process" aspects of the functioning of educational systems were also included (Stern, 1986; Taeuber, 1987). This second trend can thus be characterized as a movement towards more comprehensive indicator *systems*, first adding output measures and context measures to the more traditional measurement of inputs and resources, and secondly by a growing interest in "manipulative input factors" and process-characteristics.

The third trend is somewhat related to the second one, as far as the interest in process characteristics is concerned. Traditionally indicator systems have concentrated on macro-level data, such as national illiteracy rates, the proportion of pupils that have passed their final secondary examinations, school etc. When we think of process-indicators as referring to the procedures or techniques that determine the transition of inputs into outputs, interest in process-indicators naturally leads to an interest in what goes on in schools. So, the third trend in conceptualizing indicator systems is to measure data at more than one aggregation level (national system, school, perhaps even the classroom), for examples see Scheerens et al. 1988; Taeuber, 1987).

Implicit in the above is the notion that a context-input-process-output model as introduced in Chapter 2, and somewhat further elaborated in Figure 9.1, is the best analytical scheme to systemize thinking on education indicators. In the next section this basic scheme will be related to the classifications used in World Bank documents and the OECD-INES project. In the ensuing sections a closer look will be taken at education process indicators, particularly when these are used within a context of program evaluation.

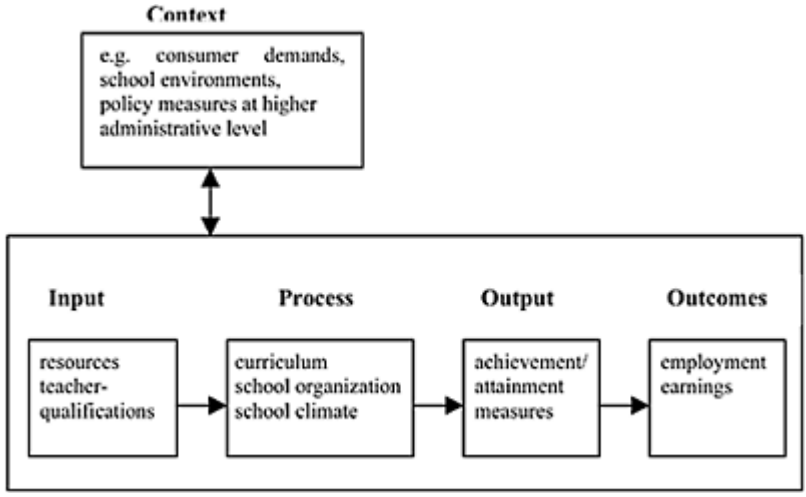


Figure 9.1 Context-input-process-output-outcome model of schooling.

9.2 Classifications

In the brochure titled 'Performance monitoring indicators. A handbook for task managers' by the Operations Policy Department of the World Bank (1996) an elaborated framework for the use of performance indicators is presented. The "handbook" presupposes a detailed rational planning process of projects. This includes a *"problem and beneficiary analysis"* (e.g. needs analyses, stakeholder motivations, whether the problem requires external development assistance), an *"objective analysis"* (statement of objectives, analysis of means to attain objectives, identification of target groups and planning in terms of time and location) and a *"finalization of project design and indicators"*. ("In this step planners carefully examine the project to ensure that all its elements are logically related. Planners also assess the integrity of indicators and realism of targets at this stage, taking into account all project assumptions and baseline data, and finalize their plans", ibid, p. 12).

Particularly this last step underlines the position that the identification of indicators is to be seen as an integral part of project planning. The "Handbook" uses an elaborate conceptual framework in which three main types of indicators are distinguished: *risk indicators, direct indicators* and *efficacy indicators*. A "free" interpretation of these three main categories is that risk indicators examine the larger context of projects[1], direct indicators are about inputs, outputs, outcomes, impact and "relevance" of the project while efficacy indicators further analyze the total set of indicators with respect to the criteria efficiency, effectiveness and sustainability of impacts. Figure 9.2, cited from the "Handbook", p. 13, illustrates the framework.

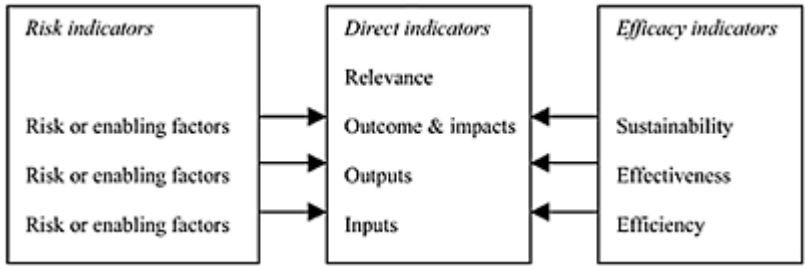


Figure 9.2 Categorization of indicators, World Bank, 1996.

The "direct" indicators closely fit the general classification presented in Figure 9.1, with the exception of "context" indicators which in the Handbook's framework fit in the risk indicator main category and the missing out of process indicators.

The way the four types of direct indicators are defined is as follows:

"*Input indicators* measure the quantity (and sometimes the quality) of resources provided for project activities" (e.g. the human resources included in the implementation unit)—ibid, p. 11.

"*Output indicators* measure the quantity (and sometimes the quality) of the goods or services created or provided through the use of inputs" (e.g. "clients vaccinated", "miles of roads built").

"*Outcome and impact indicators* measure the quantity and quality of the results achieved through the provision of project goods and services" (e.g. "reduced incubation of disease", "improved farming practices").

"*Relevance indicators"* refer to "intended outcomes on higher-order objectives that are not captured by direct outcome indicators" (e.g. improved national health care, increased farm profits and reduced food costs).

When comparing these definitions to the use of the terms process, output and outcome indicators in Figure 9.1 it is clear that the two sets of definitions can be "mapped" on to each other as depicted in Figure 9.3.

Basic systems model, Figure 9.1	**"Handbook" definitions, see Figure 9.2**
process indicators	output indicators
output indicators	outcome indicators
outcome indicators	impact indicators
(not covered)	relevance indicators

Figure 9.3 Two classifications compared, basic systems model and the "Handbook for Task Managers".

In other World Bank Publications, e.g. Carvalho & White, 1996, definitions of indicator types are somewhat like the ones on the left hand side in Figure 9.3.

"*Input* indicators measure the "means" by which projects are implemented".

"*Process* indicators measure the extent to which the project is delivering what it is intended to deliver" ("In a primary education project process indicators would be: the number of schools rehabitated, the pupil/desk ratio, the pupil textbook ratio, the pupil/exercise book ratio, and the volume of library services" (ibid, p. 9).

"*Impact* indicators measure the project's impact upon the living standards of the poor in a borrowing country".

This latter qualification differs from the systems model interpretation in Figure 1 in the sense that what is called "process indicators" by Carvalho & White is similar to education *input* indicators in Figure 9.1.

The conclusion so far is that there are important semantic differences between classification schemes of indicators, and that it would be desirable to reach more uniformity in basic terminology. A point that goes further than mere semantics is the fact that in the World Bank documents that were referred to the notion of process indicators as referring to core transition processes or "throughput" (transition of inputs into outputs over time) is altogether absent.

[1] "Risk indicators measure the status of the exogenous factors identified as critical through the risk and sensitivity analysis (risks and enabling factors) performed as part of a project's economic analysis" (ibid, p. 14).

The OECD Education Indicators project (INES—see the "Education at a Glance" publications) uses a more substantive categorization, which is evident from the table of contents in the Education at a Glance Publications (OECD, 1999).

The main categories are:

A) The demographic, social and economic context of education (e.g. Literacy skills of the adult population)
B) Financial and human resources invested in education (e.g. Educational expenditure per student)
C) Access to education, participation and progression (e.g. Overall participation in formal education
D) The transition from school to work (e.g. Youth unemployment and employment by level of educational attainment)
E) The learning environment and the organization of schools (e.g. total intended instruction time for pupils in lower secondary education)
F) Student achievement and the social and labor-market outcomes of education (e.g. Mathematics achievement of students in 4th and 8th grades, and Earnings and educational attainment)

These 6 categories can be classified in various ways. The context-input-processoutcome scheme is the first way to do so. Accordingly category A is in the context domain, category B refers to inputs, categories C, D and E refer to different interpretations of the process dimension, and category F to an output/outcome dimension. See Figure 9.4.

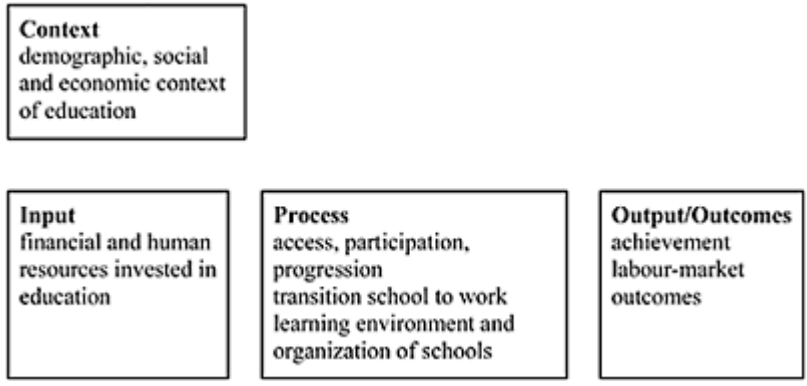


Figure 9.4 Ordering of the OECD-INES education indicator set, according to a context-input, process and outcome scheme.

In Figure 9.4 arrows between the boxes have been omitted since, only in a very loose sense, these categories are expected to be linked in a causal way. In fact each category is used in a descriptive sense and interrelationships between indicators have hardly been analyzed so far.

A second way to look upon the OECD indicator set is by distinguishing “stock” and “flow” types of indicators. Categories A and B are typically stock indicators, whereas categories C and D, and to some extent F, refer to flows.

A stock indicator describes a relevant educational aspect at one point in time in quantitative and qualitative terms (e.g. the number of qualified teachers in school year x). A flow indicator refers to the transition of an educational unit, e.g. a student or a teacher, to a different part of the system (e.g. the number of teachers that have left the profession in school year y).

It should be noted that classifying “transition” or “flow” indicators as process indicators adds a third interpretation to process indicators. To summarize we have to far encountered three types of process indicators:

a. as transformation processes (see the systems model in Fig. 9.1);
b. as checks on program implementation (Carvalho & White, 1996);
c. as flows of units through the educational system (OECD).

Before saying more about process indicators the conceptualization of educational indicators will be elaborated by examining evaluative contexts, aggregation levels and the time dimension.

9.3 Evaluative Contexts, Aggregation Levels and the Time Dimension; Towards Further Conceptualization of Education Indicators

9.3.1 Evaluative contexts

There are three different evaluative contexts in which education indicators can be used. Sometimes indicators can be used for more than one context of application at the same time:

a. Monitoring the state of education at national or district level
b. Program evaluation
c. School self evaluation

The way the OECD indicators are used is an example of monitoring at the national system level with the interesting added advantage of international comparative information, which could be used as “benchmarks”.

To the extent that loans from International Lender and Support Organizations in the education sector are used for system-broad reforms or reforms in complete subsectors, like primary or secondary education, program evaluation would largely coincide with monitoring at systems level. A simple design for the evaluation of such large-scale reforms would be *two* “inventories” of the education sector, one immediately before and one after program implementation. It could be remarked in passing that international comparison might offer interesting possibilities for the evaluation of projects of International Lending Organizations, to the extent that the nature, context and timeframe of projects in different countries would be comparable.

To the extent that education indicators are based on data collected at lower aggregation level than the national system, namely at the level of schools, teachers and pupils, they can even be used for purposes of school self evaluation. A simple example is feeding back information to schools, whereby schools could then compare their own position on certain indicators to national averages or other standards.

9.3.2 Aggregation levels

Educational systems have a hierarchical structure where administrative levels are "nested". Indicator systems usually ignore this hierarchical structure by using statistics that are defined at national level or formal characteristics of the system. Examples are: pupil teacher ratio computed as the ratio of all pupils and all teachers in a country and teacher salaries defined on the basis of nationally determined salary-scales. Even when considering use of indicators at national level only, there are two main advantages to use data at lower aggregation levels:

- disaggregate data allows for examining variation between units, e.g. the variance between schools in success rates on examinations;
- disaggregate data allows for better adjustments and more valid causal inferences; the best example in education is the use of so called "value-added" performance indicators based on achievement test scores adjusted for prior achievement and/or other relevant pupil background characteristics (also see Chapter 13).

When it is the intention to relate, for example, school organizational characteristics to pupil achievement, disaggregate data at pupil level is required to carry out appropriate multi-level analyses.

Particularly when indicators are used for program evaluation purposes, the above mentioned advantages of disaggregate data are important, because they provide firmer ground to answer causal questions about program effectiveness.

A final added advantage is that the relevance of indicator systems for lower administrative levels (e.g. school districts and individual schools) grows when disaggregate data is available.

9.3.3 Timeframe

For the purposes of evaluating World Bank programs experimental and quasiexperimental designs have been proposed (Ezemenari, Rudqvist & Subbarao, 1998).

Although there is no question about it that (quasi-)experimental designs should be used whenever possible (compare Campbell's famous idea of "Reforms as Experiments", Campbell, 1969), they are often not feasible.

Using educational indicators in a longitudinal way, whereby the same units are measured at several points in time, is a viable alternative to experimentation.

9.4 The Function of Educational Process Indicators

In previous sections various interpretations of educational process indicators were referred to. In this section process indicators that reflect malleable conditions of basic transformation processes in education will be placed central (see Figure 1). School organizational functioning and teaching and learning at classroom level are examples of such educational transformation processes.

In general it could be said that such process indicators shed some light on what happens in the "black box" of schooling. Process indicators are interesting from the point of view of policy and management since they refer to conditions that are malleable and thus the subject of active policies to improve education.

In a later section the perspective of school effectiveness research will be presented as the most likely rationale for identifying and selecting process indicators. Accordingly those process indicators will be selected that show positive associations with educational output and outcomes. Ideally such process indicators should be able to predict output (as in "education production functions": instruments in "process" conditions predicting increments in output according to an exact function). To the extent that such instrumental knowledge would be complete process indicators could rightly be used as substitutes of output indicators. Given the fact that the education production function is debated and, more generally, school effectiveness knowledge is "incomplete" to say the least such a strong instrumental interpretation is not realistic.

This leaves two further possibilities for the use of process indicators:

- as "annex" to output indicators, whereby in each and every situation of their use the association between process and output indicators would have to be explored with the intention to "explain" differences in outcomes between schools and between educational systems;
- a weaker interpretation of instrumentality, where process indicators are seen as instances of educational good practice, and, in this way, could lead to valuejudgements about educational quality even in the absence of output data.

Within the context of program evaluation process indicators are sometimes defined as checks on the actual implementation of the program. This interpretation of process indicators is easily reconcilable with the one used throughout this chapter. Implementation checks are a more basic and administrative type of monitoring, whereas process indicators as defined above, are referring to more generic causal processes of organizational functioning and teaching and learning. When process indicators are used over and above implementation checks, they say more about *why* an (implemented) program works. Figure 9.5 illustrates this. When program evaluation as compared to "monitoring" is the evaluative context both types of process indicators could be used next to each other.

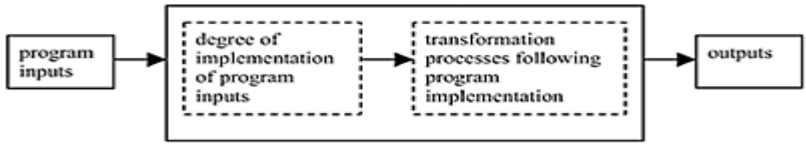


Figure 9.5 Implementation checks and process indicators.

9.5 A First Overview of Education Indicators

9.5.1 System level formulation

Formulation at system level is dealt with first. In Figure 9.6 the overall framework used in the OECD-INES project is shown as an example of system level application.

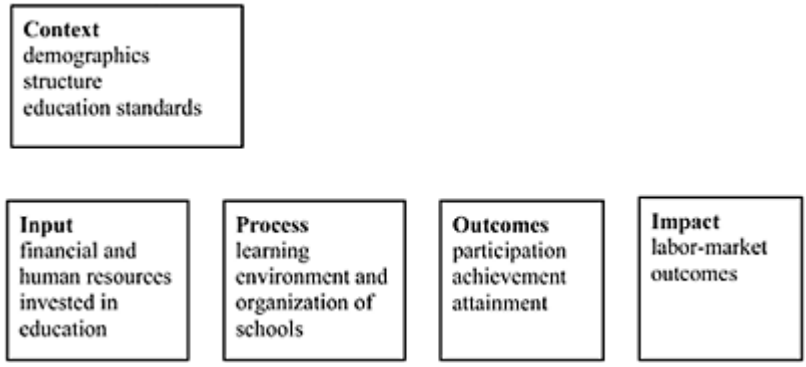


Figure 9.6 Categorization of system-level education indicators.

Categories of indicators are defined according to the position in the model.

Context indicators (defined at the level of national education systems) refer to characteristics of the society at large and structural characteristics of national education systems. Examples are:

- demographics; e.g. the relative size of the school-age population;
- basic financial and economic context; e.g. the GDP per capita;
- education goals and standards by level of education; e.g. higher completion rates, more equitable distribution of university graduates;
- the structure of schools in the country, as characterized by means of the International Standard Classification of Education (ISCED).

Input indicators at system levels refer to financial and human resources invested in education. Examples are:

- expenditure per student;
- expenditure on Research and Development in Education;
- the percentage of the total labor force employed in Education;
- pupil teacher ratios per education level;
- characteristics of the stock of "human resources", in terms of age, gender, experience, qualifications and salaries of teachers.

Process indicators at system level are characteristics of the learning environment and the organization of schools that are either defined at system level or based on aggregated data collected at lower levels. Examples are:

- the pattern of centralization/decentralization or the "functional decentralization" specified as the proportion of decisions taken in a particular domain that is taken by a particular administrative level;
- priorities in the intended curriculum per education level, expressed, for example, as the teaching time per subject;
- priorities in the education reform agenda, expressed, for example as the proportion of the total education budget to specific reform programs;
- investments and structural arrangements for system level monitoring and evaluation at a given point in time.

Output or outcome indicators at system level refer to statistics on access and participation, attainment statistics and aggregated data on educational achievement. Examples are:

- participation rates in the various education levels (primary, secondary and tertiary);
- progression through the education system, expressed for example of the proportion of students that gets a diploma in the minimum formal time that is available for a program;
- drop out rates at various levels of the education system;
- average achievement in basic curricular domains, for example in subjects like mathematics, science, literacy, measured at the end of primary and/or secondary school;
- cross curricular competencies and "life skills", like problem solving, basic literacy and social skills.

Impact or long-term outcome indicators refer to changes in other sectors of the society that can be seen as the effects of education. Examples are:

- the impact of education on youth unemployment, e.g. by categorizing youth unemployment by level of educational attainment;
- the position of school leavers with a certain level of certification on the labor market;
- income related to education and training level;
- delinquency rate per level of educational attainment.

9.6 School Level Formulation of Educational Indicators

As described in the chapter on international developments, the inclusion of *process* indicators led to an interest in indicators at the school and classroom level. Including the school as the central unit in sets of indicators also led to a different interpretation of context, input and even outcome indicators. From this perspective context indicators refer to conditions in the immediate environment of the school, like for example the press for achievement from an external body, like an external school board or the municipality. Input indicators are the specific financial, material and human resources as defined for each and every individual school. At school level background characteristics of the students, like aptitudes and socioeconomic status form a particular type of input characteristics. As there is likely to be an interest in the performance of schools, irrespective of the innate abilities and background of the students, such input indicators are used to determine what is known as the *added value* of schooling. In order to determine this added value "gross" output indicators are adjusted for these background characteristics of the students. This is done by means of particular statistical analyses.

In selecting process indicators at school level reference is usually made to what is known from research on school effectiveness and educational productivity. This body of research has yielded a set of factors that is positively associated with relatively high performance of schools, in other words it indicates "what works" in education. In subsequent chapters this research literature will be discussed in more depth.

Following the structure as depicted in Figure 9.1 examples of the various types of indicators are provided below. The specific orientation of this categorization of indicators can be summarized as follows:

- processes are defined at school and classroom level;
- context and input indicators are seen as having a direct influence on the school level;
- output indicators may take the form of value added school performance, e.g. achievement that is adjusted for differences in intake characteristics of pupils between schools.

School context indicators are conditions from the school environment that are expected to stimulate school performance. Examples are:

- achievement stimulants from higher education levels, e.g. whether or not specific achievement standards are set by the municipality or the school district;
- consumer demands, e.g. whether or not parents have free choice of schools;
- community involvement; for example the amount of discretion that local school boards have concerning the conditions of labor of teachers;
- parental involvement in school matters, measured , for example as the degree of actual involvement of parents in various school activities (the teaching and learning process, extra-curricular activities and support activities).

School input indicators refer to the financial, human and material resources of a particular school. Examples are:

- school financial resources, for example the per pupil expenditure for a particular school;

- teacher qualifications and experience, for example the proportion of qualified, under-qualified and over-qualified teachers in a particular school;
- class size; measured as the average of the number of students per class;
- school managerial "overhead" which can be measured as the proportion of full time equivalent teaching staff that is deployed for other than teaching activities;
- facilities and equipment of the school, for example the student/computer ratio.

Process indicators of school functioning refer to malleable conditions of schooling and instruction, i.e. those conditions that are under the control of the school's management and staff. Sub-categories refer to the curriculum, the school's policy and mission, leadership, criteria of organizational effectiveness, climate and instructional conditions. Examples are given below.

Achievement oriented school policy

- whether or not schools set achievement standards;
- the degree to which schools follow (education) careers of pupils after they have left the school;
- whether or not schools report achievement/attainment outcomes to local constituencies.

Educational leadership

- the amount of time principals spend on educational matters, as compared to administrative and other tasks;
- whether or not principal's appraise the performance of teachers;
- the amount of time dedicated to instructional issues during staff meetings.

Continuity and consensus among teachers

- the amount of changes in staff over a certain period;
- the presence or absence of school subject-related working groups or departments (secondary schools);
- frequency and duration of formal and informal staff meetings.

Orderly and safe climate

- statistics on absenteeism and delinquency;
- ratings of school discipline by principals, teachers and pupils.

Efficient use of time

- total instruction time and time per subject matter area;
- average loss of time per teaching hour (due to organization, moving to different rooms, locations, disturbances);
- percentage of lessons "not given", on an annual basis.

Curriculum and opportunity to learn

- the time per subject as indicated in the school's timetable (intended school curriculum);
- teacher or student ratings of whether each item of an achievement test was taught or not (the implemented school curriculum).

Evaluation of pupils' progress

- the frequency of use of curriculum specific tests at each grade level;
- the frequency of use of standardized achievement tests;
- the actual use teachers make of test results.

Ratings of teaching quality

- quality of instruction as rated by peers (other teacher);
- quality of instruction as rated by students.

Organizational effectiveness criteria

- teacher and student satisfaction;
- staff turnover.

School outcomes indicators are performance indicators measured at the end of a period of schooling, which may be adjusted for pupil intake characteristics. Examples are:

- student achievement results in basic subjects adjusted for prior achievement and socio-cultural or socioeconomic status;
- success rates at the end of a period of schooling;
- drop-out rates;
- proportion of students that go to specific categories of follow-up or further education; for example, from the perspective of primary or lower secondary schools, the proportion of students that goes to academic or vocational streams or programs in upper secondary education.

10
Perspectives on School Effectiveness[1]

10.1 Introduction

Starting out from common sense notions of an effective school being roughly the same as a "good" school the more precise meaning of school effectiveness used in empirical research studies is developed. Perspectives from various disciplines, most notably economics and organizational science, are used to render nuance and difference in focus. Despite various perspectives a relatively simple scheme consisting of a set of malleable conditions of schooling (causes) and a small range of types of criteria (effects) is considered as the core of the concept.

10.2 A General Definition

School effectiveness refers to the performance of the organizational unit called "school". The performance of the school is, most likely, expressed as the *output* of the school, which in its turn is measured by looking at the average achievement of the pupils at the end of a period of formal schooling. The question of school effectiveness is

interesting because it is well known that schools differ among themselves in performance. How much they differ is the next question and a more refined and precise version of this question is how much schools differ when they are more or less equal as far as the innate abilities and socioeconomic background of the pupils are concerned.

A somewhat different statement of the principle of "fair" comparison between schools is the aim to assess the *added value* of a period of schooling. This means assessing the impact of schooling on pupils' achievement that can be uniquely attributed to having attended school A as compared to school B. In school effectiveness research queries do not end by just assessing the "net" or value-added differences between schools. For this branch of educational research the really interesting questions only start after having established that there is significant variation. *Why* does school A do better than school B, if the differences are not due to differences in the student population of the two schools, is the issue.

. [1]Parts of this chapter are an updated version of Chapter 1 of J.Scheerens (1992). *Effective Schooling. Research Theory and Practice*, published by Cassell (London).

Different strands of educational effectiveness research have concentrated on different types of variables to answer this question. Economists have concentrated on resource inputs, such as per pupil expenditure. Instructional psychologists investigated classroom management, such as time on task and variables associated with instructional strategies. And general education experts and educational sociologists looked at school organizational conditions, such as leadership style

Before going on in explaining these different strands of educational effectiveness research and their subsequent integration into multi-disciplinary and multi-level educational effectiveness studies, a few basic characteristics of the emergent definition of school effectiveness should be highlighted.

It should be noted, first of all, that school effectiveness is an empty concept with respect to the kind of operational measures of school performance that are chosen. Since the literary meaning of effectiveness is *goal attainment* the implicit assumption is that performance measures reflect important educational objectives. Of course, opinions about what these are may differ, and consequently an easy line of attack on school effectiveness research is that it has failed to address important educational objectives. In actual practice achievement in basic subjects like arithmetic/mathematics, science and vernacular or foreign languages, are the effect measures chosen in the large majority of all strands of empirical educational effectiveness studies. Secondly, school effectiveness refers to comparative rather than absolute standards. “Effects” are expressed in terms of adjusted mean differences between schools or in terms of percentage of “explained” variation between schools. The implication is that school effectiveness studies, carried out within a particular national context, do not say anything about the actual level of educational achievement in that country. In terms of performance levels being an effective school in country X could mean something altogether different in country Y.

A final implicit aspect in the general description of school effectiveness and school effectiveness research to be noted is that it is a causal concept. Some authors therefore make an explicit difference between *school effectiveness* research on the one hand and *school effects* research on the other (cf. Purkey & Smith, 1983). In school effectiveness research not only are differences in overall performance assessed, but the additional question of causal attribution is raised: what school characteristics lead to relatively higher performance, after characteristics of the student populations have been held constant?

In summing up school effectiveness is seen as the degree of goal attainment schools realize, in comparison with other schools that are “equalized” in terms of student-intake, as a result of the values of certain conditions that are malleable by the school itself or the immediate school context.

10.3 Economic Definitions of Effectiveness

In economics concepts like effectiveness and efficiency are related to the production process of an organization. Put in a rather stylized form a production process can be summed up as a “turnover” or transformation of “inputs” to “outputs”. *Inputs* of a school or school system include pupils with certain given characteristics and financial and material aids. *Outputs* include pupil attainment at the end of schooling. The

transformation *process* or *throughput* within a school can be understood as all the instruction methods, curriculum choices and organizational preconditions which make it possible for pupils to acquire knowledge. Longer term *outputs* are denoted with the term *"outcomes"*, see Table 10.1.

Table 10.1 Analysis of Factors on the Education Production Process.

Inputs	**Process**	**Outputs**	**Outcomes**
Funding	Instruction methods	Final primary school test scores	Dispersal on the labor market

Effectiveness can now be described as the extent to which the desired level of output is achieved. *Efficiency* may then be defined as the desired level of output against the lowest possible cost. In other words, efficiency is effectiveness with the additional requirement that this is achieved in the cheapest possible manner. Cheng (1993) offers a further elaboration of the effectiveness and efficiency definitions, by incorporating the dimension of short term output versus long term outcomes. In his terms: *technical* effectiveness and technical efficiency refer to "school outputs limited to those in school or just after schooling (e.g. learning behavior, skills obtained, attitude change, etc.)". *Social* effectiveness and efficiency are associated with "effects on the society level or the life-long effects on individuals (e.g. social mobility, earnings, work productivity)" (ibid, p. 2). When crossing these two dimensions four types of school output are discerned; see Table 10.2.

It is vitally important for the economic analysis of efficiency and effectiveness that the value of inputs and outputs can be expressed in terms of money. For determining efficiency it is necessary that input costs like teaching materials and teachers' salaries are known. When the outputs can also be expressed in financial terms efficiency determination is more like a cost-benefit analysis (Lockheed, 1988, p. 4). It has to be noted, however, that a strict implementation of the above-mentioned economic characterization of school effectiveness runs up against many problems.

Table 10.2 Distinction Between School Effectiveness and School Efficiency, Cited From Cheng, 1993, p. 4.

	Nature of school output	
Nature of school input	In school/Just after schooling Short-term effects Internal (e.g. learning behavior, skills obtained)	On the society level Long term effects External (e.g. social mobility, earnings, productivity)
Non-monetary (e.g. teachers, teaching methods, books)	*School's Technical Effectiveness*	*School's Societal Effectiveness*

Monetary (e.g. cost of books, salary, opportunity costs)	*School's Technical Efficiency* (internal economic effectiveness)	*School's Societal Efficiency* (external economic effectiveness)

These already start with the question of how one should define the "desired output" of a school, even if we concentrate on the short term effects. For instance, the "production" or returns of a secondary school can be measured by the number of pupils who successfully pass their school-leaving diploma. The unit in which production is measured in this way is thus the pupil having passed his or her final examination. Often, however, one will want to establish the units of production in a finer way and will want to look, for instance, at the grades achieved by pupils for various examination subjects. In addition, there are all types of choices to be made with regard to the scope of effectiveness measures. Should only performance in basic skills be studied; is the concern also perhaps with higher cognitive processes and should not social and/or affective returns on education be established? Other problems related to economic analysis of schools are the difficulty in determining monetary value on inputs and processes and the prevailing lack of clarity on how the production process operates (precisely what procedural and technical measures are necessary to achieve maximum output).

Relevant to the question on how useful one regards the characterizing of effectiveness in economical terms is the acceptability of the school as a metaphor for a production unit.

10.4 Organization-Theoretical Views on Effectiveness

Organizational theorists often adhere to the thesis that the effectiveness of organizations cannot be described in a straightforward manner. Instead, a pluralistic attitude is taken with respect to the interpretation of the concept in question. By that it is assumed that it depends on the organization theory and the specific interests of the group posing the question of effectiveness, which interpretation will be chosen (Cameron & Whetten, 1983, 1985; Faerman & Quinn, 1985). The main perceptions on organization which are used as background for a wide range of definitions on effectiveness will be briefly reviewed.

10.4.1 Economic rationality

The already mentioned economic description of effectiveness is seen as deriving from the idea that organizations function rationally—that is to say, purposefully. Goals which can be operationalized as pursued outputs are the basis for choosing effect criteria (effect criteria are the variables by which effects are measured, i.e. student achievement, well-being of the pupils etc.). There is evidence of economic rationality whenever the goals are formulated in the sense of outputs of the primary production process of the school. In the entire functioning of a school other different goals can also play a part, such as having a clear-cut policy with regard to increase the number of enrolments. Also with regard to this type of objective a school can operate rationally, only this falls outside the specific interpretation given to economic rationality. Effectiveness as defined in terms of economic rationality can also be identified as the productivity of an organization. In

education the rational or goal-orientated model is mainly propagated via Tyler's model that can be used for both curriculum development and educational evaluation (Tyler, 1950). From the remaining perceptions on organization, to be discussed shortly, the economic rationality model is dismissed as being both simplistic and out of reach. It is well-known in the teaching field how difficult it is to reach a consensus on goals and to operationalize and quantify these. From the position that other values besides productivity are just as important for organizations to function, the rational model is regarded as simplistic.

10.4.2 The organic system model

According to the organic system model, organizations can be compared to biological systems which adapt to their environment. The main characteristic of this approach is that organizations openly interact with their surroundings. Thus, they need in no way be passive objects of environmental manipulation but can actively exert influence on the environment themselves. Nevertheless, this viewpoint is mainly preoccupied with an organization's "survival" in a sometimes hostile environment. For this reason, organizations must be flexible, namely to assure themselves of essential resources and other inputs. Therefore, according to this viewpoint flexibility and adaptability are the most important conditions for effectiveness in the sense of survival. A result of this could be that the effectiveness of a school is measured according to its yearly intake, which could partly be attributed more or less to intensive canvassing or schoolmarketing.

No matter how remarkable this view on effectiveness may seem at first glance, it is nevertheless supported by an entirely different scientific sphere—microeconomics of the public sector. Niskanen (1971) demonstrated that public sector organizations are primarily targeted at maximizing budgets and that there are insufficient external incentives for these organizations—schools included—to encourage effectiveness and efficiency. In this context it is interesting to examine whether canvassing activities of schools mainly comprise of displaying acquired facilities (inputs) or presenting output data like previous years' examination results.

Finally, it should also be mentioned that it is conceivable that the inclination towards inputs of the organic system model coincides with a concern for satisfying outputs, namely in those situations where the environment makes the availability of inputs dependent on quantity and/or quality of earlier realized achievements (output).

10.4.3 The human relations approach of organizations

If in the open system perception of organizations there is an inclination towards the environment, with the so-called human relations approach the eye of the organization analyst is explicitly focused inward. This fairly classical school of organizational thought has partly remained intact even in more recent organizational characterization. In Mintzberg's concept of the professional bureaucracy, aspects of the human relations approach reoccur, namely in emphasizing the importance of the well-being of the individuals in an organization, the importance of consensus and collegial relationships as well as motivation and human resource development (Mintzberg, 1979). From this perception, job satisfaction of workers and their involvement with the organization are

likely criteria for measuring the most desired characteristics of the organization. The organizational theorists who share this view regard these criteria as effectiveness criteria.

10.4.4 The bureaucracy

The essential problem with regard to the administration and structure of organizations, in particular those like schools which have many relatively autonomous sub-units, is how to create a harmonious whole. For this appropriate social interaction and opportunities for personal and professional development—see the human relations approach—provide a means. A second means is provided by organizing, clearly defining and formalizing these social relations. The prototype of an organization in which positions and duties are formally organized is the "bureaucracy". According to this perspective certainty and continuity concerning the existing organization structure is the effectiveness criterion. It is well-known that bureaucratic organizations tend to produce more bureaucracy. The underlying motive behind this is to ensure the continuation, or better still, the growth of one's own department. This continuation can start operating as an effect criterion in itself.

10.4.5 The political model of organizations

Certain organizational theorists see organizations as political battlefields (Pfeffer & Salancik, 1978). Departments, individual workers and management staff use the official duties and goals in order to achieve their own hidden—or less hidden agendas. Good contacts with powerful outside bodies are regarded as very important for the standing of their department or of themselves. From a political perspective the question of the effectiveness of the organization as a whole is difficult to answer. The interest is more for the extent by which internal groups succeed in complying with the demands of certain external interested parties. In the case of schools these bodies could be school governing bodies, parents of pupils and the local business community.

It has already been mentioned that organizational concepts on effectiveness not only depend on theoretical answers to the question of how organizations "are pieced together" but also on the position of the factions posing the effectiveness question. On this point there are differences between these five views on organizational effectiveness. With regard to the economic rationality and the organic system model, the management of the organization is the main "actor" posing the effectiveness question. As far as the other models are concerned, department heads and individual workers are the actors that want to achieve certain effects.

In the table below the chief characteristics of the organization-theoretical perceptions on effectiveness are summarized.

Table 10.3 Organizational Effectiveness Models.

theoretical background	effectiveness criterion	level at which the effectiveness	main areas of attention

		question is asked	
(business) economic rationality	productivity	organization	output and its determinants
organic system theory	adaptability	organization	acquiring essential inputs
human relations approach	involvement	individual members of the organization	motivation
bureaucratic theory; system members theory; social psychological homeostatic theories	continuity	organization + individual	formal structure
political theory on how organizations work	responsiveness to external stakeholders	subgroups and individuals	independence, power

The diversity of views on effectiveness which organizational theory makes leads to the question which position should be taken. Should we indeed operate from a position of there being several forms of effectiveness, should a certain choice be made, or is it possible to develop from several views, one all-embracing concept on effectiveness?

For a discussion on these questions the reader is referred to Scheerens, 1992 and Scheerens and Bosker, 1997. From the perspective of educational planning in developing countries the most probable and fruitful positions appears to be the one where productivity, in terms of quantity and quality of school output, is seen as the ultimate criterion and the other criteria are seen either as pre-conditions (responsiveness) or "means" (criteria referring to organizational conditions such at teacher satisfaction). In applied use of the school effectiveness knowledge base, such as the design and use of monitoring and evaluation systems, to be discussed in subsequent chapters, the broader organizational perspective on effectiveness can serve at the conceptual background for the development of education indicators.

10.5 Modes of Schooling, as Points of Impact for Attaining Effectiveness

In the previous section it was established that the overall concept of school effectiveness may be differentiated according to normative criteria related to various schools of thought in organizational science. These schools of thought led to a discussion about the choice of criteria or types of "effects" to be measured. Bearing in mind that school effectiveness is a causal concept, the dimension of *causes* or *means* should be considered as well, next to the type of effects.

In doing so the question that is dealt with concerns the *distinction of all possible features of the functioning of schools that are malleable in order to reach the effects that are aimed for*. Such a broad perspective is needed to obtain as complete a picture as

possible on elements and aspects of schooling and school functioning that are potentially useable in improving effectiveness.

According to well-known distinctions in organizational science (e.g. Mintzberg, 1979; De Leeuw, 1982) the following categories can be used as a core framework to further distinguish elements and aspects of school functioning:

- goals
- the structure of positions or sub-units ("Aufbau")
- the structure of procedures ("Ablauf")
- culture
- the organization's environment
- the organization's primary process

These antecedent conditions will be referred to as *modes of schooling*. Modes are considered as conditions that, in principle, are malleable by the school itself or by outside agencies that have control over the school. The overall effectiveness equation, consisting of antecedent conditions on the one hand and effects on the other can be depicted as in Figure 10.1.

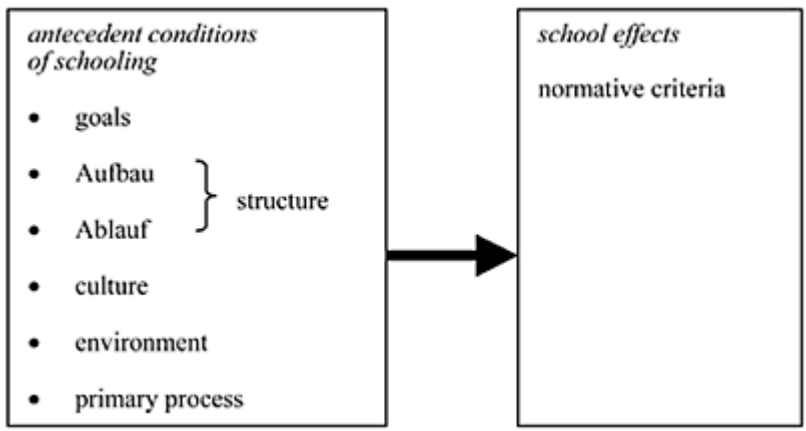


Figure 10.1 Schematic representation of school effectiveness.

Among the modes goals have a specific role. In organizational effectiveness thinking goals can be seen as the major defining characteristic of the effectiveness concept itself. In the previous section it was established that different goal areas, or effectiveness criteria, can be used to operationally apply effectiveness assessment.

When "goals" are not taken as "given" in effectiveness assessment, but as options, or directions the organization can choose, this further emphasizes the relativity of the organizational effectiveness concept. The question whether an organization chooses the "right" goals or objectives can be seen as a fundamental question that proceeds the question of instrumental rationality, concerning the attainment of "given" objectives. In this respect the well-known distinction between "doing the right things" and "doing

things right" is at stake. In its turn the question of the "rightness" of a particular choice of organizational goals can be seen as instrumental to meeting the demands of stakeholders in the external environment of the organization. In the case of schools, for instance, these may be demands from the local community or from parents' associations.

Further options of choice with respect to goals are:

- various priorities in further specification of the overall goals (in the case of schools, for instance, the relative priority of cognitive versus non-cognitive objectives and the relative emphasis on basics versus "other" subjects);
- the levels or standards of goal attainment that are striven for: to the degree that schools are relatively autonomous they may set absolute standards, that every pupil should achieve or they may adapt achievement standards to the initial achievement level of pupils;
- whether or not attainment levels are differentiated for different sub-groups of pupils.

Finally, it can be seen as an assignment of organizations to ensure that goals or attainment targets are shared among the members of the organizations. This is particularly relevant for organizations like schools, where teachers traditionally have a lot of autonomy. In control-theory the phenomenon of unifying the goals of organizational sub-units (i.e. departments and individual teachers, in the case of schools) is known as "goal coordination".

Table 10.4 Modes of Schooling.

Goals	
• goals in terms of various effectiveness criteria	
• priorities in goal specifications (cognitive—non-cognitive)	
• aspirations in terms of attainment level and distribution of attainment	
• goal coordination	
Aufbau (position structure)	
• management structure	
• support structure	
• division of tasks and positions	
• grouping of teachers and students	
Ablauf (structure of procedures)	
• general management	
• production management	planning
• marketing management	coordinating
• personnel management (among which hrm, hrd)	controlling
• financial & administrative management	assessing
• cooperation	

Culture

• indirect measures

• direct measures

Environment

• routine exchange (influx of resources, delivery of products)

• buffering

• active manipulation

Primary process

• curricular choices

• curriculum alignment

• curriculum in terms of pre-structuring instructional process

• pupil selection

• levels of individualization and differentiation

• instructional arrangements in terms of teaching strategies and classroom organization

It is beyond the scope of this book to discuss the various modes of schooling in detail. Table 10.4 provides a schematic overview of the most important subcategories. A more detailed presentation is provided in Scheerens and Bosker, 1997, Chapter 1.

"Pupil selection" is a condition that would generally fall outside the definition of school effectiveness, since the specific interest in the value that is added by schooling over and above the impact of the innate abilities of pupils precludes the consideration of this option. Yet, depending on the regulations determined by higher administrative units, it is definitely a condition that schools may be able to manipulate. Selectivity referring to a way of regulating education that can be seen as the most important competitor to the philosophy that schooling makes a difference through dedication of leadership and staff and through the choice of superior technology.

The sub-set of modes of schooling that has been the focus in empirical school effectiveness research, will be treated more fully in the next chapter, where the results of various strands of educational effectiveness research are summarized. Running ahead of this presentation it can be said that empirical school effectiveness research has concentrated on production management, co-operation, aspects of culture and all sub-categories of the primary process. The fuller set of modes, derived from organization theory, is considered useful to indicate as complete a picture as possible of conditions that may be used as points of impact for school improvement.

10.6 Summary and Conclusions

In this chapter delineating the conceptual map of school effectiveness started out by referring to economic definitions of effectiveness. Comparisons with economic

definitions of effectiveness and efficiency pointed out that the bulk of current empirical school effectiveness research considers the relationship between nonmonetary inputs and short term outputs, in Cheng's (1993) terminology: technical effectiveness.

Organization theoretical approaches to organizational effectiveness indicated a range of models, each emphasizing a different type of criteria to judge effectiveness, with productivity, adaptability, involvement, continuity and responsiveness to external stakeholders, as the major categories. Comparison of this range of effectiveness criteria to the implicit model used in most empirical school effectiveness studies showed that the productivity criterion is the predominant criterion in actual research practice. This position can be legitimized from the point of view of a means to end ordering of the criteria, with productivity taken as the ultimate criterion (Scheerens, 1992). Such a position is contested, however, by other authors who see the criteria as "competing values" (Faerman & Quinn, 1985), or who opt for a more dynamic interpretation where the predominance of any single criterion would depend on the organization's stage of development (Cheng, 1993).

In recognizing that effectiveness is essentially a causal concept, in which means to end relationships have a similar meaning as cause-effect relationships, there are in fact three major components in the study of organizational effectiveness:

- the range of effects;
- the points of impact of actions to attain particular effects (indicated as modes of schooling);
- functions and underlying mechanisms that explain why action impinged on certain modes lead to effect-attainment.

In this chapter modes of schooling were described while using the following main categories of organization's anatomy as a basic framework:

- goals
- organization structure, both with respect to the structure of positions, and the structure of procedures (including management functions)
- culture
- environment
- primary process/technology

Each of these main categories was treated as an area that, in principle, can be manipulated or influenced by the school or an external change agent. When comparing the list of modes to the current practice of empirical school effectiveness research, it appeared that the structure of procedures (particularly school management), culture, and instructional conditions have received most of the attention.

Van Kesteren (1996, p. 94) includes most of the pluriformity of perspectives that have been discussed in this chapter in his definition of organizational effectiveness:

> *"Organizational effectiveness is the degree to which an organization, on the basis of competent management, while avoiding unnecessary exertion, in the more or less complex environment in which it operates, manages to control internal organizational and environmental conditions, in order to provide, by means of its own characteristic transformation process, the*

> *outputs expected by external constituencies"* (translated from Van Kesteren, 1996, p. 94).

As is clear from this definition, as form the overall discussion in this chapter, school effectiveness is primarily seen as an issue for individual schools (school management perspective). At the same time research considers *schooling* and factors that are, generalized over individual schools, associated with relatively high "value-added" performance. Depending on the patterns of centralization and decentralization in a country (which may be different for different domains of educational functioning, like the curriculum or finance) above-school administrative levels or other constituencies have discretion over some of the effectiveness enhancing conditions. From the perspective of educational planning at the national level it is important to take this issue of functional (de)centralization into consideration. For example, it should be decided, depending on overall policy and structural and cultural conditions, whether or not key-effectiveness enhancing modes of schooling are left completely "free" to the individual schools, or central stimulation measures would be preferred.

11
A Review of the Research Evidence on School Effectiveness, from Developed and Developing Countries

11.1 Introduction; the Overall Design of Educational Effectiveness Studies

The elementary design of school effectiveness research is the association of hypothetical effectiveness enhancing conditions of schooling and output measures, mostly student achievement. A basic model from systems theory, where the school is seen as a black box, within which processes or "throughput" take place to transform this basic design. The inclusion of an environmental or context dimension completes this model (see Fig. 11.1). The major task of school effectiveness research is to reveal the impact of relevant input characteristics on output and to "break open" the black box in order to show which process or throughput factors "work", next to the impact of contextual conditions. Within the school it is helpful to distinguish a school and a classroom level and, accordingly, school organizational and instructional processes.

Research tradition in educational effectiveness varies according to the emphasis that is put on the various antecedent conditions of educational outputs. These traditions also have a disciplinary basis. The common denominator of the five areas of effectiveness research that will be distinguished is that in each case the elementary design of associating outputs or outcomes of schooling with antecedent conditions (inputs, processes or contextual) applies. The following research areas or research traditions will be considered in summarizing the research results obtained in developed countries:

1. Research on equality of opportunities in education and the significance of the school in this.
2. Economic studies on education production functions.
3. The evaluation of compensatory programs.
4. Studies of unusually effective schools.
5. Studies on the effectiveness of teachers, classes and instructional procedures.

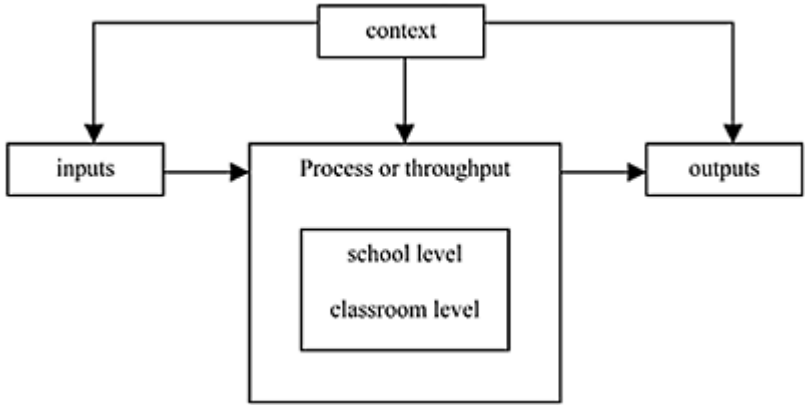


Figure 11.1 A basic systems model of school functioning.

In developing countries there is a strong predominance of studies of the education production function type. Relatively few of these have been expanded by including school organizational and instructional variables.

PART 1:

EVIDENCE FROM INDUSTRIALIZED COUNTRIES

11.2 Results Obtained in Various Strands of Educational Effectiveness Research

re 1) School effectiveness in equal educational opportunity research

Coleman's research into educational opportunity, about which a final report known as the Coleman report was published in 1966, forms the corner-stone for school effectiveness studies (Coleman et al., 1966). While this study was intended to show the extent to which school achievement is related to students' ethnic and social background, the possible influence of the "school" factor on learning attainment was also examined.

In the survey three clusters of school characteristics were measured: (a) teacher characteristics; (b) material facilities and curriculum; and (c) characteristics of the groups or classes in which the pupils were placed. After the influence of ethnic origin and socioeconomic status of the pupils had been statistically eliminated, it appeared that these three clusters of school characteristics together accounted for 10 percent of the variance in pupil performance. Moreover, the greater part of this 10 percent variance was due to the third cluster that was operationalized as the average background characteristics of pupils, which means that again the socioeconomic and ethnic origin—now defined at the

level of the school—played a central role. In reactions to the Coleman report there was general criticism on the limited interpretation of the school characteristics. Usually, only the material characteristics were referred to, such as the number of books in the school library, the age of the building, the training of the teachers, their salaries and expenditure per pupil. Nevertheless there were other characteristics included in Coleman's survey, such as the attitude of school heads and teachers towards pupils and the attitude of teachers towards integrated education, i.e. multiracial and classless teaching.

Other large-scale studies that were primarily focused at providing data on equality of opportunity, like the one by Hauser, Sewell and Alwin (1976), also indicated a relatively high correlation between socioeconomic and ethnic family characteristics and learning attainment, and a small or even negligible influence from school and instruction characteristics. The outcomes were criticized by educationalists for the rather narrow choice of school characteristics and on methodological grounds (cf. Aitkin & Longford, 1986), for multi-level associations not being properly modeled and analyzed.

re 2) Economic studies on educational production functions

The focus of economic approaches towards school effectiveness is the question of what manipulative inputs can increase outputs. If there was stable knowledge available on the extent to which variety of inputs is related to variety of outputs it would also be possible to specify a function which is characteristic of the production process in schools—in other words, a function which could accurately indicate how a change in the inputs would affect the outputs.

This leads to a research-tradition that is identified both by the term input-output studies as by the term research into education production functions. The research model for economics-related production studies hardly differs from that for other types of effectiveness research: the relationship between manipulative school characteristics and attainment is studied while the influence of background conditions like social class and pupils' intelligence is eliminated as far as possible. The specific nature of production-function research is the concentration on what can be interpreted in a more literal sense as input characteristics: the teacher/pupil relationship, teacher training, teacher experience, teachers' salaries and expenditure per pupil. In more recent observations of this research type one comes across the suggestion to take effectiveness predictors known from educational psychology research into account (Hanushek, 1986). It should be noted that the Coleman-report (Coleman et al. 1966) is often included in the category of input-output studies. In view of its emphasis on the more material school characteristics, the association is an obvious one.

The findings of this type of research have often been referred to as being disappointing. Review studies like the one by Hanushek (1986) tend to produce the same conclusions: inconsistent findings throughout the entire available research and scant effect at most from the relevant input variables.

From reanalysis of Hanushek's (1986) data-set, Hedges et al. (1994), however, conclude that there is an effect of per pupil expenditure of "considerable practical importance" (an increase of PPE by $510 would be associated with a 0.7 s.d. increase in student outcome).

But this conclusion in its turn is contested by Hanushek.

From Hanushek's, 1997, "vote count" overview of education production functions it appears that "Teacher pupil ratio", out of 277 estimates had 15% positive significant associations with student achievement. For teacher education, teacher experience, teacher salary and expenditure these percentages were respectively 9% out of 171 analyses, 29% out of 207 studies, 20% out of 119 studies and 27% out of 163 studies.

Hanushek's interpretation of these results is that there can be little confidence that adding more of any of the specific resources or, for that matter of the financial aggregates, will lead to a boost in student achievement. The variable that shows relatively the highest proportion of positive effects is teacher experience, but here, "reverse causation" could be at play, since more experienced teachers might have selected schools with better performing pupils (ibid, p. 144).

In other reviews, e.g. Verstegen and King (1998), a more positive interpretation is given on largely the same set of studies that was analyzed by Hanushek (1997). During the last decade several studies drew attention to the fact that certain resource input factors did show significant positive associations with pupil achievement or other educational outcomes. The most important of these are the studies by Card & Krueger (1992), which indicated a positive association between school resources and differences in earnings among workers, Hedges, Laine and Greenwald (1994) who conducted a statistical meta-analysis on a sub-set of Hanushek's 1979 data set and found significant effects for several resource input variables, among which a rather large positive effect of Per Pupil Expenditure, Ferguson (1991), who found particularly large effects of variables related to teacher qualifications (specifically scores on a teacher re-certification test), and Achilles (1996) who reported the sustained effects of reduced class-size (14–16 as compared to 22–24) in Kindergarten and the first three grades of primary school) on student achievement.

That these differences in interpretation are to a certain degree of the kind: "the cup is half full" as compared to "the cup is half empty" is illustrated by Verstegen & King's (1998) representation of Hanushek's, (1997) findings. By only reporting statistically significant positive and statistically significant negative estimates by omitting the large proportions of studies showing insignificant results, and "blowing up" the relatively small numbers of studies showing significant results to percentages, these authors appear to be keen to see (or construct) the bright side of things. Unfortunately, as in other types of educational effectiveness studies, the critics and those who present the more conservative interpretation appear to have the best arguments. Hanushek, 1997, presents most of them:

- when outcome measures, such as student achievement scores are properly adjusted for student background characteristics, and "value added" outcome indicators are used, the number of positive effects declines;
- if data at high aggregation levels (e.g. individual states) is used misspecification bias is likely to produce overstatement of effects (this criticism would apply to both the Ferguson and Card & Krueger studies). This problem frequently occurs for the variable Per Pupil Expenditure which is usually only defined at the district level;
- in statistical meta-analysis the null-hypothesis that is addressed is that resources or expenditure differences never, under whatever circumstances, affect student performance; clearly this hypothesis is to be rejected also in cases where only a minority of studies shows a significant positive association with the outcome variable.

Many of the recent contributions to summarizing the research evidence on education production function studies mention the need to search for answers to the question "why money does or does not matter", for example by looking for combinations and interactions between resource input levels and school organizational and instructional variables. In a recent collection of articles on class size (Galton, 1998) reference is made to differences between educational cultures in the degree to which large classes are considered a burden to teachers.

Another desirable extension of the basic education production function type of study would be to address questions of cost-effectiveness more directly, by comparing cost-effectiveness or even cost-benefit ratio's for different policy measures. A comparison of education production function studies between industrialized and developing countries is particularly interesting, since a "restriction of range" phenomenon (little variance in, for example, teacher salaries between schools) might suppress the effects in relatively homogenous school systems. Results of education production function studies in developing countries will be presented in a subsequent section.

re 3) The evaluation of compensatory programs

Compensatory programs may be seen as the active branch in the field of equal educational opportunity. In the United States compensatory programs like "Head Start" were part of President Johnson's "war on poverty". Other large-scale American programs were "Follow-Through"—the sequel to Head Start—and special national development programs that resulted from Title 1 of the Elementary and Secondary Education Act, enacted in 1965. Compensatory programs were intended to improve the levels of performance of the educationally disadvantaged. In the late sixties and early seventies there were also similar programs in the Netherlands like the Amsterdam Innovation project, the Playgroup Experiment project, Rotterdam's Education and Social Environment (OSM) project and the Differentiated Education project (GEON) of the city of Utrecht.

Compensatory programs manipulate school conditions in order to raise achievement levels of disadvantaged groups of pupils. The level in which this is achieved demonstrates the importance of the school factor—and in particular the conditions and educational provisions within it.

However, it proved to be not that simple to redress the balance with effective compensatory programs. In fact no overwhelming successes could be established. There was heated debate on the way available evaluation studies should be interpreted.

The key question is: what results can be realistically expected from compensatory education given the dominant influence in the long run of family background and cognitive aptitudes on pupils' attainment level? Scheerens (1987, p. 95) concluded that the general image provided by the evaluation of compensatory programs reveals that relatively small progress in performance and cognitive development can be established immediately after a program finishes. Long-term effects of compensatory programs cannot be established by and large. Moreover, it has been occasionally demonstrated that it was the "moderately" disadvantaged in particular that benefited from the programs, while the most educationally disadvantaged pupils made the least progress, relatively speaking.

In view of the variety of compensatory programs the evaluation studies gave some insight into the relatively best type of educational provision. When comparing the various components of Follow Through, programs aimed at developing elementary skills like language and mathematics and which used highly structured methods turned out to be winners (Stebbins et al., 1977; Bereiter & Kurland, 1982; Haywood, 1982). More recent corroboration of this conclusion is evident from the evaluation of a structured program on elementary reading in the United States "Success for All" (Slavin, 1996).

As will appear later, there is a remarkable similarity between these characteristics and the findings of other types of effectiveness research. In any case, when interpreting the results of evaluations of compensatory programs one should be aware that the findings have been established among a specific pupil population: very young children (infants or first years of junior school) from predominantly working-class families.

re 4) Effective schools research

Research known under labels like "identifying unusually effective schools" or the "effective schools movement" can be regarded as the type of research that most touches the core of school effectiveness research. In Coleman's and Jencks' surveys the inequality of educational opportunity was the central problem. In economicrelated input-output studies the school was even conceived as a "black box". In the still to be discussed research on the effectiveness of classes, teachers and instruction methods, education characteristics on a lower aggregation level than the school are the primary research object.

Effective school research is generally regarded as a response to the results of studies like Coleman's and Jencks' from which it was concluded that schools did not matter very much when it came down to differences in levels of achievement. From titles such as "Schools can make a difference" (Brookover et al., 1979) and "School matters" (Mortimore et al., 1988) it appears that refuting this message was an important source of inspiration for this type of research. The most distinguishing feature of effective schools research was the fact that it attempted to break open the "black box" of the school by studying characteristics related to organization, form and content of schools.

The results of the early effective schools research converged more or less around five factors:

- strong educational leadership;
- emphasis on the acquiring of basic skills;
- an orderly and secure environment;
- high expectations of pupil attainment;
- frequent assessment of pupil progress.

In the literature this summarizing is sometimes identified as the "five-factor model of school effectiveness". It should be mentioned that effective schools research has been largely carried out in primary schools, while at the same time studies have been largely conducted in inner cities and in predominantly working-class neighborhoods.

In more recent contributions effective schools research became more integrated with education production function and instructional effectiveness research, in the sense that a mixture of antecedent conditions was included, studies evolved from comparative case-

studies to surveys and conceptual and analytical multi-level modeling took place to analyze and interpret the results. Numerous reviews on school effectiveness have been published since the late seventies. Examples are Purkey and Smith (1983) and Ralph and Fenessey (1983). More recent reviews are those by Levine and Lezotte (1990), Scheerens (1992), Creemers (1994), Reynolds et al. (1993), Sammons et al. (1995), and Cotton (1995).

The focal point of interest in the reviews is the "what works" question; typically the review presents lists of effectiveness enhancing conditions.

There is a fairly large consensus on the main categories of variables that are distinguished as effectiveness enhancing conditions in the reviews, also when earlier and more recent reviews are compared.

Table 11.1 summarizes the characteristics listed in the reviews by Purkey and Smith (1983), Scheerens (1992), Levine and Lezotte (1990), Sammons et al. (1995), Cotton (1995).

Consensus is largest with respect to the factors:

- achievement orientation (which is closely related to "high expectations");
- co-operation;
- educational leadership;
- frequent monitoring;
- time, opportunity to learn and "structure" as the main instructional conditions.

Table 11.1 Effectiveness Enhancing Conditions of Schooling in Five Review Studies (italics in the column of the Cotton study refers to sub-categories).

Purkey & Smith, 1983	**Levine & Lezotte, 1990**	**Scheerens, 1992**	**Cotton, 1995**	**Sammons, Hillman & Mortimore, 1995**
Achievementoriented policy	Productive climate and culture	Pressure to achieve	Planning and learning goals	Shared vision and goals
Cooperative atmosphere, orderly climate		Consensus, cooperative planning, orderly atmosphere	Curriculum planning and development	A learning environment, positive reinforcement
Clear goals on basic skills	Focus on central learning skills		Planning and learning goals *school wide emphasis on learning*	Concentration on teaching and learning
Frequent evaluation	Appropriate monitoring	Evaluative potential of the	Assessment (district, school,	Monitoring progress

		school, monitoring of pupils' progress	classroom level)	
In-service training/staff development Strong leadership	Practiceoriented staff development Outstanding leadership	Educational leadership	*Professional development* collegial learning School management and organization, leadership and school improvement, leadership and planning	A learning organization Professional leadership
Time on task, reinforcement, streaming	Salient parent involvement Effective instructional arrangements	Parent support Structured, teaching, effective learning time, opportunity to learn	Parent community involvement Classroom management and organization, instruction	Home school partnership Purposeful teaching
High expectations	High expectations		Teacher student interactions	High expectations Pupil rights and responsibilities

Purkey & Smith, 1983	**Levine & Lezotte, 1990**	**Scheerens, 1992**	**Cotton, 1995**	**Sammons, Hillman & Mortimore, 1995**
			Distinct-school interactions Equity Special programs	
		External stimuli to make schools effective Physical and material school characteristics Teacher experience School context characteristics		

Behind this consensus on general characteristics hides considerable divergence in the actual operationalization of each of the conditions. Evidently concepts like "productive, achievement-oriented climate and educational leadership are complex concepts and individual studies may vary in the focus that different elements receive.

re 5) Studies on instructional effectiveness

As the most relative strands of research on teaching and classroom processes for the topic at hand are studies on characteristics of effective teachers, and studies that go under the label of "process-product studies". This latter category of studies was also inspired by Carroll's (1963) model of teaching and learning and off-springs of this model, such as the models of mastery learning (Bloom, 1976) and "direct teaching" (e.g. Doyle, 1985).

The research results have been reviewed by, among others, Stallings (1985), Brophy and Good (1986), and Creemers (1994) and quantitatively synthesized in meta-analyses by Walberg (1984), Fraser et al. (1987) and Wang, Haertel and Walberg (1993). These latter authors incidentally have also included variables outside the classroom situation, like the student's relationships with peers, and the home environment (e.g. television viewing) in their analyses which they label under the heading of "educational productivity".

In the sixties and seventies the effectiveness of certain personal characteristics of teachers was particularly studied. Medley and Mitzel (1963), Rosenshine and Furst (1973) and Gage (1965) are among those who reviewed the research findings. From these it emerged that there was hardly any consistency found between personal characteristics of the teacher like warmheartedness or inflexibleness on the one hand, and pupil achievement on the other. When studying teaching styles (Davies, 1972), the behavioral repertoire of teachers was generally looked at more than the deeply-rooted aspects of their personality. Within the framework of "research on teaching" there followed a period in which much attention was given to observing teacher behavior during lessons. The results of these observations, however, in as far as they were related to pupil achievement, seldom revealed a link with pupil performance (see Lortie, 1973, for instance). In a following phase more explicit attention was given to the relation between observed teacher behavior and pupil achievement. This research is identified in the literature as "process-product studies". Variables which emerged "strongly" in the various studies were the following (Weeda, 1986, p. 68):

1. *Clarity:* clear presentation adapted to suit the cognitive level of pupils.
2. *Flexibility:* varying teaching behavior and teaching aids, organizing different activities etc.
3. *Enthusiasm:* expressed in verbal and non-verbal behavior of the teacher.
4. *Task related and/or businesslike behavior:* directing the pupils to complete tasks, duties, exercises etc. in a businesslike manner.
5. *Criticism:* much negative criticism has a negative effect on pupil achievement.
6. *Indirect activity:* taking up ideas, accepting pupils' feelings and stimulating self-activity.
7. *Providing the pupils with an opportunity to learn criterion material*—that is to say, a clear correspondence between what is taught in class and what is tested in examinations and assessments.
8. Making use of *stimulating* comments: directing the thinking of pupils to the question, summarizing a discussion, indicating the beginning or end of a lesson, emphasizing certain features of the course material.
9. *Varying the level* of both cognitive questions and cognitive interaction.

In later studies effective teaching time became a central factor. The theoretical starting points of this can be traced back to Carroll's teaching-learning model (Carroll, 1963). Chief aspects of this model are:

- actual net learning time which is seen as a result of: perseverance and opportunity to learn;
- necessary net learning time as a result of: pupil aptitude, quality of education and pupil ability to understand instruction.

The mastery learning model formulated by Bloom in 1976 was largely inspired from Carroll's model, and the same goes for the concept of "direct teaching".

Doyle (1985) considers the effectiveness of direct teaching, which he defines as follows:

1. Teaching goals are clearly formulated.
2. The course material to be followed is carefully split into learning tasks and placed in sequence.
3. The teacher explains clearly what the pupils must learn.
4. The teacher regularly asks questions to gauge what progress pupils are making and whether they have understood.
5. Pupils have ample time to practice what has been taught, with much use being made of "prompts" and feedback.
6. Skills are taught until mastery of them is automatic.
7. The teacher regularly tests the pupils and calls on the pupils to be accountable for their work.

The question whether this type of highly structured teaching works equally well for acquiring complicated cognitive processes in secondary education as for mastering basic skills at the primary school level was answered in the affirmative (according to Brophy & Good, 1986, p. 367). However, progress through the subject matter can be taken with larger steps, testing need not be so frequent and there should be space left for applying problem-solving strategies flexibly. Doyle (ibid) emphasized the importance of varying the learning tasks and of creating intellectually challenging learning situations. For the latter an evaluative climate in the classroom, whereby daring to take risks even with a complicated task is encouraged, is a good means.

In the domain of classroom organization Bangert, Kulik & Kulik's meta-analysis (1983) revealed that individual teaching in secondary education hardly led to higher achievement and had no influence whatsoever on factors like the self-esteem and attitudes of pupils. "Best-evidence-syntheses" by Slavin (1996, p. 57) indicated a significantly positive effect of co-operative learning at the primary school level.

Meta-analyses by Walberg (1984) and Fraser et al. (1987) found the highest effects for the following teaching conditions:

- reinforcement
- special programs for the educationally gifted
- structured learning of reading
- cues and feedback
- mastery learning of physics
- working together in small groups

It should be noted that more recently developed cognitive and particularly constructivist perspectives on learning and instruction challenge the behavioristically oriented approach and results of the process-product research tradition (Duffy & Jonassen, 1992; Brophy, 1996). According to the constructivist approach independent learning, meta-cognition (e.g. learning to learn), "active learning", learning to model the behavior of experts ("cognitive apprenticeship") and learning from real life situations ("situated cognition") should be emphasized. The effectiveness of teaching and learning according to these principles has not been firmly established as yet. Authors who have addressed this issue (Scheerens, 1994; De Jong & Van Joolingen, 1998) however, point out that a straightforward comparison with more structured teaching approaches may be complicated, since constructivist teaching emphasizes different, more higher order cognitive objectives. Moreover, structured versus "active" and "open" teaching had probably be better conceived as a continuum of different mixes of structured and "open" aspects, rather than as a dichotomy.

11.3 Integration

Of the five effectiveness-oriented educational research types, which were reviewed, two focused on "material" school characteristics (such as teacher salaries, building facilities and teacher/pupil ratio). The results were rather disappointing in that no substantial positive correlations of these material investments and educational achievement could be established in a consistent way across individual studies. On the basis of more recent studies these rather pessimistic conclusions have been challenged, although methodological critique indicates that the earlier pessimistic conclusions are more realistic. In-depth process studies connected with large-scale evaluations of compensatory programs pointed out that programs which used direct, i.e. structured, teaching approaches were superior to more "open" approaches. The research movement known as research on exemplary effective schools (or briefly: effective schools research) focused more on the internal functioning of schools than the earlier tradition of input-output studies.

These studies produced evidence that factors like strong educational leadership, emphasis on basic skills, an orderly and secure climate, high expectations of pupil achievement and frequent assessment of pupil progress were indicative of unusually effective schools.

Research results in the field of instructional effectiveness are centered around three major factors: effective learning time, structured teaching and opportunity to learn in the sense of a close alignment between items taught and items tested. Although all kinds of nuances and specificities should be taken into account when interpreting these general results they appear to be fairly robust—as far as educational setting and type of students is concerned. The overall message is that an emphasis on basic subjects, an achievement-oriented orientation, an orderly school

Table 11.2 General Characteristics of Types of School Effectiveness Research.

		independent variable type	**dependent variable type**	**discipline**	**main study type**
a.	(un)equal opportunities	socioeconomic status and IQ of pupil, material school characteristics	attainment	Sociology	Survey
b.	production functions	material school characteristics	achievement level	Economics	Survey
c.	evaluation compensatory programs	specific curricula	achievement level	interdisciplinary pedagogy	quasiexperiment
d.	effective schools	"process" characteristics of schools	achievement level	interdisciplinary pedagogy	case-study
e.	effective instruction	characteristics of teachers, instruction, class organization	achievement level	educational psychology	Experiment observation

environment and structured teaching, which includes frequent assessment of progress, is effective in the attainment of learning results in the basic school subjects. Table 11.2 summarizes the main characteristics of the five research traditions.

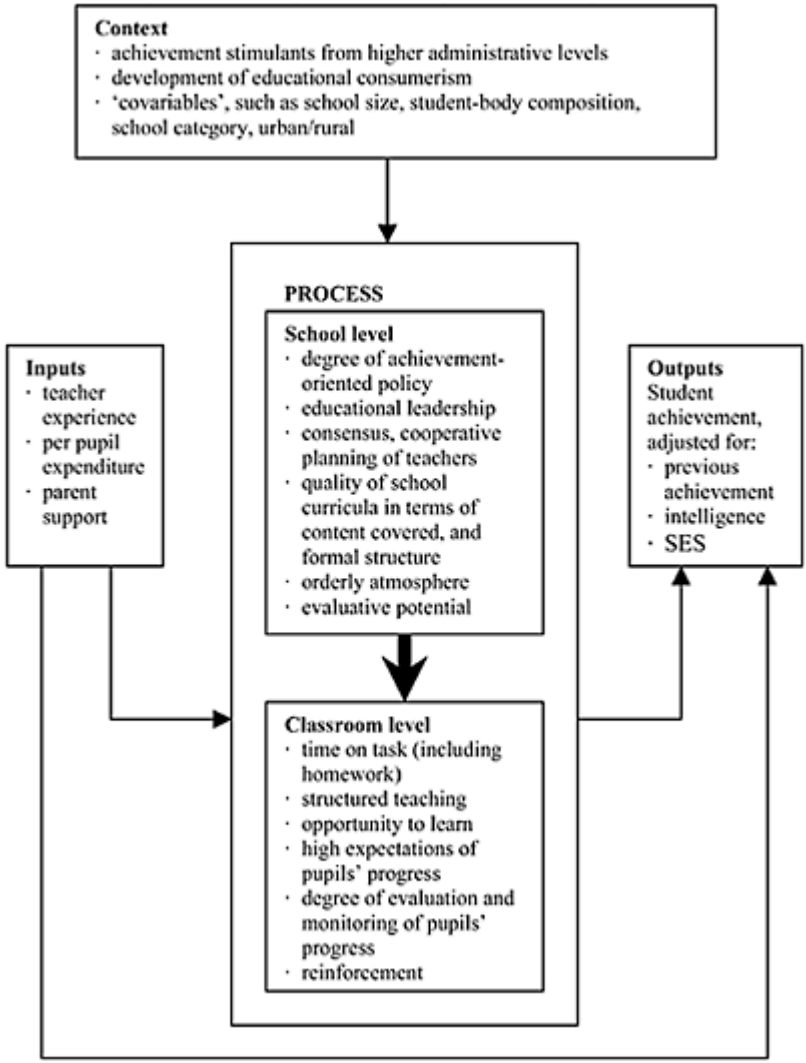


Figure 11.2 An integrated model of school effectiveness (from Scheerens, 1990).

In recent school effectiveness studies these various approaches to educational effectiveness have become integrated. Integration was manifested in the conceptual modeling and the choice of variables. At the technical level multi-level analysis has contributed significantly to this development. In contributions to the conceptual modeling of school effectiveness, schools became depicted as a set of "nested layers" (Purkey & Smith, 1983), where the central assumption was that higher organizational levels

facilitated effectiveness enhancing conditions at lower levels (Scheerens & Creemers, 1989). In this way a synthesis between production functions, instructional effectiveness and school effectiveness became possible, by including the key variables from each tradition, each at the appropriate "layer" or level of school functioning [the school environment, the level of school organization and management, the classroom level and the level of the individual student]. Conceptual models that were developed according to this integrative perspective are those by Scheerens (1990), Creemers (1994), and Stringfield and Slavin (1992). Since the Scheerens model was used as the starting point of the meta-analyses described in subsequent sections it is shown in Figure 11.2.

The choice of variables in this model is supported by the "review of reviews" on school effectiveness research that will be presented in the next section.

Exemplary cases of integrative, multi-level school effectiveness studies are those by Mortimore et al. (1988), Brandsma (1993), Hill et al. (1995), Sammons et al. (1995) and Grisay (1996).

11.4 Summary of Meta-Analyses

In Table 11.3 (cited from Scheerens & Bosker, 1997) the results of three metaanalysis and a re-analysis of an international data set have been summarized. The results concerning resource input variables are based on the re-analysis of Hanushek's (1979) summary of results of production function studies that was carried out by Hedges, Laine and Greenwald, 1994. As stated before this re-analysis was criticized, particularly the unexpectedly large effect of per pupil expenditure.

The results on "aspects of structured teaching" are taken form meta-analyses conducted by Fraser, Walberg, Welch and Hattie, 1987. The international analysis was based on the IEA Reading Literacy Study and carried out by R.J.Bosker (Scheerens & Bosker, 1997, ch. 7). The meta-analysis on school organizational factors, as well as the instructional conditions "opportunity to learn", time on task", "homework" and "monitoring at classroom level", were carried out by Witziers and Bosker and published in Scheerens and Bosker, 1997, Ch. 6. The number of studies that were used for these meta-analyses varied per variable, ranging form 14 to 38 studies. The results in columns 2 and 3 are expressed as correlations between the input or process variable in question and student achievement in mathematics or language. Normally a correlation of .10 is interpreted as "small"; .30 is "medium" and .50 or more is large (Cohen, 1969). The "plusses" in the first column indicate that research reviews mention these factors as positively associated with achievement.

Table 11.3 Review of the Evidence From Qualitative Reviews, International Studies and Research Syntheses.

	Qualitative reviews	**International analyses**	**Research syntheses**
Resource input variables:			
Pupil-teacher ratio		−0.03	0.02
Teacher training		0.00	−0.03
Teacher experience			0.04
Teachers' salaries			−0.07
Expenditure per pupil			0.20
School organizational factors:			
Productive climate culture	+		
Achievement pressure for basic subjects	+	0.02	0.14
Educational leadership	+	0.04	0.05
Monitoring/evaluation	+	0.00	0.15
Cooperation/consensus	+	−0.02	0.03
Parental involvement	+	0.08	0.13
Staff development	+		
High expectations	+	0.20	
Orderly climate	+	0.04	0.11
Instructional conditions:			
Opportunity to learn	+	0.15	0.09
Time on task/homework	+	0.00/-0.01 (n.s.)	0.19/0.06
Monitoring at classroom level	+	−0.01 (n.s.)	0.11 (n.s.)
Aspects of structured teaching:			
–cooperative learning			0.27
–feedback			0.48
–reinforcement			0.58
Differentiation/adaptive instruction			0.22

The results in this summary of reviews and meta-analyses indicate that resource-input factors on average have a negligible effect, school factors have a small effect, while instructional have an average to large effect. The conclusion concerning resource -input factors should probably be modified and "nuanced" somewhat, given the results of more recent studies referred to in the above, e.g. the results of the STAR-experiment concerning class-size reduction.

There is an interesting difference between the relatively small effect size for the school level variables reported in the meta-analysis and the degree of certainty and consensus on the relevance of these factors in the more qualitative research reviews.

It should be noted that the three blocks of variables depend on types of studies using different research methods. Education production function studies depend on statistics and administrative data from schools or higher administrative units, such as districts or states. School effectiveness studies focussing at school level factors are generally carried out as field studies and surveys, whereas studies on instructional effectiveness are generally used on experimental designs. The negligible to very small effects that were found in the re-analysis of the IEA data-set could be partly attributed tot the somewhat "proxy" and superficial way in which the variables in question were operationalized as questionnaire items. An additional finding from international comparative studies (not shown in the table) is the relative inconsistency of the significance of the school effectiveness correlates across countries, also see Scheerens, Vermeulen and Pelgram, 1989 and Postlethwaite and Ross, 1992.

PART 2:

EVIDENCE FROM DEVELOPING COUNTRIES

In this part of the chapter the evidence about effectiveness enhancing conditions of schooling in developing countries will be reviewed. The review sets out by referring to earlier review articles, particularly those by Hanushek (1995) and by Fuller and Clarke (1994), which in itself incorporates results of reviews by Fuller (1987), Lockheed and Hanushek (1988), and Lockheed and Verspoor (1991). Next a schematic description of 13 studies conducted after 1993 is provided. Conclusions are drawn about the state of the art of educational effectiveness research in developing countries, in terms of predominance of the type of factors that are studied, outcome comparison with results from industrialized countries, relevant research innovations and implications for policy and practice applications.

11.5 Production Function Studies in Developing Countries

Hanushek (1995) summarized the effects of resources in 69 studies in developing countries.

Table 11.4 Percentages of Studies With Positive Significant Associations of Resource Input Variables and Achievement for Industrialized as Compared to Developing Countries (Sources: Hanushek, 1995, 1997).

Input	**Industrialized countries**	**Developing countries**
	% sign. positive associations	% sign. Positive associations
Teacher/pupil ratio	15%	27%
Teacher's education	9%	55%
Teacher's experience	29%	35%
Teacher's salary	20%	30%
Per pupil expenditure	27%	50%

When the number of positive associations of resource factors with achievement found in this study in developing countries are compared with the percentages cited in section 11.2, (re c) for industrialized countries the comparison shown in Table 11.4 results.

The relevance of facilities in education in developing countries, not shown in the comparison, amounts to no less than 70 when expressed as the percentage of significant positive studies.

The larger impact of these resource input factors in developing countries can be attributed to larger variance in the independent as in the dependent variables. Both human and material resources in education in industrialized countries are distributed in a relatively homogeneous way among schools, in other words: schools do not differ that much on these variables. Regarding the outcome variables (e.g. educational achievement) Riddell (1997) has shown that schools in developing countries vary on average 40% (raw scores) and 30% (scores adjusted for intake variables). This is a considerably larger variation than is usually found in industrialized countries; where values of 10% to 15% between school variance on adjusted outcomes are more common (cf. Bosker & Scheerens, 1999).

The positive outcomes of production function studies in developing countries make intuitive sense (if basic resources and facilities are not present this will obviously be detrimental to the educational endeavor as a whole). At the same time the outcomes give rise to interesting interpretations when they are brought to bear on the principles of micro-economic theory. Jimenez and Paquea (1996), for example, present findings that support the thesis that local involvement in school finance stimulate both achievement orientation as economy in spending. Their study on public primary schools in the Phillipines provided evidence that efficiency gains (less costs, while maintaining quality

standards) were obtained in settings where the community provided extra funding and schools were held accountable for this. Pritchett and Filmer (1997) point at the political advantages of spending on human resources (diminishing class size in particular) as compared to spending on instructional materials, despite the much larger efficiency of the latter approach, while Picciotto (1996) criticizes the narrow set of educational performance criteria that is used in most education production function research and states that "program design must be informed by assessments of overall educational performance against societal objectives; by evaluations of the relevance of the objectives themselves and by judicious design of institutions to deliver the needed services" (ibid, 5). Microeconomic theory has interesting conjectures with respect to control mechanisms in education as well; where the argument is that bureaucratic control measures are expensive and faulty and community involvement and "direct democracy" would present a better alternative. Currently these conjectures should be appreciated for their heuristic function in stimulating further research. The evidence is not sufficiently inclusive, however, to allow for an overall assessment of consumerbased versus bureaucratic control. Moreover, outcomes are more likely to be contingent on other situational factors, like the traditional structure of the educational systems and cultural aspects.

When studies are becoming more theory-driven and cost-benefit analyses are more frequently included, production function research is to be considered as a viable approach to school effectiveness studies in both developed and developing countries. Particularly so in developing countries because of generally lower levels and greater variability of school inputs.

11.6 Reviews of School Effectiveness Research in Developing Countries

Fuller and Clarke (1994) carried out a major review of school effectiveness studies in developing countries. The review considered about 100 studies and drew upon earlier reviews by Fuller (1987), Lockheed & Hanushek, 1988, Lockheed & Verspoor and an analysis of 43 studies in the period 1988–1992 conducted by the authors themselves.

Only studies that controlled achievement for students' family background were included; and only significant associations at the 5% level were reported. They found that there were about three times as many studies carried out in primary schools as compared to secondary schools. Also, financial, material and human resource input variables were investigated more frequently than school and classroom process variables, with the exception of instructional time. This predominance of relatively easily assessable input characteristics is also evident from the fact that variables like class size and teacher training were studied about four times more frequently than school organizational characteristics and about twice as frequent as instructional characteristics like "instructional time" and "specific pedagogy" (Fuller & Clarke, 1994).

On the basis of their review of significant positive effects Fuller and Clarke (ibid) conclude that rather consistent school effects can be found in three major areas: *availability of textbooks and supplementary reading material, teacher qualities* (e.g. teachers' own knowledge of the subject and their verbal proficiencies) and *instructional time and work demands placed on students*.

Policy relevant factors that showed inconsistent or lack of effects appeared to be class size and teacher salaries.

Fuller and Clarke's review once more underline the predominance of production function type of effectiveness studies in developing countries. Riddell (1997), in a more methodologically oriented review, observes that a "third wave" of school effectiveness research in developing countries is "in danger of being lost without ever having been explored". By this third wave she refers to, what I have described as "integrated school effectiveness studies", comprising resource inputs, organizational factors and instructional characteristics, in which multi-level modeling is a vital methodological requirement.

An interesting set of suggestions that Fuller and Clarke develop in their interpretation of the research evidence, is to pay more attention to cultural contingencies when studying school effectiveness in developing countries. Such contingencies might help in explaining why school and classroom level variables "work" in one country but not in the next. They distinguish four broad categories of cultural conditions:

a. the local level of family demand for schooling;
b. the school organization's capacity to respond to family demand "while offering forms of knowledge that are foreign to the community's indigenous knowledge" (Fuller & Clarke, 1994, p. 136);
c. the teacher's capacity and preference for mobilizing instructional tools;
d. the degree of consonance between the teacher's pedagogical behavior and local norms regarding adult authority, didactic instruction and social participation within the school (ibid, p. 136).

These ideas, as well as the appeal to overcome other weaknesses of school effectiveness studies (lack of cost benefit analyses, shortage of longitudinally designed studies) have demanding implications for the design of studies. According to Riddell (1997) Fuller and Clarke fail to present clear research alternatives.

From a review of 12 more recent effectiveness studies carried out in developing countries (Scheerens, 1999) reconfirmed the predominance of the production function approach with a restatement of the importance of equipment, particularly textbooks and the human resource factor (teacher training). According to the author instructional and pedagogical theory appeared to be practically missing as a source of inspiration for educational effectiveness studies in developing countries. In the four studies that did look into some school organizational and instructional variables, the impact of these variables was relatively low. This (limited) review of 12 studies confirms the results of an earlier review by Anderson, Ryan and Shapiro (1989) who stated that "variations in teaching practice in developing countries are only rarely found to be associated with variations in students learning". Cultural contingencies, as referred to by Fuller and Clarke, or lack of variation in teaching practices in some developing countries could be offered as hypothetical explanations for these outcomes.

11.7 Scope and Limitations of the School Effectiveness Model for Educational Planners

Although the integrated model of school effectiveness is comprehensive in that it encompasses input, process, output and context conditions and recognizes the multilevel structure of educational systems it has a number of limitations.

1. The model has the level of the individual school as its focus, and leaves important issues of a proper functioning of national education systems untreated; I shall refer to this as the *aggregation limitation*. When *subsidiarity**) is applied and schools are autonomous this limitation is counterbalanced to a degree, since, by definition, the school would have more formal responsibilities.
2. The model has a strongly instrumental focus, treating educational goals and objectives as largely "given". Extending the model according to the larger perspective of organizational effectiveness, as briefly referred to in part I, can partly compensate for this limitation, by taking into account the responsiveness of the school vis-à-vis changing environmental constraints. It is again dependent

*) See discussion and explanation of this concept further on.

on the pattern of functional decentralization in an educational system, to what extent adoption mechanisms at school level are important as compared to the provision of such levels at the macro level. We shall refer to this limitation as the *instrumentality limitation.*

3. Although the model is amenable to include questions of equity and efficiency, the actual research practice has not lived up to expectations in this area. Moreover, the way school effectiveness research is dealing with these issues is also determined by the other two limitations concerning level of aggregation and instrumentality. The argument is that, particularly in developing countries, these issues deserve to be dealt with from a broader perspective than the school effectiveness model. This limitation will be referred to as the *relatively narrow quality orientation.*

re 1) aggregation limitations

As indicated in Figure 11.2, where an "integrated" model was shown, school effectiveness is seen as including malleable conditions at various levels of education systems. The bulk of these malleable conditions is situated at the school level. This results points at the focus, perhaps also to be seen as a limitation of empirical school effectiveness research. The component which includes contextual conditions is less well developed. The model concentrates on contextual conditions that can be linked to stimulation of achievement orientation at school level. Examples are the setting of achievement standards and the stimulation of educational consumerism. The practice of reporting school performance through public media links both. So "standard setting" and stimulating accountability, by introducing evaluation and feedback mechanisms are measures of (national) educational administrators included in the "integrated" school effectiveness model. Clearly this is not all that national education planners can do to stimulate the overall quality of schooling. Other major issues are:

- privatization and decentralization,
- creating vertical coordination between levels of schooling (e.g. in the sense of ISCED-levels),
- setting standards for teacher training and providing teacher training;
- providing sufficient access to schooling (which may involve trade-offs between "quantity and quality" of schooling in developing countries, and providing an equitable distribution of scarce educational resources.

The issue of decentralization deserves some further attention in this context, because it points at contexts where the importance of school level conditions is enhanced, which means that the malleable conditions laid bare by school effectiveness research gain in relevance. First, some clarification will be provided with respect to the concepts of "functional decentralization" and "subsidiarity". These concepts provide a basis to determine the relative importance of the school as a decision-making level in education systems, and moreover differentiate the answer to this question according to particular domains of decision-making. In the history of education in the Netherlands the term subsidiarity was used to refer to a specific way in which denominational pressure groups in education linked to see the relationship between the state and corporations representing interest groups in the educational field. According to the subsidiarity principle the state should not interfere in matters that can be dealt with by organized units of professionals. In the original case these organized units were the denominationally based corporations or pressure groups of representatives in the education field, their umbrella organizations in particular. "Subsidiarity" was the term preferred by the RomanCatholic denomination, while the Protestants spoke of "sovereignty in one's own circle". Leune (1987, 379–380) points at the corporatistic nature of this kind of concepts. According to the subsidiarity principle the state only acts subsidiary, that is, it only interferes as a replacement, when needed. A simple example of subsidiarity is a driving-instructor, who takes over the steering of a vehicle when the trainee makes a mistake, but in all other cases quietly watches without interference. Within the context of the European Commission the term subsidiarity is used to express the principle that what can be accomplished by the member states should not be done by the central organs of the Union.

Of course it is debatable to what extent subsidiarity should be applied to schooling, in other words which functions the schools could accomplish without interference from higher administrative levels. The concept of functional decentralization, already introduced in Chapter 4, helps in nuancing this discussion by taking into account that a system can decentralize in some domains, but not in others.

Although various classifications are available in the literature (cf. Van Amelsvoort & Scheerens, 1997) the most commonly recognized educational domains are:

- the curriculum (including goals and standards)
- finance
- the conditions of labor and personnel policy
- school management
- teaching methods
- quality control

A well-known pattern of functional decentralization is a liberalization of finance (e.g. block grants), management (cf. "school-based management"), and teaching methods,

accompanied by a centralized core curriculum. In actual practice it appears hard to relax central regulations concerning the conditions of labor of educational personnel, under conditions of collective bargaining by trade unions.

A further qualification with respect to the degree of decentralization is possible by recognizing that sometimes government units are merely dispersed ("deconcentration"), that decision-making authority is sometimes only partly shed ("delegation") and in other cases is completely given to local bodies ("devolution") (cf. Bray, 1994).

Although the empirical evidence is scarce, there appears to be some support for the hypothesis that functional centralization on curriculum standards and assessment enhances educational performance (e.g. Conley, 1997). Setting achievement standards and assessing student achievement relate favorably to effectiveness enhancing conditions at the school level. Having clear, accessible objectives can add to the overall purposefulness and achievement orientation of the school. It can, likewise, be seen as a supportive condition for "instructional leadership", and, if information is properly fed back to stakeholders, as a basis for organizational learning, accountability and improved "consumerism".

A further hypothesis, regarding developing countries is that the lower the level of schooling of parents and the poorer the catchment area of the school the more effective these measures of functional *centralization* are likely to be.

In summary, this section has underlined that there are important categories of measures of system level educational policy that are *not* covered by the school effectiveness model. So the school effectiveness approach should definitely not be seen as a panacea for all educational problems, particularly as far as developing countries are concerned.

To the extent that systems become functionally decentralized, particularly in the pedagogic and school management domain the malleable conditions of schooling, which research has identified as stimulating effectiveness, gain importance.

re 2) instrumentality limitation

Another aspect of the school effectiveness model is the "goal immanent" orientation. A function of "goal detection" or adaptation of goals according to changing societal and contextual conditions is missing. When the school effectiveness model is broadened in scope, by taking into account additional criteria such as responsiveness, participant satisfaction and formal structure (cf. Faerman & Quinn, 1985) this situation is improved. In developing countries material support from the local community appears to be particularly important, and part of the schools' effort would be needed to acquire this support.

Given its technical and instrumental orientation the school effectiveness model is not strongly oriented towards incentives, and trade-offs between task-related and person-related interest. This is one of the reasons to attempt to connect microeconomic theory and school effectiveness modeling (cf. Scheerens & Van Praag, 1998).

Again, in developing countries "adaptability" and provisions of conditions that create incentives for good performance also deserve to be dealt with at macro level.

re 3) relatively narrow quality orientation

The school effectiveness models is, at its core, an instrumental model of direct school outputs (as compared to more long term, societal outcomes of schooling), in other words quality is addressed as technical effectiveness. The origin of school effectiveness research lies in improving education in poorer "inner city" districts in US cities, and, among studies, there is definitively a bias towards less "privileged" educational contexts, and therefore the research findings have a certain relevance to creating more equal educational provisions. Equity is more directly addressed in studies on so-called "differential effectiveness", where the effectiveness of a school is differentiated according to sub-groups; i.e. boys/girls and children with high and low SES backgrounds. These studies are scarce, and the results inconclusive, however. The same applies to studies that have addressed cost-effectiveness. This state of affairs underlines a previous conclusion that the school effectiveness model inadequately addresses equity and efficiency of educational provisions at large and that, particularly in developing countries, these issues should be addressed primarily at the level of macro level educational policies.

11.8 Summary and Conclusions

In this chapters five strands of educational effectiveness research were discussed. The general conclusion, when reviewing the bulk of the research, was that in developed countries the impact of resource-input factors is fairly small. This outcome was interpreted against the background of relatively small variation in these variables in developed countries. On the basis of recent studies, human resource inputs, particularly teacher qualifications, deserve reconsideration, however. In developing countries the significance of the impact of resource input factors was established in a larger proportion of studies. Several reviewers have pointed at the larger between school differences in developing countries (Bosker & Witziers, 1996, Riddell, 1997), which could explain the differences between developed and developing countries in these research outcomes.

Compensatory programs, school improvement projects and studies of unusually effective schools in developed countries have concentrated on a similar set of relevant school-organizational variables. Reviewers agree on the relevance of factors like: achievement oriented school policy, educational leadership, consensus and cooperation among staff, opportunities for professional development of staff and parental involvement. When subjected to statistical meta-analysis, the impact of these school-organizational factors is relatively small to "medium". In developing countries these factors have been studied infrequently; what results are available show insubstantial impact.

At classroom level instructional and teacher effectiveness studies have indicated medium to large effects of variables like: time on task, content covered or "opportunity to learn", and aspects of structured teaching like; frequent monitoring of students' progress, feedback, reinforcement and cooperative learning. A limitation of these research outcomes is that they have not addressed other than subject-matter based learning objectives in traditional school subjects. On the other hand such learning objectives are likely to remain relevant and these outcomes, which support a behavioristic

interpretation, are sufficiently robust to be considered vis-à-vis constructivist perspectives on learning and instruction. Again, results depend mostly on studies in developing countries. From the limited number of studies in developing countries that was considered no substantive impact of instructional factors was apparent. More detailed and in depth studies of instructional variables in the context of developing countries, also in relationship to cultural background factors, as suggested by Fuller and Clarke, 1994, are considered as quite relevant for future research.

In the course of this chapter quite a few limitations of the research findings have been pointed out, also with respect to the interpretation and use of these findings in developing countries. The question of the robustness of the knowledge base on school effectiveness should, once again, be considered.

What is to be noted, first of all, is that in developed countries the margins to which schools can make a difference appear to be relatively small when expressed in the usual social scientific criteria for effect-sizes. Between school variances in developing countries are generally larger.

When interpreted in a more "practical" way, for example by comparing the 10% most effective schools to the 10% least effective schools, for a country like the Netherlands, would make for a difference of one or two levels of the hierarchically categorized secondary school-system. This means, for example, that pupil A in effective school X with the same ability level as pupil B in ineffective school Y would get the advice to go to a secondary school of the lowest level, while B would get the advice to go visit a secondary school at level 3 (the Dutch system has currently 4 difficulty levels of secondary schools). Other authors have expressed this difference in terms of one grade-level (Purkey & Smith, 1983). It should also be noted that this societal effect would be there for all the pupils in these 10% higher or lower scoring schools.

The next question is the degree to which the net between school variance in pupils' achievement is attributable to the malleable conditions of schooling that are considered as the "independent" variables. In a typical "integrated" school effectiveness study, which contains school level and classroom level variables, as the study by Brandsma, 1993, the relevant proportion was about 60%. This means that a relatively large proportion of the between school variance (say the variation between school average scores on a particular achievement test) is explained by the variables that have been selected on the basis of school effectiveness models. As stated in the above, however, this between school variance is usually only a relatively small proportion of the total variance in pupil achievement (on average about 10% in industrialized countries and much larger (up to 30–40%) in developing countries. An important alternative source of variance is the "contextual" effect of e.g. the average initial aptitude of the students. Within the small margins of the variance that lies between schools in developed countries, this appears to be a fair support for the variables that have been proposed as hypothetical effectiveness enhancing conditions.

In developing countries research appears to support the common sense notion that provision of basic resources, particularly among the most deprived schools, makes most of the difference. In this context the challenge for the future lies in more frequent and in-depth study of instructional conditions.

A final observation regards the larger impact of factors closer to the actual teaching and learning process as compared to more 'distal' factors like school organizational and

school environmental conditions. From the perspective of national policy-making and planning these results should be weighted against the efficiency of bringing about changes at a higher level in the system (which contains fewer units). If there is evidence for a positive, although small, significant impact of a particular style of school leadership, “instructional” or “educational” leadership as this research literature shows, a training course for head teachers could be more cost-effective than training all the teachers in the country.

Interpreting the factors considered in various strands of educational effectiveness research as “levers” for change and improvement requires an exploration of relevant theory, which will be the subject of the next chapter.

12
The Meaning of the Factors That are Considered to Work in Education[1]

12.1 Introduction

The core of the empirically supported knowledge base on educational effectiveness is a set of factors that have been shown to be positively associated with pupils' achievement in basic school subjects. Before addressing the question about the firmness of the empirical support for these factors and the strength and direction of their association with achievement, a closer look will be taken at the conceptual meaning of the most commonly mentioned factors. Referring to the conceptual map of school effectiveness developed in Chapter 10, it should be noted that the total set of factors comprises both conditions at school and conditions at classroom level and that some factors have a structural, whereas others have a more cultural nature. The central concepts in educational effectiveness are therefore only partially objective and descriptive (those related to structure). An important part has to do with attitudes, perceptions and normative positions (those related to culture).

In this chapter an attempt will be made to capture the operational core of the factors that are usually mentioned in the reviews on school effectiveness research. This will be done by taking a close look at the contents of the actual instruments that have been

developed within the context of empirical school effectiveness studies and as part of instruments for school self-evaluation.

The following school effectiveness studies were used as the basis for this inventory: the Junior School Project (Mortimore et al., 1988), the Differential School Effectiveness Project (Sammons et al., 1995), the OECD-INES International Survey of Schools (Scheerens & Ten Brummelhuis, 1996), the School Improvement and Information Service (Hill et al., 1995), the Third International Mathematics and Science Study (TIMSS) (Knuver & Doolaard, 1996; Universiteit Twente, 1995a, 1995b), the Stability of School Effects Study (Doolaard, 1996), Study into school and classroom characteristics secondary education (Van der Werf & Driessen, 1993). Apart from these school effectiveness studies five Dutch school self-evaluation instruments were analyzed (Hendriks & Scheerens, 1996) as well as the Case/IMS self-evaluation system (Keefe, 1994).

[1] Reprinted from The Foundations of Educational Effectiveness, J.Scheerens & R.J.Bosker, pp. 99–138, 1997, with permission from Elsevier.

The general factors summarized in Table 12.1 were analyzed:

Table 12.1 General Effectiveness Enhancing Factors.

1.	achievement orientation/high expectations/teacher expectations
2.	educational leadership
3.	consensus and cohesion among staff
4.	curriculum quality/opportunity to learn
5.	school climate
6.	evaluative potential
7.	parental involvement
8.	classroom climate
9.	effective learning time (classroom management)
10.	structured instruction
11.	independent learning
12.	differentiation, adaptive instruction
13.	keeping records on pupils' progress
14.	feedback and reinforcement

The elements found in the operational definitions and instruments concerning these factors will be summarized for each factor. In addition, an impressionistic view on the 'core' of each factor will be given.

It should be noted that the selection of effectiveness enhancing factors closely resembles the factors included in Scheerens' (1992) model, presented in Chapter 10: factors that were represented in the set of instruments that was analyzed and which are not included in Scheerens' model are: adaptive instruction, classroom climate and independent learning. When comparing the factors in Table 12.1 to the modes of schooling in Table 1.5, it is clear that these factors are only a subset of the modes, the major distinction bring the missing out of environmental conditions in Table 12.1.

12.2 Achievement Orientation/High Expectations

Within the set of operational definitions that was considered the following main components could be distinguished:

Table 12.2 Components and elements of achievement orientation/high expectations.

Achievement-oriented school policy/high expectations
A clear focus on the mastery of basic subjects
• a relatively high curricular emphasis on basic subjects as compared to other subjects
• a relatively high curricular emphasis on basic subjects as compared to general pedagogical aims like personal, cultural, and social development
• high emphasis on basic subjects now as compared to five years earlier
• emphasis on Value added' or progress
• in which areas has progress been made during the last 5 years?
• knowledge transfer and academic development have precedence over general development
• explicit statement of minimum competency levels in basic subjects
• explicit measures to improve quality of education in basic subjects
High expectations (school level)
• school policy is aimed at reaching minimum competency objectives for all pupils
• all teachers stimulate pupils to reach a highest possible score on an assessment test in the highest grade
• to-day pupils do as well as formerly
• stating relatively ambitious achievement levels motivates teachers and pupils
• explicit statement of high expectations on pupils' achievement in policy plans, in communications between head teachers and teachers and by means of rewarding pupils for outstanding performance, or good progress at each level of achievement
• becoming an effective school is the central mission of the school
High expectations (teacher level)
• teachers believe that high expectations on pupils' achievement stimulate school effectiveness
• the degree to which teachers strive for high pupils' achievement
• the degree to which teachers believe that his/her own perceptions influence achievement
• teachers' attitude towards the degree to which pupils' performances can be improved
• the degree to which teachers strive for minimum competency levels
• the degree to which teachers require high achievement of each pupil
• the degree to which teachers believe that objectives and standards can be reached
• teachers emphasize that performance can always be improved
• teachers stimulate pupils to work harder
• teachers pay attention to good performance and reward good achievement

• the degree to which pupils experience that teachers have high expectations on their performance
'Keeping and using records on pupils' achievement
• the school keeps achievement records on all pupils
• the school uses achievement records to compare itself with other schools and with earlier performance

• a clear focus on the mastery of basic subjects;
• fostering high expectations on pupils' achievement, at school and teacher level;
• the use of records on pupils' progress.

Table 12.2 contains an overview of elements that were distinguished as a further specification of these three major components.

Elements of achievement orientation, or pressure for achievement that are not contained in this overview, but have been mentioned in the literature are:

• "placing 'attainment' on the agenda of staff meetings and in talks between the school head and individual staff';
• "employing achievement pressure as a criterion when recruiting new teaching staff';
• "implementing resources, including testing systems, that make it easier to introduce an achievement-oriented policy" (Scheerens, 1992, p. 87).

It is clear that the general concept of achievement orientation and fostering high expectations comprises overt policy choices, attitudes, behaviors and structural facilities. The core idea is the determination to get from pupils what they are worth, in term of aptitudes and home environment. Standard setting in a way that pupils are challenged, but not demotivated because the standards are either too high or too low, appears to be the main structural measure in a 'balanced' interpretation of achievement orientation. 'Balanced' in the sense that no mono-maniacal preoccupation with achievement, regardless of ability levels, is implied, but care is taken of individual differences between pupils.

12.3 Educational Leadership

In the operational definitions and instruments that were analyzed a first general division in conceptions of educational leadership can be made between:

a. general leadership skills applied to educational organizations:

• articulated leadership
• information provision
• orchestration of participative decision making
• coordination

b. instructional/educational leadership in a narrower sense, i.e. leadership directed at the school's primary process and its immediate facilitative conditions:

• time devoted to educational versus administrative tasks
• the head teacher as a meta-controller of classroom processes

• the head teacher as a quality controller of classroom teachers
• the head teacher as a facilitator of work-oriented teams
• the head teacher as an initiator and facilitator of staff professionalization

Table 12.3 contains an overview of elements belonging to these nine sub-categories of educational leadership.

Table 12.3 Components and Elements of Educational Leadership.

a) general leadership skills
Articulated leadership
• the school leader has a clear and explicit view on how the school has to be managed
• the school leader provides clear and unambiguous leadership
• the degree to which head teachers take the lead
• the school leader has considerable discretion
• the school leader plays a major role in hiring new teachers, initiating new policy, initiating new curricular options and teaching methods
The school leader as an information provider
• degree, timeliness and quality of information provision
• adequate dissemination of information
• the head teacher informs parents, parents' association and board regularly
• the head teacher channels information so that it reaches the relevant people involved
• the head teacher sees to it that there is sufficient information on the work of colleagues in order to reach sufficient coordination of tasks
• the school leader informs the teaching staff about the board's decisions
The school leader as an orchestrator of participative decision making
• the school leader uses a clear decision-making procedure
• decisions are taken on the basis of sound and well-grounded information
• decisions are supported by a sufficient number of staff
• the time needed to take decisions is fair
• it is clear in our school who decides on what subject
• decisions are taken by the whole team
• head teachers feel they can control matters at school
• the school leader engages teachers in the choice of new subject matter and teaching methods
• the classroom teacher has a say in decisions about his/her classroom
• the school leader engages personnel in the school's policy making

• the school leader engages parents in decision making

• the school leader sees to it that decisions taken are carried through

• innovation is not hindered by decision making

• the head teacher sees to it that clear decisions are made in staff meetings

• the school leader is firm in adhering to rules and agreements

• the school leader feels that engaging teachers in decision making stimulates school effectiveness

• the school leader engages the staff in drawing up the guideline for running the school

• the school leader engages department heads in matching teachers and classes, staff appraisal, and policy decisions

• the school leader engages teachers in decisions on matching teachers and classes, provision of teaching aids and materials, the development of school guidelines, the recruitment of new personnel

• there are forums in the school to express views and opinions

• procedures for teacher appraisal are developed in conjunction with the staff

• ease of communication with the school leader as seen from the perspective of the staff

The school leader as a coordinator

• the school leader as an initiator of staff meetings

b) Instructional leadership

Time devoted to educational versus administrative tasks

• the number of hours a head teacher teaches

• total number of hours for managerial, non-teaching activities

• division of school leader activities over administrative/organizational, instructional leadership, contacts with parents, own professional development

• the number of times per year/month a head teacher attends lessons, discusses pupils' functioning with teachers

• teachers are content with the relative emphasis the head teacher spends on instructional versus other leadership tasks

• the degree to which teachers are satisfied with stimulating effectiveness enhancing leadership

The school leader as a meta-controller of classroom processes

• the school leader is aware of pupils' progress

• the school leader initiates consultations about the progress of individual pupils

• the school leader uses records on pupils' progress as a basis to set teaching priorities, modification of curricula and methods, adaptation of teaching

• methods and placing pupils in ability groups

• the school leader stimulates the systematic counselling of pupils with learning and behavioural

problems throughout the school

• the degree to which the school leader takes corrective action on the basis of test results

• the degree to which the school leader emphasizes specific attention to be given to weak pupils

• the school leader requires that teachers keep records on pupils' progress

The head teacher as a counsellor and quality controller of classroom teachers

• teachers are happy with their relationship with the school leader

• teachers experience support, appreciation, counselling and feedback from the school leader

• the school leader knows about educational practice in each classroom

• the school leader regularly asks teachers about their work

• the school leader attends lessons and talks about them with teachers

• the school leader appraises teachers

• the school leader shows his/her appreciation if teachers do a particularly good job

• the school leader encourages teachers to exploit their talents

• the school leader supports teachers who need help in carrying out improvement measures

• the school leader guides and counsels teachers during staff meetings by inquiring about how things go in classrooms in a detailed way, by discussing strong and weak points with teachers, by advising them on how to optimalize instruction, by setting successful teachers as examples, and by stimulating the further development of teachers

• the school leader stimulates teachers to improve their professional craftsmanship

• the school leader may try to modify teaching strategies

• the degree to which the school leader encourages teachers and gives them feedback and recognition

• the number of times the head teacher informally communicates with one or more staff members

• frequency of counselling contacts with beginning teachers

• the school leader uses records on pupils' achievement in appraisal interviews with teachers

• frequency of the school leader attending lessons

• any type of information gathering with respect to the quality of teachers

The school leader as a facilitator of work-oriented teams

• the school leader encourages the staff to work as a team

• the school leader encourages a clearly established division of tasks among staff

• special skills of teachers are taken into account when tasks are divided among staff

• the school leader monitors the general orientation of the various subject matter areas

• the school leader sees to it that different learning routes are aligned

• the school leader monitors the attainment of educational objectives

• the school leader has an open mind with respect to initiatives to improve the quality of education
• the school leader takes appropriate action when desired educational and organizational aspects are not fulfilled
• the school leader and team talk about desired changes at school
• the school team is invited to put forward improvement proposals
• a supportive attitude of the head teacher with respect to the implementation of new methods of work
The school leader as an initiator and facilitator of staff professionalization
• the school leader emphasizes the importance of team development and further education
• the school leader tries to further educate him/herself by means of courses and study of literature
• the head teacher encourages further education of teachers in a selective, targeted way
• there is an explicit policy for furthering training of teachers
• who decides about further training of teachers?
• percentage of staff that has followed courses for further training as a teacher
• percentage of staff that has followed courses during out of school hours/during school hours
• has the school leader taken part in courses aimed at his/her own professionalization?

Of the two dimensions that were distinguished as part of the general concept of educational leadership, the second, namely leadership focused on the school's primary process, should be considered as central. The other dimension addresses the specific demands required for leading and controlling organizations in which professionals at the operating core need to have a considerable degree of autonomy.

As a whole educational leadership can be seen as a phenomenon that needs to strike a balance between several extremes: direction versus giving leeway to autonomous professionals, monitoring versus counselling and using structures and procedures versus creating a shared (achievement-oriented) culture. Sammons, Hillman and Mortimore (1995) in this context refer to the leading professional.

The system-theoretical concept of meta-control is perhaps the most suitable to express the indirect control and influence an educationally or instructionally oriented school leader exercises on the school's primary process. Of course this does not imply that the head teacher is looking over the teachers' shoulder all the time, but he or she is 'involved' in important decisions on objectives and methods, and visibly cares about overall achievement levels and individual pupils' progress. From the set of components that were listed in Table 12.3 it is evident that the meta-control of the school leader is exercised in a non-authoritarian way, expressing concern about pupils, individual staff members, and team work.

Some authors who define educational leadership, say more about structural conditions surrounding the instructional process, whereas others are more focused on cultural aspects. Irwin (1986, p. 126) belongs to the former category in mentioning the following aspects of educational leadership: the school leader:

- functions as an initiator and coordinator of the improvement of the instructional programme;
- states a clear mission of the school;
- has a task-oriented attitude;
- establishes clear objectives;
- supports innovation strategies;
- stimulates effective instruction;
- is quite visible in the organization;
- sees to it that pupils' progress is monitored regularly;
- delegates routine tasks to others;
- regularly observes both the work of teachers and pupils.

Leithwood and Montgomery (1982, p. 334) mention the following more cultural aspects of educational leadership:

- stimulation of an achievement-oriented school policy;
- commitment to all types of educational decisions in the school;
- stimulating cooperative relationship between teachers, in order to realize a joint commitment to the achievement-oriented school mission;
- advertising the central mission of the school and obtaining of support of external stakeholders.

In more recent views on educational leadership, inspired by the concept of the learning organization, motivating staff by providing incentives and creating consensus on goals are emphasized. Mitchell and Tucker's concepts of transactional leadership and transformational leadership (Mitchell & Tucker, 1992) form a case in point. Staff development and the 'human resource' factor are further underlined in these approaches. These newer perspectives do not create a sharp break with the longer existing conceptualizations of educational leadership, but emphasize the cultural and the staffing mode of schooling.

Scheerens (1992, p. 89) draws attention to the point that the rather heavy requirements of an educational leader do not necessarily rest on the shoulders of just one individual:

> *"At first glance the description of 'educational leadership' conjures up an image of a show of management strength: not only the routine work necessary for the smooth running of a school, but also active involvement with what is traditionally regarded as the work sphere of the routine assignments leave sufficient time for the more pedagogic tasks. Nevertheless, this leadership does not always have to come down to the efforts of one main leader. From the school effectiveness research of Mortimore et al (1988) it emerges that deputy heads in particular fulfil educational leadership duties. Delegation can go further than this level: it is desirable that, given the consensus of a basic mission for the school, there is as broad as possible a participation in the decision making. In the end certain effects of pedagogic leadership such as a homogeneous team, will fulfil a self-generating function and act as a substitute for school leadership (according to Kerr's (1977) idea of 'substitutes for leadership'."*

12.4 Consensus and Cohesion Among Staff

Given the traditional autonomy of teachers it is clear that consensus, cohesion and sufficient continuity for pupils when they pass from one teacher to the next, should not be taken for granted in schools. Therefore, in many school effectiveness studies, the degree to which schools succeed in building coherence and consistency is seen as a hypothetical explanation for the fact that some schools do better than others.

In the operational definitions and instruments that were analyzed the following components of consensus and cooperation were distinguished:

- Types and frequency of meetings and consultations.
- The contents of cooperation.
- Satisfaction about cooperation.
- The importance attributed to cooperation.
- Other indicators of successful cooperation.

Table 12.4 contains an overview of the elements that were distinguished as part of each of these five components of consensus and cooperation.

Table 12.4 Components and Elements of Consensus and Cooperation in Schools.

Types and frequency of meetings and consultations
• Number of formal staff meetings with the head teacher
• Frequency of informal meetings among groups of teachers
• Informal contacts between staff
The contents of cooperation
Items considered important in cooperation at school:
• pedagogical mission
• educational concept
• school aims, objectives
• pedagogic actions
• planning and implementation of lessons
• acquiring teaching methods and materials
• discussing pupils' achievement
• establishing entrance behavior at the beginning of the school year
• treatment of pupils with learning difficulties
• educational change and innovation
• subject matter choice, assignments, achievement test, homework, preparation of lessons, observation of lessons

• counselling of beginning teachers
Satisfaction about cooperation
• satisfaction in relation to colleagues with respect to allocation of duties and coordination concerning:
– variety in interests
– professional competence
– supporting school improvement
– involvement in pupils' learning and satisfaction
– the amount of curriculum/'techniques'-discussion in team meetings
– acceptance, support and opportunity to cooperate
– cooperation at school and within the team
The importance attributed to cooperation
• To which degree do head teachers agree on the importance of the following activities, as effectiveness-enhancing conditions:
– the necessity of aligning the curriculum of subsequent grade levels
– similarity in teaching approach among grades and classrooms
– a common policy with respect to pupils with special learning and behavioral problems
– the use of pupil records to be passed from one grade level teacher to the next
– the importance of cooperation within departments
Other indicators of successful cooperation
• explicit policy aimed at furthering cooperation among staff
• encouragement of consultations on lesson goals, teaching strategies, use of equipment
• explicit division of tasks and coordination activities
• an established practice of team teaching
• consensus among staff, within departments
• frequent discussions about curriculum and teaching approach

In the way consensus and cooperation is measured, facts, actual cooperation and frequency of sessions where staff meet and cooperate, as well as perceptions and attitudes on cooperation are both included. With respect to the substance of cooperation both agreement on overall mission and educational philosophy as well as consultation on "technical" aspects of teaching and instruction are measured.

There appears to be no agreement on areas of cooperation that are thought to be particularly relevant. Across studies a broad range of cooperation activities and topics to cooperate on are chosen.

12.5 Curriculum Quality and Opportunity to Learn

The curriculum has been described as the "blue print" for the functioning of the primary process in education. In articulating the curriculum and by indicating clear targets, the curriculum could function as a powerful coordination mechanism (i.e. a form of standardization). On the other hand such standardization is usually balanced by opportunity for teachers to exercise their own professional autonomy.

The degree to which content that is actually taught (sometimes described as the "implemented curriculum") corresponds to the test or examination of items used to assess achievement (the achievement curriculum) is usually taken into account in international comparative studies under the label "opportunity to learn".

Examination of the instruments in this area led to the following categories:

- The way curricular priorities are set.
- Choice of methods and textbooks.
- Application of methods and textbooks.
- Opportunity to learn.
- Satisfaction with the curriculum.

Table 12.5 Components and Elements of Curriculum Quality and Opportunity to Learn.

The way curricular priorities are set
• the extent to which subject matter provision is determined (i.e. guidelines are developed) by the ministry, the school board, the school team
• knowledge about core objectives arithmetic/math and science, the school work plan
• the importance of a good range of extra-curricular activities for the school's effectiveness
• the importance of:
– provision improvement for extending special needs in ordinary schools
– improving preparation for the post-graduate course/profession-oriented education
• attention for:
– acquiring unconventional behavior
– subject integration factual subjects
– realistic math education
– introducing computers
– the attainment targets
• attention for learning study skills
Choice of methods and textbooks
• availability of books for language and math
• well-functioning methods for spelling, decoding, reading comprehension, composition writing

and math, meaning:
– a clear line with regard to subject matter content
– clear directives for instruction and testing
– a step-by-step approach for the low achievers
– a clear distribution of minimum competency goals over school years
• which language methods (in which group)
• which arithmetic-math methods (in which group)
• method for science
Application of methods and textbooks
• knowledge of the manual for arithmetic/math/science methods
• the time the method is being used
• considering transfer to other methods
• which part and which chapter in the beginning of the school year
• which part and which chapter now
• keeping sequence in the method
• % of subject matter dealt with at the end of the school year
• progress in method at the end of the school year
• other material for arithmetic/math/language/science than prescribed in method
• use of a calculator
• % of pupils being in a position to use a calculator
Opportunity to learn
• % of time for arithmetic/math/science spent on method
• division of lessons to subject matter components
• other subject matter areas (within the subject)
• number of lessons per subject matter area
• which test items link up with education taught so far (for arithmetic/math and science
Satisfaction with the curriculum
• education gets shape and content in accordance with the schools' vision and goals
• the extent of satisfaction with the curriculum now and 5 years ago
• satisfaction with the curriculum and the teaching materials
• satisfaction with the choice of subjects offered
• effectiveness of the curriculum's coordination within in the school

• successes with respect to extra-curricular activities and curriculum development over the past 5 years
• the degree to which the work at school is considering interesting
• the extent to which a curriculum is modern
• lessons:
– number of lessons that stir the imagination
– diversity of subjects

When overviewing the elements in the instruments, summarized in Table 12.5, the core elements appear to be:

– a clear focus of the curriculum;
– coordination and alignment of the curriculum (relationship goals and curricular choices, correspondence among grade levels, classes and teachers);
– test curriculum overlap, or "opportunity to learn".

12.6 School Climate

The concept of school climate can be seen as a synonym of school culture. In the history of school effectiveness research two aspects of culture and climate have received emphasis: orderliness and achievement orientation. In the earlier presentation achievement orientation has been treated as a characteristic of explicit, or even official policy. Achievement-oriented climate refers more to internalized norms and views of individual staff members shared with their colleagues, also in less formal relationships. A third aspect of school climate is the experience of the general "goodness" of all kinds of internal relationships and the satisfaction this give to staff and pupils.

Table 12.6 Components and Elements of School Climate.

Aspect a) Orderly atmosphere
The importance given to an orderly climate
• good discipline, pupil behavior and an orderly and safe learning environment are effectiveness enhancing conditions
• inconsistent approach of pupil behavior and discipline and bad pupil behavior impede the school's effectiveness
• the school having a corresponding philosophy with respect to an orderly climate
• the school head finds it important to create a quiet, orderly environment
• the extent to which a school head attaches importance to a task-oriented atmosphere
• the extent to which a teacher pursues an orderly climate

Rules and regulations

• clear rules for pupils, pupils know where they stand

• clear (written) rules for:

- clothing and physical care of pupils

- pupils doing paid jobs

• formally recording and applying rules with respect to a.o. lateness, disturbing the lesson, absenteeism

• the extent to which school rules are recorded per subject

• rules and sanctions with respect to discipline are well-understood by staff and pupils and are not consistently offended

• the extent to which behavioral rules are honest and are being maintained

• proportion of teachers using the following behavioral rules (a.o. looking after pupils leaving the classroom orderly, seeing to it that the classroom is left behind clean)

• the way rules are being applied in case of lateness, disturbing the lesson, cheating and truancy

• improving and maintaining behavioral rules is an important objective for the school

Punishment and rewarding

• % of pupils being disciplinary punished last year

• number of rewards mentioned by the school head

• number of punishments mentioned by the school head

• rewards/punishments ratio

• teacher rewards work more than punishment

• teacher rewards behavior more than punishment

• forms of rewards by school head (a.o. praise)

• forms of punishments by school head (a.o. verbal warnings, confinement)

• a clearly applied system of punishment and rewarding at the school

Absenteeism and drop-out

• registration of pupils' presence/absenteeism

• control of absentee registration by teachers

• the frequency school heads or teams are being confronted with the following behavior (of grade 6)

– being late at school

– being illegal absent

– staying away from a lesson

• measures to avoid structural cancelling of lessons as much as possible

• policy in case a teacher is absent
• measures with respect to truancy
• policy aimed at preventing early school leaving
• measures to prevent early school leaving
• measures taken when a pupil seems to become an early school-leaver
Good conduct and behavior of pupils
• other pupils do not encourage a child teasing another child
• teachers and pupils see to it that teaching-learning processes are undisturbed
• teachers create a learning environment in which pupils can work in a task-oriented way
• see to it that nobody disturbs a teacher during the lesson
• the pupils behave well when the teacher leaves the classroom
• the lessons are not often disturbed by noise down the hall
• level of pupil-sound in the classroom
• level of pupil movement in the classroom
• teachers' audibility in the classroom
• pupils' behavior around the school
• strengthening pupils' behavior
• the level of unaccepted pupils' behavior now and 5 years ago
• important successes and problems with respect to pupils' behavior and discipline now and 5 years ago
• the school's high standards of pupil behavior
• the frequency school heads or team are being confronted with the following behavior (of grade 6)
– vandalism
– theft
Satisfaction with orderly school climate
• a quiet, orderly learning environment at school
• the school yard, the group classrooms and the common apartments form an orderly and attractive play/learning environment for the pupils
• the school supplies a supporting and secure environment
• pupils and teachers feel secure at school
• there is a safe and orderly climate in my group
• satisfaction with respect to safety at school, behavior in the classroom, the school and teachers being attentive

• satisfaction with respect to pupils' behavior
• degree of satisfaction with pupils' behavior now and 5 years ago
• the extent to which teachers set an example in their behavior to pupils
• satisfaction with respect to precautions/the way the school handles vandalism, drugs, alcohol and tobacco
Aspect b) Climate in terms of effectiveness orientation and good internal relationships
Priorities in an effectiveness-enhancing school climate
• effectiveness enhancing conditions for a school
– a caring pastoral environment
– positive inter-personal relationships for staff and students
– the encouragement of a positive attitude to school (pride in school)
– shared goals and values by staff and students
– high level of pupil motivation
– students satisfaction
• effectiveness enhancing conditions for your school
– students feel valued as people
– encouragement of student responsibility
Perceptions on general effectiveness enhancing conditions
• effectiveness enhancing conditions of a school:
– teacher motivation
– teacher commitment/effort
– personal effectiveness of teaching staff
– commitment/enthusiasm of teaching staff
• effectiveness restricting conditions of a school:
– heavy workload
– low staff morale
– lack of commitment and enthusiasm by some staff
– high teaching staff absence rates
Relationships between pupils
• how do you feel about relationships between pupils
• communication between pupils
• pupils want to belong to the school and to each other
Relationships between teachers and pupils

• how do you feel about relationships between pupils and teachers
• contacts with pupils are open and pleasant
• the teacher/pupil social relations are good
• the team tries to understand pupils' needs
• communication with teachers
• teachers like pupils, support them, want them to associate nicely, know what every pupil wants, treat them fair, etc.
• did the school have success with respect to better relationships between teachers and pupils the past 5 years
• team functioning with respect to controlling pupils (firm but friendly relations)
Relationship between head teacher and pupils
• communication between head teacher and pupils
• head teacher listens to ideas/opinions/complaints from pupils about the climate and atmosphere)
Relationships between members of staff
• relationships between teachers
• feeling member of a group
• joint informal meetings occur a few times per year
• colleagues lending a ready ear for personal problems
• feeling a joint/shared responsibility
• break new colleagues in to make them feel at home
• mutual relations aimed at learning from each other
• staff behave according to joint agreed rules
• deviation of habits/rules is allowed
• not blaming colleagues for mistakes
• meetings are characterized by openness and commitment
• the extent of mutual confidence allows to publicly express feelings of (dis)pleasure
• considering people's wishes when taking decisions
• colleagues pressing to put up with a decision in case of a different point of view
• a minority opposing a majority in team meetings
• problems are most of the time dealt with by the team within a reasonable time
• there are conflicts
• team morale (degree of energy, enthusiasm and spirit)
• team functioning now and 5 years ago with respect to:

– solidity of work relationships

– commitment

– shared support

– morale

– working hard

– stress level

Relationships: the role of the school head

• relationships between school head and teachers

• the school head:

– trusts his team members

– can easily be approached

– progresses job satisfaction

– takes suggestions and ideas of teachers with respect to work climate and sphere serious

– pays attention to solving/improving mutual relations in case of conflicts

• the behavior of school head evokes conflict

Engagement of pupils

• pupils have a say in what happens at school

• pupils co-decide about what happens at school

• pupils are proud of the school and show responsibility

• did the school have success with respect to pupils' responsibility the past 5 years

Appraisal of roles and tasks

• teaching/other tasks

• role clarity (clearly described tasks)

• job variety

• degree of job satisfaction

Job appraisal in terms of facilities, conditions of labor, task load and general satisfaction

• sufficient facilities (methods/materials) to efficiently carry out work

• salary and (secondary) conditions of labor

• competent authority passing on to a rewarding system based at personal commitment and motivation of teachers

• importance of part-time appointments

• opportunities for career enhancement

• task load (general anticipatory and perceived psychosocial mental strain):

– in general
– own task load
• satisfaction with respect to working-hours
• teachers believe they are overworked and under pressure
• average absenteeism of team members now and 5 years ago
• quality of working life
• satisfaction with respect to working with pupils
• enthusiasm for the work/the school (now and 5 years ago)
• attention for extra curricular activities
• feeling valued in functioning as a teacher
• opinion with respect to teachers' motivation
• successes/problems with respect to teachers' motivation during the past 5 years
Facilities and building
• classrooms/school/school building/playground clean, neat and well equipped
• sufficient space in/around the school
• sufficiently good facilities in and around the school
• no problems with respect to the school's entrance and with respect to stairs and halls in the school
• service quality in the area of safety, advice, care, health and canteen/stay-over facilities

Indicators on the school climate range from perceptions and normative views to behavioral characteristics and factual circumstances like a set of explicit behavioral rules, absenteeism statistics and characteristics of the school building.

Rules about proper behavior and discipline express the conviction and effort of schools to suppress disruptive and negative, non-task related activities as much as possible. In school effectiveness thinking "good relationships" and satisfaction are considered instrumental to enhanced school effectiveness, and not just as "aims in themselves".

The main sub-categories express the breadth of scope of the school climate concept.

12.7 Evaluative Potential

The concept of "evaluation potential" (Scheerens, 1987) expresses the aspirations and possibilities of schools to use evaluation as a basis for learning and feedback at the various levels within the organization, also taking into account limitations and constraints. Aspects of this concept are:

• priority given to assessment and monitoring;

- evaluation technology (e.g. standardized pupil monitoring systems or computerized "test service systems");
- use of evaluation results and records at the school level.

In Table 12.7 the main components and elements of "evaluative potential" as part of available instruments have been listed.

Table 12.7 Components and Elements of Evaluative Potential.

Evaluation emphasis
• school-wide policy with respect to marketing/assessment and regularly monitoring pupils' progress are effectiveness enhancing conditions
• an inconsistent approach of 'student assessment' restricts effectiveness
• the quality of education is regularly put on the agenda
• the quality of education is a central factor when discussing possible changes
• the majority of the staff is very committed and prepared to deal with quality issues
Monitoring pupils' progress
• a strong emphasis on the evaluation of test results
• agreements and/or rules at school level with respect to testing/registration
• at our school pupils' progress is regularly tested/we handle a good testing system for progress registration to register problems with pupils in time and to take appropriate measures
• the extent to which a department head evaluates the learning progress in the department
• in groups 1 and 2 attention is paid to early signalizing so-called "pupils at risk" with regard to speech-language, social-emotional, auditive, visual-spatial and motor development, concern for more cognitive activities and the task and work attitude
• the extent to which reading and arithmetic are tested
• evaluation of pupils' progress takes place by means of standardized progress tests
• what is pupils' assessment based on (national standards, comparison with other schools, progress of the child itself)
• does the school handle achievement standards for individual pupils/standards at school level
• (written) rules for promotion to the next year/retention yes/no
• decision on promotion/retention based on opinion teacher
• is the school posted on pupils' functioning in further education
The use of pupil monitoring systems
• pupils' progress being administered in a pupil monitoring system at school level
• evaluating pupils' progress in basic skills at least twice a year by means of a pupil monitoring system

• registration of pupils' progress in individual pupil files, in group surveys, in central pupil monitoring system
• which pupil monitoring system is being used and do all teachers use the same pupil monitoring system
School process evaluation
• has the school been assessed during the past 5 years by means of an instrument for school self evaluation
• which aspects are structural tested/evaluated, analyzed and, if necessary, improved:
– pupil satisfaction
– teacher achievement on the basis of pupil data
– teacher satisfaction on the basis of…
– functioning of the school management
– resource expenditure
– courses and teaching
– provision of education
– new teaching methods
– dissemination of innovations
– the process of educational improvement
– implemented changes
– policy formation
• comment upon each other's functioning in a positive way
Use of evaluation results
• the school being aware of possible level of changes in pupil performance during the past 5 years
• the school being aware of it's position with respect to pupil performance with regard to other schools having a comparable pupil population
• for how many subjects is it possible to compare the present average achievement level to 5 years ago
• for how many subjects does the school compare pupil progress with other schools
• discussing pupils' progress and development regularly and systematically
• evaluation of pupil performance:
– leads to adjustment of instruction and learning strategies
– supports assignment to ability groups
– changes in teaching strategies
• comparisons in achievement are being used for educational improvement

• using former pupil data for educational improvement
Keeping records on pupils' performance
• is keeping records on pupils' performance dealt with in the school work plan
• if yes, indications for keeping records on pupils' performance concern the recording of it
• teachers keep records on pupils' development and progress
• does the teacher keep records on language progress
• total number of registrations by teacher
• how often does keeping records on individual pupil's progress in documents open to the school head occur
• method of registration of learning progress:
a. standardized data
b. judgement by individual teacher
c. both a and b
d. there is no registration
• registration school progress:
– not
– in individual pupil file
– in group summary
– in central pupil monitoring system
• are pupils' data kept up with through the entire school career
• if yes, by means of automatized computer system
• frequency in which summaries of registration data are presented:
– per pupil
– per teacher
• group summaries of pupils' achievement are made
• use summaries per pupil/teacher for…
• record results written assignment
• record test results
• execute an error analysis
• process pupils' achievement in pupil monitoring system at school level
• frequency of written reports to parents (per school year/group)
• quality of reporting of pupils' progress (all-embracing, exploratory and valuable information on pupil's progress)

- the school pays a lot of attention to reporting towards pupils and parents
- written pupils' report when pupils pass to next school year

Satisfaction with evaluation activities

- the degree of satisfaction with the student assessment/monitoring system now and 5 years ago
- during the past 5 years, did the school succeed in establishing:

– improved record-keeping/student profiles

– improved monitoring of pupils' progress

- the team's satisfaction with respect to the amount of attention paid to improving education

One of the problems in measuring schools' involvement in evaluation is the diversity in evaluation methods, which range from very informal procedures like the marking of assignments to the regular use of standardized achievement tests. Also, there are several objectives of school-based evaluation:

- monitoring of "normal" progress in pupils' achievement;
- diagnosing learning difficulties;
- assessment of whole school, department or classroom/teacher performance;
- school diagnosis as a basis for prospective innovations and school improvement activities;
- assessment to meet external accountability requirements;
- assessment to be used as a basis for "marketing" the school and informing parents and other stakeholders.

The main aspects of "evaluative potential", distinguished in the introductory section on this factor, orientation, technique and use, are clearly reproduced in the instruments that were analyzed.

12.8 Parental Involvement

Continuity in home and school learning and an active involvement of parents in school matters is considered relevant in various strands of school effectiveness research. Both actual involvement and effort of the school to facilitate involvement are usually included in instruments for measuring this alleged effectiveness-enhancing factor.

Table 12.8 Components and Elements of Parental Involvement.

Emphasis in school policy
• strong parental support as an important condition for school effectiveness
• little parental support impedes effectiveness
• school heads and teachers are open for suggestions from parents
• the school emphasizing the importance of parental involvement with respect to education and pedagogical affairs
• the school being open for parents attending lessons
• the school has a parents' association of which parents can become a member on a voluntary base
• are parents in parents' committees, parents' councils or participation councils reflecting the pupils' population and is this aimed for
• agreements with respect to home visits
• facilities for parents to be present in the school
• parents' complaints are taken seriously
• agreement with the following pronouncements:
– parental involvement is considered positive
– parents are allowed to influence education's organizational structure
– parents are allowed to influence educational contents
– the school's and parents' responsibilities should be clearly defined
– disappointing achievement is often due to parents not supporting the school
• a parent activity program is drafted yearly
• the school stimulates that as many parents as possible attend the individual talks about their child's progress
• the school pays specific attention to parents who are hard to reach
• the school encourages parents to help and support children at home
Contacts with parents
• a good written information exchange between school and parents (school newspaper, monthly bulletin, etc.)
• does the school inform parents about:
– progress yes/no
– educational content aims
– pedagogical/educational starting-points of the school
– important changes with respect to content/structure of education

– subjects dealt with in participation councils meetings
• parental involvement when deciding on:
– policy
– curriculum
– school planning
– finances
– personnel
– school organization
• the school head is available for parents at fixed times
• parents can drop in with the school head any time
• number of parents' evenings for discussing individual pupils' progress
• number of parents' evenings for discussing general subjects
• parents' attendance at parents' evenings about learning progress/general subjects
• all parents are visited at home at least once a year
• the extent to which parents seek for/wish information with respect to their child(ren)'s progress
• teachers give parents concrete instructions with respect to supporting learning and developing skills of the children
• % of parents involved in:
– instructional/learning process
– other school activities (e.g. library/documentation center policy)
– out-of-school activities
– other supporting activities
– homework and homework conditions
Satisfaction with parental involvement
• the school's satisfaction with contacts with parents
• the school's satisfaction with respect to parents' assistance with school activities
• teachers feel they can rely on parents
• the team can well be approached by parents
• parents' satisfaction with respect to the speed they are informed about pupils' progress
• parents' satisfaction with respect to quality of report cards
• improving parental involvement is an important goal of the school

A feature that does not receive much attention in the list of elements cited in Table 12.8 is a particular emphasis on contacts with parents from cultural minority and lower

SES backgrounds (compare e.g. Cotton, 1995). Stringfield and Teddlie (1990) indicate that in neighborhoods where the majority of parents has a generally uninterested or negative attitude to schooling, buffering against parental influence rather than seeking involvement might be a more suitable policy.

12.9 Classroom Climate

Like in school climate orderliness, good relationships and satisfaction are the main components of classroom climate.

Table 12.9 Components and Elements of Classroom Climate.

Relationships within the classroom
• classroom scores on:
− relationships between pupils
− relationships between teacher and pupil
• appreciation for teacher as a companion
• situation with respect to relationships between teacher and pupil now and 5 years ago
• warmth towards pupils (a more rewarding than punishing position)
• attitude teacher towards pupils (treat pupils as responsible, having pupils experience success)
• empathy (the extent to which a teacher comprehends the pupils and take care of them)
Order
• fairness/firmness (control in the classroom)
• classroom scores on:
− order in the classroom
• rules in the group are clear for each pupil
• creation of an orderly, quiet work environment
• situation with respect to control (firm but friendly relations) on pupils now and 5 years ago
Work attitude
• work attitude in the classroom
• in the group there is a (serious) atmosphere, aimed at learning
• see to it that pupils are working task-oriented on their assignment
• teacher energy/enthusiasm (teacher interested and enthusiastic with respect to the curriculum offered
• pleasure in mathematics

• the used of mathematics

• fear and difficulty

Satisfaction

• Classroom fun factor

The fun factor is to give an indication of whether or not it was an enjoyable experience to be a pupil in a particular teacher's class. The 'fun factor' is the sum of all 'yes' responses to the eight items that follow:

− Did the teacher smile often

− Was there positive physical contact with pupils

− Did the teacher show a sympathetic interest in the children other than as learners

− Did the teacher chat to the pupils about non-work matters on any occasion during the day (Whether pupil or teacher initiated)

− Was communication between children generally cheerful

− Was the children's behavior generally relaxed

− Were there any jokes and/or was there any laughter in which the teacher was involved? (this does not include jokes at the expense of other pupils)

− Was there any sign that pupils wanted to be in the classroom outside of class teaching time, either before or after sessions

In comparison to the components that were distinguished in "school climate" the achievement orientation component is missing. This aspect, however, is more or less covered in another factor, namely teacher expectations.

Pedagogical aspects, like the enactment of moral values, in classroom interaction are also underrepresented in the set of instruments that was analyzed.

12.10 Effective Learning Time

Learning time can be interpreted as a measure of the quantity of exposure to "educational treatment" at school. Time can be assessed at school and at classroom level, and a distinction is to be made between "planned time" (e.g. the time per subject matter area in the timetable) and "implemented time" or "time on task".

When summarizing the elements found in the set of instruments that were analyzed the following components were distinguished:

• Importance of effective learning time.
• Monitoring of absenteeism.
• Time at school level.
• Time at classroom level.
• Classroom management (avoiding and minimizing ineffective "time consumers").
• Homework.

The elements are further distinguished in Table 12.10.

Table 12.10 Components and Elements of Effective Learning Time.

Importance of effective learning time
• emphasis on
− developing better policy and better procedures to enlarge instruction time
• impeding/progressing school effectiveness:
− good registration of presence and absenteeism
− good class management
− give high priority to homework
Monitoring of absenteeism
• % of pupils truanting
• the way the school handles absenteeism and lateness
• satisfaction with respect to pupils' presence now and 5 years ago
Time at school level
• number of school days
• number of teaching days/hours
− number of teaching days per school year
− number of full teaching days per school week
− number of semi teaching days per school week
− total number of hours per school week
− length of a school day
• % of canceling of lessons
• number of days with no lessons due to structural causes
• % of total number of hours indicated on the table
• measures to restrict canceling of lessons as much as possible
• policy with respect to unexpected absenteeism of a teacher
• (in school work plan) agreements on substituting teachers
Time at classroom level
• number of lessons on timetable per school year
• a lesson consists of how many minutes
• amount of teaching hours for language/arithmetic

• amount of minutes for arithmetic/physics per week

• duration last arithmetic lesson in minutes

• accuracy with respect to starting and finishing lessons in time now and 5 years ago

• number of lessons that are cancelled

• satisfaction with respect to available amount of time for working in the classroom

Classroom management

• attention for classroom management in the school work plan

− with respect to lesson preparation

− rules and procedures for the lesson's course

• situation with respect to aiming at work in the classroom (now and 5 years ago)

• average % of teachers spending time on:

− organization of the lesson

− conversation (small talk)

− interaction with respect to the work

− supervision (pupil activities/behavior)

− feedback/acknowledgement

• average time during lesson spent on discussing homework, explaining new subject matter, maintaining order

• sources of loss of time during lessons:

− pupils do not know where to find equipment

− disturbances due to bad behavior of pupils

− frequent interruptions

− loss of time due to lengthy transitions from one activity to the next

− unnecessary alterations in seating arrangements

− frequent temporarily absence of pupils during lessons

− waiting time for individual guidance

− many (more than 3) teacher interventions to keep order

− lack of control on pupils' task related work

Homework

• attention for assigning homework at school/agreements in school work plan

• homework after last (arithmetic) lesson: yes/no

• number of homework assignments per week

• type of homework (arithmetic/language) (reading/composition writing)

• amount of homework
• amount of time needed for homework (per day)
• extra homework for low-achieving pupils
• successes and problems now and 5 years ago with respect to:
− prioriting homework
− a consistent homework policy

The attitudinal component “importance of effective learning time” might as well be considered as an aspect of “achievement orientation”, or an “achievement-oriented climate”. The rest of the components and elements appear to be amenable to construction of a one dimensional index in which official time per subject constitutes an upper limit, and actual observed time on task during lesson hours, to which time for homework may be added, the lower limit.

12.11 Structured Instruction

Although, as will be discussed in more detail in subsequent chapters, there are diverging instruction-theoretical and pedagogical perspectives on “good instruction”, in school effectiveness research the view that instruction should be well-structured and closely monitored predominants. In the set of instruments that were analyzed the following components could be distinguished:

- Importance of structured instruction.
- Structure of lessons.
- Preparation of lessons.
- Direct instruction.

Table 12.11 Components and Elements of Structured Instruction.

Importance of structured instruction
• emphasis in school’s policy on
− the quality of teaching
− encouraging pupils to take responsibility for their own learning process (teacher independent learning)
− emphasizing exam preparation
− sufficient ‘challenge’ for both high and low achieving pupils
• to what extent agreed upon:
− whole class instruction gives the best results
− discovery learning mainly needs to happen outside the school

− pupils acquire less knowledge when different pupils do different tasks

− repeating a year often benefits pupils' development

− the high-achieving pupil is especially the victim of individualized education

− individualized education benefits all pupils

− when dividing pupils into groups achievement will do as criterion

Structure of lessons

• direct instruction divided in:

− looking back daily

− presenting subject matter

− guided practice

− giving feedback and correction

− independent practice

− looking back weekly/monthly

• teacher uses a lesson plan

Preparation of lessons

• lesson preparation building upon:

− lessons formerly taught

− written plan

− other teachers/math specialists

− text books

− standardized tests

• most important information source for planning arithmetic/math lessons (lesson content, way of presentation, homework, tests)

− core objectives

− school work plan

− manual

− text book

− other source books

• the subject matter is the central factor when teaching

Direct instruction

• attention for instruction in the school work plan

• indications in school work plan with respect to:

− clear objectives of instruction

− construction of the instruction

− way of presenting subject matter

− the use of instructional materials

• explanation or help to individual/groups of pupils in or outside the lesson

• teachers deal with subject matter that corresponds to the lesson's aim

• teacher explains at the beginning of the lesson to what prior knowledge the subject matter corresponds

• teacher gives pupils the chance to raise questions about the last lesson

• teacher explains beforehand what pupils have to know at the end of the lesson

• teacher knows what to achieve with the lesson

• lesson objectives are clear to pupils

• teacher applies instructional methods to increase pupil's achievement

• teacher deals with only one subject matter component at the time

• explanation in small successive steps

• teacher takes next step when preceding step is understood

• teacher gives concrete examples

• it appears from pupils' reactions that the teacher explains the subject matter clearly

• teacher poses intellectual questions that invite pupils to participate actively

• after posing a question the teacher waits to let the pupils think

• teacher gives many pupils a turn

• a lot of interaction between teacher and pupils

• pupils respond well to questions posed by the teacher

• teacher have pupils practised under guidance

• teacher continues until all pupils have mastered the subject matter

• explanation is clear

• teacher involves pupils in instruction

• teacher takes care that pupils are concentrated during instruction

• during instruction immediate feedback to answers of pupils

• the lesson displays a clear structure

• at the end of instruction summary of subject matter (by teacher/pupils)

• pupils get tasks they can handle

• group work, if appropriate

• teacher's activities (controlling) when pupils work on assignments

- teachers take time to help pupils with tasks
- pupils know which tasks are to be carried out
- teacher sees to it that pupils work in a concentrated way during assignments
- teacher sees to it that pupils work task-oriented during assignments
- from pupils' reactions it appears that everyone knows what he or she has to do
- there is sufficient control on pupils doing the assignments they are supposed to do
- pupils work at a good pace
- % of time during lessons in which assignments are discussed
- analysis of mistakes
- checks on homework

Monitoring

- is monitoring of pupils' achievement mentioned in the school work plan
- indications concerning:

– pupils' written assignment

– the use of tests

- % of lessons containing tests
- the number of tests, hearings
- types of tests per school year (a.o. posing questions in class, own tests, curriculum-embedded tests)
- which procedures are used to assess pupils' achievement with respect to arithmetic
- progress in pupil learning outcomes is measured by means of (curriculum-embedded) tests
- teacher uses checklist for oral hearing of pupils
- the way the teacher prepares pupils for tests
- teacher checks whether all pupils have reached the minimum goals
- teacher checks up on difference between expected and actual pupil achievement
- compare pupil achievement to:

– former pupil achievement

– fellow-pupil achievement

– norms and standards

- in what way is arithmetic/math work of a pupil judged (absolute criterion, class average etc.)
- are test results used for individual help, extra explanation

- taking action in connection with test results
- use learning progress for:

– preparing a program for individual pupil

– reporting to parents

– informing teacher about next group

– evaluating the school's functioning

– putting pupils into (parallel) classes

– selecting pupils for teaching programs (enrichment/remediation)

– grouping pupils within classes

– other

• the degree of pupils' progress has an effect on class level (e.g. other grouping patterns, more or less instruction etc.)

• successes/problems with respect to preparation for tests over the past 5 years

• review and correct written assignment of pupils

• use of curriculum-embedded tests

• use of curriculum-independent tests

• use of self-made tests

The main sub-factors in "structured instruction" are basic requirements of well-prepared and well-controlled teaching on the one hand an aspects of direct instruction on the other.

12.12 Independent Learning

Next to the direct instruction perspective, a new instructional paradigm based on constructivism has emerged. It emphasizes independent learning, use of meta-cognitive skills and learning embedded in authentic assignments and "real life" situations. The elements of independent learning that were found in some of the instruments that were analyzed, are summarized in Table 12.12.

12.13 Differentiation

Differentiation is aimed at instruction that is adaptive to the specific needs of subgroups of pupils. The success of differentiation is to a large extent dependent on school and classroom organization. Crucial intervening variables are time on task, and the quality of tuition during group work etc. Elements of differentiation are summarized in Table 12.13.

Table 12.12 Components and Elements of Independent Learning.

• attention for independent learning in school work plan
• teacher-independent learning is being encouraged yes/no
• if yes, indications concerning:
− relation instruction/processing time
− organization of independent learning
− other types of differentiation
• state of affairs with respect to teacher-independent learning/independent learning
• the extent to which pupils are responsible for their own work
• the extent to which pupils are responsible for their own work during a longer period
• the extent to which pupils are able to chose their own assignments
• the extent to which pupils' cooperation is encouraged by teachers
• in case of independent learning, do pupils work:
− on the same subject
− on various subjects per group of same level
− on the same subject at own level
− on various subjects at various levels
• opportunity for pupils to plan the school day themselves
• successes and problems with respect to teacher-independent learning/independent learning

Table 12.13 Elements of Differentiation.

• attention for differentiation in school work plan
• indications for differentiation concerning:
− instruction
− processing
• minimum goals per class for all pupils
• use of differentiation model: if yes, which one
• application of setting/streaming with respect to capacities in the school/ department
• how to deal with differences between pupils in arithmetic/math attainment levels during lessons (all pupils the same subject matter…)
• % of lessons in which pupils:
− work on the same subject

− work on two subjects

− work on three or more subjects

• how often do pupils work individually or in pairs

• % of teacher time spent on communication with the class, groups and individuals

• criteria with respect to subject matter provision/grouping:

− achievement

– results standardized test

– results diagnostic test

– results oral test

– teachers' recommendations

– parents' wishes

– pupils' wishes

– method's demands

• pupil grouping within the class:

– no grouping

– age groups

– level groups

– interest groups

– other

• frequency of regrouping pupils (evt. of more classes) on behalf of level groups

• problems and successes with respect to differentiation the past 5 years

• subject matter mastery adapted to slow and fast learners

Special attention for pupils at risk

• policy with regard to low-achieving pupils

• school policy is explicitly aimed at catering for a wide range of educational needs: in other words, clear directives and structural attention for pupils with problems

• catering for special individual educational needs concerning:

– diagnosing pupils "at risk"

– remedial teaching

– cooperation with special education

– drafting intervention plans

− drafting group plans

• amount of extra time teachers are prepared to spend on problem pupils
• extra provisions for problem pupils
• low-achieving pupils get more time for reflection, extra attention, instruction, help, material and exercise material
• provisions/approved methods for preventing (teaching) problems
• check systematically which subject matter is not being mastered
• group teachers having expertise with regard to diagnostic test administration
• group teachers are able to translate test data into intervention plans

The selection of elements is somewhat colored by the strong current focus on taking care of pupils with learning and behavioral problems in regular (as opposed to special) primary schools in the Netherlands.

12.14 Reinforcement and Feedback

Reinforcement and feedback are important basic conditions for learning. Elements of instruments are summarized in Table 12.14.

Table 12.14 Elements of Reinforcement and Feedback.

Reinforcement
• is feedback in connection with pupils' achievement discussed in the school work plan
• indications for feedback in connection with pupils' achievement are related to discussion by the teacher
• how often, in arithmetic/math lessons, do you take the following action when pupils answer wrongly (a.o. correct wrong answer, pose different question)
• during the lesson feedback is given and pupils' mistakes are corrected
• when pupils carried out an assignment it is discussed immediately
• the teacher explains what was wrong when he returns the tests
• teacher gives pupil as much as possible real and positive feedback to achieved results
• frequency of discussing learning progress with pupils
• low-achieving pupils get extra feedback
Feedback
• results written assignment is discussed with pupil if necessary
• results curriculum-embedded test are discussed with pupil if necessary
• results method-independent tests are discussed with pupil if necessary

• results of self-made tests are discussed with pupil if necessary
• a differentiated supply based on tests is offered
• quality/suitability of feedback
• state of affairs with respect to giving constructive feedback now and 5 years ago
• problems with respect to inadequate feedback

It should be noted that reinforcement and feedback have both cognitive and motivational implications, as a basic requirement in learning and in rewarding exertion and good performance.

12.15 Summary and Conclusions

The main components of each of the fourteen general effectiveness-enhancing factors are summarized in Table 12.15.

Table 12.15 Components of Fourteen Effectiveness-Enhancing Factors.

Factors	**Components**
Achievement, orientation, high expectations	• clear focus on the mastering of basic subjects
	• high expectations (school level)
	• high expectations (teacher level)
	• records on pupils' achievement
Educational leadership	• general leadership skills
	• school leader as information provider
	• orchestrator of participative decision making
	• school leader as coordinator
	• meta-controller of classroom processes
	• time educational/administrative leadership
	• counsellor and quality controller of classroom teachers
	• initiator and facilitator of staff professionalization
Consensus and cohesion among staff	• types and frequency of meetings and consultations
	• contents of cooperation
	• satisfaction about cooperation
	• importance attributed to cooperation
	• indicators of successful cooperation

Curriculum quality/opportunity to learn	• the way curricular priorities are set
	• choice of methods and text books
	• application of methods and text books
	• opportunity to learn
	• satisfaction with the curriculum
School climate	*a) orderly atmospheres*
	• the importance given to an orderly climate
	• rules and regulations
	• punishment and rewarding
	• absenteeism and drop out
	• good conduct and behavior of pupils
	• satisfaction with orderly school climate
	b) climate in terms of effectiveness orientation and good internal relationships
	• priorities in an effectiveness-enhancing school climate
	• perceptions on effectiveness-enhancing conditions
	• relationships between pupils
	• relationships between teacher and pupils
	• relationships between staff
	• relationships: the role of the head teacher
	• engagement of pupils
	• appraisal of roles and tasks
	• job appraisal in terms of facilities, conditions of labor, task load and general satisfaction
	• facilities and building
Evaluative potential	• evaluation emphasis
	• monitoring pupils' progress
	• use of pupil monitoring systems
	• school process evaluation
	• use of evaluation results
	• keeping records on pupils' performance
	• satisfaction with evaluation activities
Parental involvement	• emphasis on parental involvement in school policy

	• contacts with parents
	• satisfaction with parental involvement
Classroom climate	• relationships within the classroom
	• order
	• work attitude
	• satisfaction
Effective learning time	• importance of effective learning
	• time
	• monitoring of absenteeism
	• time at school
	• time at classroom level
	• classroom management
	• homework
Structured instruction	• importance of structured instruction
	• structure of lessons
	• preparation of lessons
	• direct instruction
	• monitoring
Independent learning	no sub-components
Differentiation	• general orientation
	• special attention for pupils at risk
Reinforcement and feedback	no sub-components

The range of components within factors in several cases shows that effectivenessenhancing conditions are measured in terms of:

a. priorities assigned to factors and components; i.e. attitudes, beliefs, goal statements;
b. factual state of affairs relevant to factors and components;
c. appraisal and judgement on the degree to which factors and components are realized.

Particularly with respect to the latter category (appraisal) there is the danger of reactivity in the measurement of (hypothetical) effectiveness-enhancing conditions, because the judgement on processes and antecedent conditions may be colored by knowledge about outcomes and “dependent variables”.

The divergence in choice of elements for instruments across sources (i.e. instruments used in school effectiveness studies and school diagnosis instruments) is somewhat inflated, because there are sometimes rather slight differences between elements. It

should also be noted that divergence at item-level does not preclude that elements will be correlated and be shown to be subsumable under common headings, also by means of data-analytical procedures like factor-analyses. On the other hand it is quite clear that there is little agreement, at the operational level, on the substance of the key factors that are supposed to determine school effectiveness.

A further observation is that most of the factors are broad, in the sense that there is a wide range of components and elements. This is particularly the case for educational leadership and school climate. The broadness of the factors makes it hard to decide which of the set of elements is supposed to be crucial in enhancing effectiveness. Both the divergence and the broadness of the factors makes summary review and qualitative research synthesis rather hazardous, because operationalizations of the same general factor may be quite different across studies.

A third and final observation is that the factors are not mutually exclusive. Zones of overlap exist between:

- achievement orientation in policy and climate;
- evaluative potential and monitoring as an aspect of structured teaching;
- curriculum aspects and coordination and consensus;
- educational leadership and use of students' records (also an aspect of evaluative potential);
- participatory decision-making and consensus.

The modes of schooling that are most strongly represented in the set of instruments that was analyzed are:

- school policy;
- management/leadership;
- climate;
- curriculum;
- instruction;
- relations with parents.

Modes like financial inputs and professional development of teachers were underrepresented in the set of instruments that was analyzed.

The main data provider for the instruments that were analyzed is the school leader, followed by the teacher. For some instruments inspectors or head of departments were the data provider. Most instruments are written questionnaires asking for self-reports from head teachers and teachers. Direct observation and structured content analysis of documents occurred in a small minority of cases.

The main conclusion from this analysis of instruments used in school effectiveness research is that there is great divergence among studies, that each project leader appears to be re-inventing the wheel in the area of instrument development for measuring effectiveness-enhancing school and classroom variables and that there are no commonly used standardized research instruments to measure factors that are supposed to be the core of effectiveness-enhancing conditions.

Despite the need to adapt the choice of measurements and instruments somewhat to local circumstances, it appears to be worthwhile to try and develop a set of core indicators on the most promising antecedent conditions of school effects. The

development of process indicators within the framework of the current OECD education indicator project can be seen as a first attempt to achieve this task in an international comparative context (Scheerens & Ten Brummelhuis, 1996).

13
Educational Indicators of Value Added

13.1 Introduction

In the not too distant past educational indicators were seen as fairly simplistic in nature and defined as straightforward quantitative measures of different aspects of an education system. Most commonly indicators were aggregated at the national level and in some cases at the regional, local district or school level to provide a basic summary of educational provision, take-up and costs. However, in the last ten years or so a more complex picture has developed with indicators falling into distinct categories of input, process, context and output data. This work has drawn to a large extent on the OECD indicators project (INES) which aims to develop a comprehensive system of educational indicators in four different aspects—student learning outcomes, education and labor market destinations, schools and school processes and attitudes and expectations of stakeholder groups in education (OECD, 1995). However, Scheerens (1999) has argued that current educational indicator systems are limited in approach as they have "…no aspiration to "dig deep", while employing easily measured characteristics and so-called proxy measures." He goes on to suggest that:

> Another "danger" is the use of process or throughput data as evaluation criteria, instead of explanatory conditions of educational outputs. This could easily lead to goal displacement, where the "means" in education are treated as "goals" in themselves. A technical limitation which might encourage this improper use of process indicators is the fact that the question of relating process and output indicators by means of formal statistical analysis has hardly been tackled for applied purposes. (Scheerens, 1999).

In terms of Scheerens' point in relation to the applied use of educational indicators this chapter aims to examine the recent development of educational indicators that are more sophisticated and precise in their meaning and interpretation, particularly for the purpose of school evaluation. Thus, this chapter will focus on the development of what are often called value added indicators of school effectiveness—a key aspect of policy development work in a number of countries worldwide.

The background to the development of educational indicators that are valid in the applied context of evaluating schools and systems as a whole stems from research carried out since the 1960s on the impact of schooling on student performance and the related areas of school effectiveness and improvement. In the last decade or so, the accumulation

of convincing research findings in this area as well as significant advances in methodological techniques have re-stimulated international debate and government thinking on educational policy and practice. As a consequence national (and regional) policy makers in several European countries and worldwide have focused their attention on the possibilities for improving educational practice, and indirectly, educational standards, competitiveness and performance, by encouraging more systematic approaches to school and teacher evaluation and self-evaluation (Reynolds et al., 1996). For example, the current UK labor (and previous conservative) governments have emphasized the need for schools and teachers to use evidence and data to inform their own internal evaluations of the education they provide (DfEE, 1996; DfEE, 1998a). This approach involves an on-going and systematic self evaluation of a school's educational practice and improvement processes using information drawn from a variety of sources. Moreover, at the regional level, local education authorities (LEAs) in England now have a statutory role in monitoring the quality of education and improvements in all schools in their region. At the national level, a relatively new system of external school inspections has been introduced (Matthews & Smith, 1995) and the school examination performance tables have continued to be published by the Department of Education and Employment (since 1992) as a mechanism for educational accountability. Plans are also scheduled for autumn 1998 to provide further detailed information for the purpose of school improvement to all English primary and secondary schools (SCAA, 1997; DfEE, 1998b). Multinational educational indicator systems such as OECD/INES are also now beginning to incorporate value added approaches in a move to enhance the validity of indicators for the purpose of national comparisons of school effectiveness.

These policy developments concerning both external and internal school evaluation have been informed, in part, by the findings of quantitative research studies which share a common aim; to separate and measure the school effect and that of other external factors such student prior attainment and socioeconomic status. This research utilizes what is frequently called the *value added concept* and defines school effectiveness in terms of specific student outcomes—the *relative* progress of students in a school over a particular period of time in comparison to students in other schools (Thomas & Mortimore, 1996; Thomas, 1998). In the context of these studies, school improvement is also defined specifically in terms of improvement in value added performance.

This chapter will describe the methodology of value added measures of student progress and illustrate how this approach provides an important innovation in both external school evaluation and internal self-evaluation. Further innovations in school evaluation using questionnaire data and qualitative approaches are described in the next section.

13.2 The Value Added Concept

The value added concept rests on the assumption that schools add 'value' to the achievement of their students. It is based on the idea of measuring student progress, usually in cognitive outcomes such as reading or mathematics attainment during a given period of time. However, the concept can also be applied to non-cognitive outcomes such as attitude scales or measures of vocational competence. In order to measure progress

baseline and outcome measures are required at the beginning and end of a particular time period (for example covering all or part of the primary or secondary phases of education). Of course, as students grow older one would expect progress or improvement to be made and average attainment levels to rise. Therefore, researchers use the term value added to refer to the extra value that is added by schools to student attainment over and above the progress or improvement that might be expected in a normative sense. Value added measures thus seek to establish whether students in some schools make relatively greater or less progress than those in other schools over a specified period of time. The most effective of schools would be those in which student progress exceeds expectations. Therefore, in the context of internal school self-evaluation, the purpose of value added measures is to provide teachers with meaningful, valid and accurate evidence of the relative progress of their own students. The aim is that these measures can be used by teachers to inform and reflect on their professional and educational practice.

However, in the high stakes context of external school evaluation and accountability, value added measures provide only one source of comparative information about a school's effectiveness. It is important to remember that the value of schools' educational quality is broader than what can be measured by attainment in a few specific areas of student activity. A comprehensive value added evaluation framework might also encompass measures related to numerous other aspects of a school's mission, processes and outcomes (see for example next section on innovation in the use of qualitative data for school evaluation). This broader approach to school evaluation and self-evaluation encapsulates a practical application of Scheerens' (1990) theoretical model of school functioning. In other words, similar data describing inputs, process and outputs is collected about individual schools but the primary purpose is that the information is used directly by school staff, and where necessary also external evaluators, to evaluate their educational policy, practice and improvement processes.

13.3 Data Collection Procedures and Issues

Calculating the effect a school has on student progress is complex. In part this is because the educational experiences of any individual student and the wide variety of factors influencing her or his progress can be viewed as unique and almost impossible to quantify. However, the more information it is possible to have about individual students, sub-groups of students, and all students in a school as well as comparative data across a whole population (or representative sample) of schools, the more reliable and informative any subsequent analysis is likely to be. As mentioned previously, the key evidence required to measure a student's academic progress is baseline and outcome attainment data over a specific period of time. Other background and contextual information about individual students and schools is also needed to provide statistical controls for those factors—outside the control of the school—that have a significant impact on a student's attainment, relative progress, or both (Sammons et al., 1994; Thomas, 1995).

The different types of quantitative data required to calculate value added measures of school performance are described below and Table 13.1 provides some specific examples.

Student Outcomes

Ideally student outcomes need to be measured using valid and reliable assessment instruments relevant to the academic curriculum taught in schools or other aspects of the curriculum (such as vocational training and qualifications). This is because it is important to measure only those aspects of a student's education which the school has a clear aim and statutory role to provide and develop. Of course, when aiming to measure school effectiveness it is crucial to bear in mind—and also to separate out where possible—the influence of teaching and learning at home, or otherwise outside the confines of the school (for example private tuition). It is also necessary that the assessment methodology employed to measure to student outcomes is unbiased and sufficiently reliable to ensure that students at the same level of attainment will be assessed at the same level irrespective of where the assessment takes place. For example, it may be argued that teacher assessed outcomes (such as portfolio assessment) are open to unequal standards being applied across schools and are therefore less appropriate for value added techniques than standardized and externally marked tests such as those from national examination systems and standardized attainment tests (Goldstein, 1993; Thomas et al., 1998b).

Table 13.1 Examples of Variables Required to Calculate Value Added Measures.

Student Assessment Outcomes:
• pupils' outcome attainments in maths and reading (for academic outcomes)
• pupils' outcome attitudes on five scales: engagement with school, self efficacy, behavior, pupil culture, teacher support (for attitude outcomes)
Student Assessment Baselines:
• pupils' prior attainments in maths and reading (for academic outcomes)
• pupils' prior attitudes on five scales: engagement with school, self efficacy, behavior, pupil culture, teacher support (for attitude outcomes)
Student Background:
• gender
• age
• entitlement to free school meals
• special needs
• learning support
• mobility
School Background:
• size
• status (public/private)
• type (single sex/co-educational)

School context:

- percentage of pupils entitled to free school meals
- region

Student Baselines

In order to measure a student's relative progress overall, or in any particular curriculum area, it is necessary to obtain information about their previous—or baseline—attainments, either at entry to a phase of education or some alternative and pre-defined starting point. Ideally, as a minimum, data is required concerning students baseline attainment in the core curriculum subjects of language or numeracy or both, and if possible other relevant aspects of the curriculum. Again, as noted in relation to outcome assessments, it is necessary that the baseline assessments are valid, reliable and fit for value added purposes. Also, baseline and outcome measures need to be collected and recorded in such a way that individual student records can be matched accurately over time. The use of unique student, class and school identification codes are vital in order to facilitate this matching process. Once a longitudinal data-set of individual student attainment records has been created it is then possible to proceed with a statistical analysis to examine both absolute attainment at one point in time and—most importantly—the *relative* progress of students in a particular school in comparison to students in the wider population (or a representative sample) of schools.

Background and Contextual Information

Information about student baseline and outcome attainment is absolutely essential to provide an accurate and direct measure of students *relative progress in attainment over time* and this approach is the *only* method of calculating valid value added measures of school effectiveness. However, the collection of other student and school data is also needed in order to examine the impact of various background and contextual factors that are *outside* the control of the school but that appear to influence the rate of student progress. Previous research has shown that student background characteristics such as socioeconomic status, gender, ethnicity, first language and mobility are in some cases statistically significantly related to student progress in attainment and therefore are able to provide a means of fine tuning value added measures of school effectiveness (Thomas & Mortimore, 1996). Similarly some previous studies have shown that school context data, such as the percentage of disadvantaged students in a school, has a statistically significant impact on student progress (Sammons et al., 1994). The key factor in planning the collection and analysis of additional student and school background and contextual information for a particular evaluation system is deciding whether a straightforward progress measure is required or, alternatively, a more sensitive value added measure that also controls for other factors that are outside the control of the school.

Of course, in the absence of progress data, the impact of socioeconomic and other background factors on student attainment is considerable at any one point in time (for example at either baseline or outcome). This finding is well documented and not surprising given the cumulative effects of such factors on children's educational

experiences since birth. Therefore, student and school background data may also be usefully employed to provide *contextualized* attainment measures at any one time point as an additional, or an alternative (but conceptually different and arguably less accurate) approach to measuring school performance.

Educational and School Process Information

Educational and school process information is not usually employed in the calculation of value added measures of school performance. This is because in contrast to the background and contextual data described above, measures of educational and school processes aim to quantify aspects of school life that are defined to be *within* the control of the school. Measures of this kind include, for example, teacher attitudes, experience and supply, class size, organization and grouping strategies, school ethos and organization. However, the rationale of collecting additional information about educational and school processes is to enable school and classroom process variables to be contrasted and evaluated against measures of students' relative progress (see MacBeath & Mortimore, 1994 or Sammons, Thomas & Mortimore, 1997 for examples). In this case, the aim is to provide evidence to explore and illuminate the reasons underlying differences in school effectiveness.

We have briefly outlined above the different types of the information required to calculate value added measures and Table 13.2 provides a summary of the data collection requirements for different value added and other school performance measures. The data requirements shown relate in each case to a single cohort of students. Of course, in order to examine trends over time or improvement in value added performance it is necessary to collect the same outcome and baseline data for at least three consecutive student cohorts. Similarly, data for consecutive student cohorts would also be necessary to examine trends in contextualized or raw performance measures.

Table 13.2 Data Requirements for Value Added and Other School Performance Measures.

Value Added Measures of Relative Student Progress	**Data Variables Required**
(1) Progress only measure	Outcome and baseline assessment measures
(2) Progress, background and context measure	Outcome and baseline assessment measures, student and school background and context measures
Other Performance Measures	**Data Variables Required**
(3) Contextualised attainment measure	Assessment measures at any one time point, student and school background and context measures
(4) Raw attainment measures	Assessment measures at any one time point

Overall, the pitfalls of collecting any type of quantitative data include the need for quality assurance procedures and accuracy checks to be put in place as a vital part of the data collection exercise. Of course, data errors or missing data, or both, will mean that the

results of any value added analysis will be difficult, if not impossible to interpret and these issues are discussed further in relation to the limitations of value added measures (see also Elliot, Smees & Thomas, 1998). Clearly schools need to collaborate with other schools at the local, regional and national level in order to provide comparative data in an identical format. For a detailed example, the case study of the Lancashire local education authority value added project illustrates how the on-going collection of examination and other data can be organized successfully at the regional level involving a total of 98 secondary schools (see Chapter 16).

13.4 How is Value Added Measured?

(i) Statistical methods for measuring value added

A key challenge for researchers has been to develop approaches which allow the statistical analysis to separate out the effect of the school experience on individual student outcomes (what students achieve) and the extent to which student intake characteristics (those things the students arrive at school with) affect student outcomes. A variety of statistical techniques can be employed for this purpose which vary in the sensitivity and sophistication of analysis. Table 13.3 provides a summary of three main approaches and some of the key advantages and disadvantages of the different techniques are also discussed.

Table 13.3 Summary of Statistical Methods.

Statistical Technique	**Unit of Analysis**	**Statistical Measures**
Summary Statistics	School	Mean, standard deviation
Multiple Regression	*Either* student or school	Residual (difference between observed and expected score)
Multilevel Modeling	*Both* Student and school	Residual (difference between observed and expected score)

The first approach, summary statistics (including the mean and standard deviation), can be used to calculate aggregate school level measures from student level data. These measures provide a crude picture of school performance via estimates of 'raw' levels of student attainment for each school in a sample (i.e. unadjusted for any other factors) and may be used, for example, to produce simple league tables of average school performance at one point in time (such as those published by the DfEE in England). Of course, the main disadvantage of this kind of approach is that student progress in attainment cannot be evaluated. A further disadvantage of this method is that the school is the unit of analysis and therefore detailed information about individual students is lost in the analysis.

The second approach, multiple regression analysis, is the standard statistical technique for calculating the residual difference between an observed and expected score. In the

case of measuring individual student progress, the observed score is a student's actual level of attainment and the expected score is the level that would be predicted on the basis of his or her previous—baseline—attainment. Consequently the residual score (which ranges from a positive to negative value) is interpreted in terms of whether a student is performing above or below expectation (on the basis of the overall statistical relationship between baseline and outcome attainment of all students in all schools in the sample). In essence, the residual score provides the key statistical measure of student's relative progress—the value added. One advantage of this approach is that several factors, such as baseline attainment and other student characteristics like gender and socioeconomic status, can be employed in the analysis to provide a more sensitive estimate of value added than would result from employing a single baseline predictor. However, the disadvantage of this approach is that the unit of analysis has to be either at the level of the student (i.e. where student residual scores are calculated) or the school (i.e. where school residual scores are calculated). In the former case important information about the clustering of students within a particular school is lost, and in the latter case, detailed information about individual students is lost.

The third approach, multilevel modeling, is a recent generalization of multiple regression which involves the same principle of calculating a residual value added score. However, this new technique takes account of the clustering of students within schools and allows the unit of analysis to include both the student and the school level. Thus multilevel modeling is a far more sophisticated approach than both summary statistics and standard multiple regression when the aim is to disentangle the complexity of schools' effectiveness and it is now widely recognized as the most flexible tool for examining the hierarchical nature of student attainment data (Goldstein, 1995, 1997). This approach can be used to calculate unbiased and accurate estimates of school residuals for all students (or particular groups of students such as boys or girls) as well as, crucially, the statistical significance of an individual school's results. If data is available for consecutive student cohorts this technique can also be employed to model trends in value added results over time.

(ii) Providing a realistic picture of performance

We have described briefly the various statistical techniques for examining school performance. However, we now turn to the important issue of how these methods can be used to reflect the full complexity of school effectiveness. Previous research has indicated that in order to provide a realistic picture of a school's performance a range of different value added measures are required to show the internal variations in school effectiveness across one or more dimensions (e.g. student outcomes in different aspects of the curriculum, different groups of students, different periods of time and different regions).

Different aspects of the curriculum

Of the previous studies which have examined schools' effects on different outcomes most have focused on the performance in the areas of English and mathematics (at the secondary level see Willms & Raudenbush, 1989; Goldstein et al., 1993; Thomas &

Mortimore, 1996 and at the primary level see Sammons, Nuttall & Cuttance, 1993; Thomas, 1995). The findings have indicated that schools doing well with students in one aspect are not necessarily effective in all aspects. Similar conclusions have been drawn from a study in England looking at a wider range of outcomes: six GCSE subjects and one overall GCSE measure (Thomas et al., 1997a). These findings are also reflected in research carried out in the Netherlands (e.g. Luyten, 1994) and at the post 16 level (Fitz-Gibbon, 1991). This evidence strongly suggests the need to examine school effectiveness measures across a range of academic outcomes in order to reveal the pattern of departmental or subject area performance. Clearly, using a single measure effectiveness may conceal important within school differences not only across academic aspects of the curriculum but also other aspects such as vocational or attitudinal outcomes (Scheerens, 1992; Sammons, Thomas & Mortimore, 1997; Scheerens & Bosker, 1997; Thomas, 1998).

Different periods of time

With regard to the stability, or instability, of school effects over time, the importance of this aspect of school effectiveness has been established by several researchers (e.g. Willms & Raudenbush, 1989; Gray et al., 1995; Gray, Goldstein & Jesson, 1996; Thomas et al., 1997a). In general the evidence indicates that for most schools performance is broadly similar over time but for some schools results can vary substantially indicating either improvement or decline in performance. In this context, it is important to emphasize that 'real' improvement (or decline) in performance, resulting perhaps from a shift in educational policy or practice, can only be identified by examining long-term changes in results over time (Gray, Goldstein & Jesson, 1996). Recently, researchers have noted the importance of examining in detail the performance trends of individual schools and the educational processes associated with different patterns of improvement (Gray, Hopkins & Reynolds, 1998).

Different groups of students

Research in the late 1980s also examined the issue of differential school and departmental effects for different groups of students (such as high and low attainers, boys and girls or different ethnic groups) and found that an important aspect of a school's effectiveness was whether it was equally effective for *all* student groups. However, these studies were somewhat limited in the number of schools investigated or the availability of detailed information about the background and prior achievements of the student sample. More recent research, employing detailed student level data, has confirmed and extended previous findings on this topic and established that using an overall measure of school (or departmental) performance may mask important differences in the relative progress made by different student groups, particularly those categorized by prior attainment (O'Donoghue et al., 1997; Thomas et al., 1997b; Thomas, 2001).

Different regions

Currently only a few studies have addressed the issue of regional or national differences in the size, extent and consistency of school effects or the differential impact of pupil and school background characteristics in different regional, socioeconomic and educational policy contexts. Evidence of this kind is vital to inform educational policy makers about the influence of local area, regional and national policy and practice. Gray, Jesson and Sime (1990) has compared the value added estimates for schools in six different LEAs in the UK and found substantial differences between the estimates of school variation (after controlling for student intake) for different regions. However, the conclusions that can be drawn from these comparisons are limited due to differences in the controls employed for student intake (4 LEAs were lacking prior attainment data) and the small size of school samples (30 or fewer schools in 5 LEAs). At the international level Scheerens et al. (1989) has examined the variance in student outcomes at the school and classroom level across 17 countries involved in the IEA[1], an international comparison study of mathematics attainment. The results showed considerable variation across nations in the percentage of between school and between class variance in pupil outcomes. Also, Creemers, Reynolds and Swint (1994) have described a comparative study involving 5 countries, focusing mainly on primary mathematics which is part of the on-going International School Effectiveness Research Programme (ISERP). Although this study is severely limited due to the very small samples of schools in each country (12 or fewer) the findings show important differences between countries in the size and extent of school effects after controlling for student intake. Creemers, Reynolds and Swint (1994) underline the need for further research to investigate systematically the existence and reasons underlying regional and national differences in school effects with larger samples of schools. Finally, new and very recent research has also indicated that the difference or similarity in schools' departmental results can vary across UK regions (Thomas, 2001).

Thus, overall the evidence suggests strongly the need for further evidence about school performance over time and in detail for different student groups, not just in terms of total performance but also at department (or subject) level, as well as in other outcome areas (such as vocational and affective/social) in order to describe the full complexity of school effectiveness. Moreover, the apparent differences in the range and extent of school effects, both across and within national boundaries, indicates the importance of examining separately regional and national indicators of school effectiveness as well as the educational policies that may underlie any differences observed.

(iii) New Developments in value added measures

New developments in value added research have focused on the issue of the continuity of primary school effects at the secondary level (Sammons et al., 1996; Goldstein & Sammons, 1997). Initial results indicate a lasting impact of primary school effectiveness on students' progress in secondary school. In other words, students from primary schools where the learning and teaching was effective appear to continue to make better progress at secondary school than students from less effective primary schools. This last point is also directly relevant in dealing with the fairly frequent occurrence of students changing schools *within* an educational stage. Hill and Goldstein have recently argued that this

issue has important implications for the accuracy of a school's effectiveness measures which relate to a period of time

[1]IEA=International Association for the Evaluation of Educational Achievement.

when many students may have left and other new students have started (Hill & Goldstein, forthcoming). Therefore new developments in the methodology of calculating value added measures may need explicitly to take into account previous schools attended by individual students. In the meantime, value added data should be viewed cautiously in the context that previous school effects need to be clarified via further research.

(iv) Interpretation and the Limitations of Value Added Measures

So far we have outlined the methodology of calculating value added measures and importance of providing effectiveness indicators in range of different dimensions or areas. However, when trying to make sense of value added measures it is crucial to emphasize the statistical uncertainty and limitations of any numerical data so as to avoid over-interpretation of the results. By the term—*statistical uncertainty*—we mean the uncertainty involved in estimating any average numerical score from a sample of observations, scores or measurements. Thus, when measuring school effectiveness, an individual school's value added results are estimated from the relative progress in attainment made by the sample of students in the school. This uncertainty prevents any fine distinctions to be made between the performance of most schools (see Goldstein & Healy, 1995; Goldstein & Spiegelhalter, 1996).

It is also important to consider the issue of *measurement error* when interpreting data based on measures of student attainment. *Measurement error* is the error associated with trying to obtain a 'true' measure of an individual student's attainment from an 'observed' measure of their attainment at one specific point in time. For example, if a student is distracted from an assessment task they are unlikely to complete the assessment as well as if they were fully engaged with the task. In this case, the measurement error is the difference between their 'true' level of attainment in completing the task and their 'observed' attainment.

Thus, the value of school effectiveness measures is defined to a large extent, by the quality, reliability and validity of the data analyzed. Another issue which is difficult to address is the accuracy and appropriateness of the data. For example, the indicator of student disadvantage 'eligible for free schools meals' (FSM) may be inaccurate because some parents do not apply for an eligibility means test. Furthermore, the system does not cover completely all students likely to suffer from social and economic disadvantage. Other relevant measures of socioeconomic status, such as level of parental education, occupation and income are difficult and costly to collect. Nevertheless, FSM is currently the most readily available, easily updated measure of socioeconomic disadvantage among school children.

All the above caveats and limitations point to the importance of considering the statistical significance of individual school results (and also where possible the stability of results over time) as well as other relevant data or evidence that may be available in a school in order to avoid over-interpreting the results.

13.5 Presenting Value Added Results

Using a multilevel value added approach we are able to create a profile of each school's effectiveness in a range of different areas or dimensions (e.g. across academic and attitude outcomes and across measures for different groups of students). However, the practical issue of how value added results can be presented to aid valid interpretation of the results also needs to be addressed.

Table 13.4 Profile of School X: Value Added Performance.

Outcome	Residual and Effectiveness Category
Language	-0.5----------|----------*----------0.5 0.16 (average positive)
Mathematics	-0.5----------|----------*----------0.5 0.21 (effective)
Science	-0.5----------*----------|----------0.5 -0.17 (average negative)
Language (High attainers – top 50%)	-0.5----------|----------*----------0.5 0.18 (average positive)
Language (Low attainers – bottom 50%)	-0.5----------*----------|----------0.5 -0.21 (average negative)
Engagement with school	-0.5---*----------|----------0.5 -0.38 (ineffective)
Self Efficacy	-0.5----------|--*----------0.5 0.03 (average positive)
Pupil Culture	-0.5----------|--------*----------0.5 0.09 (average positive)
Behavior	-0.5----------|----------*----------0.5 0.16 (average positive)
Socioeconomic Context (%FSM)	
School X	0----------*----------|----------100 32%
National Average	0--------*----------|----------100 9%

Note: (1) Table adapted from Thomas et al. (1998a).
(2) All pupil outcome measures have been transformed to normal scores with mean of zero and standard deviation of one.

An example of one school's profile of results is shown in Table 13.4. The residual figures shown indicate the number of standard score units above or below expectation the typical student in a school is achieving in comparison to students in other schools (after

controlling for baseline attainment and other background factors). However, given the statistical limitations of the methodology and the fact that it is impossible to take account of all factors outside the control of the school, it is important that the school residuals are always treated with some caution. Therefore, one useful approach to facilitate the interpretation of schools' effectiveness—for example when examining the impact of school's educational and improvement processes—is to categorize each school's residuals into four groups (also shown on Table 13.4).

[1]	Effective	Positive residual—significantly different from zero ($p<0.05$)
[2]	Average (positive)	Positive residual not statistically significant
[3]	Average (negative)	Negative residual not statistically significant
[4]	Ineffective	Negative residual—significantly different from zero ($p<0.05$)

In addition to examining schools' effectiveness categories previous research indicates that an important contrasting dimension in any evaluation framework should reflect explicitly the socioeconomic context of each school (e.g. Mortimore & Whitty, 1997). Therefore, contextual information about the relative level of student disadvantage is also presented on Table 13.4 (i.e. the percentage of students in the school entitled to free school meals). However, when interpreting the results it is important to bear in mind that socioeconomic context is difficult to define and measure accurately. The only indicator commonly available in England is the fairly crude contextual variable percentage of pupils entitled to free schools meals. Therefore, the methodology may be further improved by using a more sensitive or appropriate indicator of disadvantage, such as level of parental education.

13.6 Using Value Added Indicators as a Valid Tool for School Evaluation and Self-Evaluation

Value added data are helpful for school evaluation and self-evaluation by raising questions about changes and/or consistency in results over time, highlighting differences between individual departments in a school compared to the whole school value added, and allowing schools to compare themselves with other schools. A key aspect of this approach is encouraging schools' ownership of the data and ensuring confidentiality of the results (see Robertson et al., 1998). The following points summarize approaches to stimulate school and teacher self-evaluation.

Using data at the individual pupil level

- For individual students and specific groups of students (such as boys or girls or certain ethnic groups) value added results and predicted grades can provide guidance in monitoring and target-setting. However, the results should be used cautiously for an individual pupil, bearing in mind other information about an individual's particular circumstances, and the fact that past performance does not necessarily predict future performance.

Using data at the classroom level

- Examine class level progress of students and average/spread of predicted grades to inform teaching strategies and level of work.

Using data at the department or subject level

- Examine departmental, subject and/or teacher effectiveness versus summary measures of school effectiveness and the implications for whole school policies.

Using data at the school level

- Employ a wider range of value added measures to reflect more fully the aims of schooling (e.g. using student attitudes and vocational as well as academic outcomes).
- Contrast the results against other types of data available in schools such as information about the views of key groups obtained using for example teacher and parent questionnaires.
- It is important to consider the importance of confidence limits when making any comparisons between schools—if the confidence intervals of two particular schools overlap then there is no *statistically significant* difference between their performance.
- Bear in mind limitations of the methodology for individual schools. How relevant are issues of: measurement error, missing data, data accuracy and the retrospective nature of the data?

Using data at the regional level

- Examine local or regional differences in value added results between schools and the implications for local or regional education policy.

Using data at the national level

- Examine the national and/or international profile in value added results between schools and the implications for national education policy.

Using data to examine context

- Consider the local, regional or national context of the school. Value added measures cannot adjust for *all* factors outside the control of the school. Therefore contextual information such as raw baseline and outcome data and the overall socioeconomic disadvantage of students provide essential information to contrast against overall value added performance.

Using data to examine improvement

- Track changes in results over time to examine real improvements, or random fluctuations in performance, or both, in relation to school improvement initiatives.

• Examine the way teachers use data to reflect on past performance and to inform, evaluate and improve their current policy and practice.

Using data to examine equity

• Examine differential effectiveness for different groups of students (e.g. boys/girls, high/low attainers) and implications for equal opportunities.

Using data to examine curriculum

• Examine effectiveness for different year groups or age cohorts (e.g. students at different stages of a curriculum) and implications for differing rates of progress/ curriculum coverage.

13.7 Implementing a Value Added System of Evaluation: the Role of the School, Local and Central Education Authorities

The need for schools to analyze data in a more sensitive and detailed way has been emphasized, at a range of levels: individual students; various student groups; subgroups; subject level; whole school and across schools regionally or nationally. However, in order to implement a value added system of school evaluation and/or self evaluation it is important that schools are willing to collaborate with other schools at the local, regional and national level in order to provide comparative data. This approach involves the central organization of collecting, analyzing and presenting value added and other comparative feedback information (such as teacher, parent or pupil questionnaire data) to schools in a common format for the purpose of supporting and stimulating school self evaluation and improvement activities. As exemplars, numerous LEA projects are currently in progress in England and at the national level the Qualifications and Curriculum Authority (QCA) value added national project has published recommendations for providing schools with feedback on pupil attainment (SCAA, 1997). Additional examples are provided by the confidential value added projects—providing a variety and feedback data to schools—set up by the Universities of London and Durham and the National Foundation for Educational Research.

13.8 A Research Agenda

Previous research evidence from England has shown that overall raw statistics of student performance alone cannot give an accurate picture of how effective a school is at raising and maintaining the achievement of all its students, or how capable it is of sustaining its standards over time. The availability and analysis of longitudinal individual student level data is essential to allow schools and teachers to examine different aspects of their school's effectiveness and this has been illustrated by numerous studies in Europe (such as in the UK and Netherlands). Some schools that may appear to be effective in terms of the overall value added measure may not be so effective in terms of individual

departments, for different groups of students, different year groups or over different periods of time. Overall these findings indicate that internal variations in effectiveness need to be monitored at all stages of statutory education.

Value added approaches can also provide a powerful methodology for understanding the mechanisms and levers of school improvement. These innovative quantitative techniques of school and teacher self-evaluation can be employed at any stage or phase of schooling (primary, lower/upper secondary or post statutory education). The analysis—and feedback to schools—of value added data is frequently seen as an integral part of the development and improvement work carried out with schools by external consultants, school inspectors and academic researchers. A key aspect of this approach is encouraging schools' ownership of the data and ensuring confidentiality of the results. Examining the way teachers use data to reflect on past performance and to inform and evaluate their current policy and practice is a crucial aspect to understanding how schools improve. For this reason Chapter 16 provides a detailed case study of how one secondary school uses value added indicators to evaluate school and teacher performance and improvement as well as other aspects of school life.

The following key questions are relevant in terms of a future research agenda:

1. Additional dimensions of school effectiveness. What outcomes of schooling are valued and therefore need to be evaluated, in addition to those frequently reported in school effectiveness research?
2. How can effectiveness at different levels within the national education systems be measured and fed back to schools and what is the relationship between effectiveness at different levels—national, regional, local, school, department, classroom, individual?
3. What is the long term impact of school self evaluation processes on the quality of teaching and learning?

PART 5
Inspection and School Self-Evaluation

14
Monitoring on the Basis of School Inspections

14.1 Introduction

In the previous parts of this book the basic concepts and assessment methods of educational evaluation were explored and described. In this chapter, we move on to the practical application of these concepts and methods for the purposes of monitoring and inspection.

The debate on the value of school inspection as a means of monitoring the quality of education and as a tool to drive up standards is of major interest to educational policy makers and practitioners across Europe and worldwide. Key concerns relate to the challenges, problems, responsibilities and tasks for school inspection in the 21st century, both nationally and internationally and these were the main themes of a recent international meeting—Inspecting in a New Age organized by the Netherlands Inspectorate of Education for the inspectorate's 2000th anniversary (Troost, 2001). In particular the meeting focused on the internationalization of inspection as a profession and on internationalization of the outcomes of inspections.

In response to these themes Osler (2001) indicated that the most valuable asset of school inspection is when it can bring a positive influence on improvement in the quality of the learner's experience. He stated: 'It is not enough for inspection simply to lead to a report; it is necessary now to evaluate in order to bring about improvement… The professional credibility of an inspectorate…comes in large part from demonstrating a positive influence on improvement'.

However, a key feature and current issue concerning the processes and development of inspection systems is the relationship between external evaluations the main function of school inspectorates—and internal evaluations often referred to as school self-evaluation. Osler (2001) argues that external inspection is *'essential to a healthy education system'* and also that inspection *'is about ensuring that schools' self-evaluation does not become self-deception or self-congratulation'*. However, in the light of an increasing emphasis on equity and inclusion within some education systems, coupled with many education systems allowing more flexibility and autonomy in decision-making there now appears to be a weight given to school selfevaluation in some countries to enable a broader range of context specific quality criteria to be addressed. Interestingly, there has always been different approaches that can be applied in terms of the overlapping tasks and responsibilities of external and internal evaluations and these vary in different country contexts. For example, in England the new inspection framework (2003) puts a much greater emphasis on school self-evaluation than previously. Validating school self-

evaluation is seen as a major part of the inspection process and the nature and extent of inspections are 'differentiated' according to the evidence of schools' success.

We will return in detail to the nature and role of internal evaluation and school self-evaluation in subsequent chapters. However, in this chapter the aim is to focus on external evaluations and to review the key features of inspection systems. With this aim we mind, we would argue that broadly there are two crucial elements in the process school inspection. First, what criteria are employed in judging the quality of education; and second how school inspection is implemented. The former concerns the concept of quality of education; the second aspect relates to the methodology used to collect evidence and data about the quality of education and the quality of inspection (see also Standaert, 2000 for a discussion on this topic). It is not surprising that there are indeed a variety of different approaches taken by inspection systems in different countries in terms of the two elements referred to above. This chapter will provide a brief overview of these differences as well as a detailed case study of one particular system in England to illustrate the features and processes of inspection.

14.2 Systems of School Inspection

School inspection systems employed in different countries—particularly European countries—can be differentiated in terms of four specific features. The first feature is the inspection model or focus. The second is the outcome or output of the inspection process. The third relates to the length and intensity of the inspection process and the final feature is the position and location of the inspectorate within the overall education system. Each of these aspects is described below with concrete examples from different countries.

The inspection model or focus refers to the target of the inspection such as individual people (e.g. teachers, school managers and governors), institutions or systems (e.g. schools and local governors), subject areas (e.g. individual subject departments within schools) and thematic inspection (e.g. equal opportunities). The approach of inspecting schools as a whole originates from Great Britain and can be seen in the countries such as Flanders, Northern Ireland, Scotland, the Czech Republic and the Netherlands (Standaert, 2000). In contrast the approach of inspecting individual teachers rather than schools has been the focus of inspection in countries such as France and Greece. However, in relation to subject or thematic inspections, each country has its own interest regarding particular aspects of education, for instance, inspection of in-service training in the Netherlands, inspection of prison and youth offender services in England and inspection of vocational education in Northern Ireland.

Moving on to examine the outcomes of inspection systems—these also vary across countries. Different outcomes or outputs of the inspection process include both formative and summative evaluations in the sense that the former focuses on an advisory function and the latter focuses on an accountability function. However, generally there is a greater emphasis on the accountability function and it has been recently been argued that there is a need for all inspection systems to produce independent, publicly accountable, valid and reliable inspection results (Van Bruggen, 2001). Nevertheless, different aims and assumptions of the inspection process in different countries point to different outcomes and outputs as well as to different criteria being employed to measure the quality of

schooling or teachers. The difficult task of setting up criteria for the measurement of the quality of school should not be under-estimated. For example, one feature of the evolution of many education systems is decentralization where regions or schools have more power in decision-making (Osler, 2001; Dobart, 2001). One result of these shifts in policy is that different criteria may need to be established for different types of schools or regional contexts. This type of approach explicitly recognizes important differences in the qualifications, knowledge, and skills needed for learners' to be successful in different contexts (Kervezee, 2001). Dobart (2001) has also convincingly summarized the issues about quality criteria by arguing *'The inspectorate can only be able to fulfil its functions by being responsive, accountable, and by involving other parties in the development and adaptation of its own definition of quality'*. Echoing Osler's (2001) comments he goes on to state that inspectorates need to show that their *'…assessments of the quality of schools and of the system have added value for the improvement of quality in general and of the individual school in particular'*.

Of course, the outcomes or outputs of the inspection process also relate to the methods used to collect evidence to judge educational quality. A further key issue in this respect is the quality of the inspection process itself including the validity, reliability, accessibility and clarity of inspection judgements (Dobart, 2001). Both qualitative information and quantitative data is utilized to a greater or lessor extent in different countries. However, some countries have developed more sophisticated methods than others in collecting, maintaining and reporting evidence. For example, countries like the Czech Republic, the Netherlands and the UK have developed inspection databases. In particular, England has the most advanced database compared to other countries in Europe (Standaert, 2000). Nevertheless improvements to the methodology are also a concern, as Wim Kleijne pointed out in the Inspecting in the New Age meeting, when he argued that collecting qualitative data more systematically and developing better methods of analyzing this data is a challenge for the future of the Netherlands inspection system (Troost, 2001). A further aspect of the outcome of the inspection process which varies across countries is whether there is a follow up on inspection, an approach which was supported by countries that participated in a Standing International Conference of Inspectorates (SICI) workshop, held in Podebrady in 2000 (Drábek, 2000).

Turning now to the length and intensity of the inspection process—this feature refers to the period of time and extent of inspection resources (e.g. manpower) specified for each inspection as well as the interval between different inspections with the same target (e.g. school or teacher). For example, in the case of England, in order to reduce the burden placed on schools by the inspection process the length and intensity of school inspection has been differentiated mainly on the basis of a school's performance and previous inspection judgements. The period of inspection for schools judged to be effective is shorter than that for ineffective schools (Ofsted, NR2003–4, 2003).

Finally, the position and location of the inspectorate within the overall education system also varies across countries. As noted previously, one important factor in this respect is the shift of authority and decision making power from central to local government, a trend that enables schools to have more power in making decisions. For example, in Hungary, local authorities have replaced traditional centralized school inspectorates since 1986 and now control all administrative matters (Dobart, 2001). In contrast, Inspectorates at the national level linked to the central administration are clearly

apparent in Austria, the Czech Republic, Denmark, Hessen, Ireland, Northern Ireland, North Rhine-Westphalia, Portugal, and Scotland (Standaert, 2000).

In spite of the differences in inspection systems observed across countries, there is also clearly a number of important common aims or features or both. Inspection as a form of evaluation plays a powerful role in maintaining and striving to improve the quality of education, in most if not all countries. Nevertheless, it is not surprising that it is difficult to find a common definition of inspection criteria or to reach an agreement on a particular school inspection model across different countries, given the differences in national traditions, culture and aspirations—among other factors that may influence a country's educational goals. Interestingly, in the context of European Union, there is now more than ever before an impetus for common educational goals, which one may expect to result in greater similarities between European inspection systems in the future. However, the most important current concern is that via the co-operation, discussion and analysis of information by different countries, each country may develop its own quality assured school inspection system based on its own political, social, cultural and educational context (Osler, 2001). Also important is the way policy makers and stakeholders in different countries find to address the challenge of how to combine school inspection with the ideal that all schools are good enough to provide all children and students in the society with an excellent education (Kervezee, 2001). With these two final points in mind it is notable that Standing International Conference of Inspectorates (SICI) has been facilitating the cooperation and discussion between countries to enhance the understanding of education and inspection.

The key issues and features of different inspection systems have been described briefly above, to illustrate these a case study of school inspection in one country England—is described and discussed in detail in the next section.

14.3 English Case Study of School Inspection

The Background of the UK Inspection System

Historically, Her Majesty's Inspectors (HMI) were appointed to inspect publicly funded schools in England in 1839 (Ofsted, NR188C, 2002). One and half centuries later, in 1992 the UK government formally set up a new inspection department—the Office for Standards in Education (Ofsted[1])—to inspect all schools regularly in order to raise standards of achievement and improve the quality of education (HMSO, 1992). Not only should Ofsted respond to the management of the inspection system but it is also obliged to ensure the high quality of the inspection process. Consequently, Ofsted arranges scheduled independent inspection of schools through inviting contractors who have qualified the Quality Assurance Standard to tender for inspection services. Inspection contractors are appointed on the basis of value for money in terms of quality as well as price, and their previous performance wherever possible (Ofsted, 2003).

[1] On 1st September 1992, Ofsted a non-ministerial government department independent from the Department for Education & Skills officially the Office of Her Majesty's Chief Inspector of Schools in England, was established to administer the new inspection system (Ofsted, NR188C, 2002).

Ofsted publishes the school inspection Framework and its Handbooks to help both inspected schools and their inspection teams to understand the inspection process and its work. Meanwhile, Ofsted continually reviews and revises the Framework and its Handbooks[2] to improve school quality via on-going improvements to the inspection process. In order to enhance the quality and effectiveness of inspectors work, training courses are organized, which lead to a formal assessment by HMI. Also training information is published in Ofsted's regular publication—Update—with the purpose of keeping inspectors, inspection

providers, LEA and others well informed about up-to-date policy developments, practices and other related matters within Ofsted.

The first inspection of secondary schools under the new system took place in September 1993, followed one year later by the inspections of primary and nursery and special schools (Ofsted, 1999c). Under the governance of the School Inspections Act 1996[3], schools have to be inspected at least once on a six-year cycle (Ofsted website, Jan., 2003). As a result, all schools had been inspected at least once by July 1998 (Ofsted, 1999c) and Ofsted is on the way to complete the second full inspection cycle by 2004 (Ofsted, NR188B, 2002). To date, the role of Ofsted to inspect and to monitor educational standards has also broadened to include a wide range of educational settings in addition to primary, secondary and special schools. For example, initial teacher training courses; nursery education settings; local education authorities; education and training for 16–19 year-olds; further education and sixth-form colleges; and prison/youth offender institutions are also now inspected by Ofsted.

The Process of the School Inspection in England: Roles and Tasks

The purpose of inspection is to review four themes originally defined under Section 10 of the School Inspections Act 1996 (Ofsted, 1999c):

- the educational standards achieved in the school;
- the quality of the education provided by the school;
- whether the financial resources made available to the school are managed efficiently; and
- the spiritual, moral, social and cultural development of pupils at the school.

Ofsted has developed these themes into an Evaluation Schedule (see Table 14.1), which contains the guidelines to assist inspectors in conducting school inspections.

[2] The most recent inspection framework, which originally was developed and published in 1992, and has been revised several times since then, is published on 31st January 2003. It is expected that the new Handbooks effective from September 2003 for primary and secondary schools will be published by late May 2003 (Ofsted website, Dec., 2002).

[3] The School Inspections Act 1996 mentioned in this chapter refers to the School Inspections Act 1996 which is amended by subsequent legislation, the School Standards and Framework Act 1998 and the recent Education Act 2002.

. The Evaluation Schedule covers the whole range of inspection work and in spite of being artificially compartmentalized, should be treated as a unified map (Ofsted, 2003). Accordingly, a good inspection should provide an independent, external judgement of the school in terms of the quality criteria stated under the four broad themes listed in Table 14.1The process of school inspection can be divided into three phases: prior to inspection, during the inspection and post inspection. A variety of data collection proformas, questionnaires and reports are filled in or produced by schools, parents, pupils or inspectors as part of the inspection process.

Normally, schools are informed about their inspection six to ten weeks before it happens. They are also informed about the type of inspection in terms of whether it will be a short inspection (for the most effective schools) or a full inspection (for all other schools).

Table 14.1 Ofsted Evaluation Schedule.

The Effectiveness Of The School
1. How successful is the school?
2. What should the school improve?
The Standards Achieved By Pupils
3.1 How high are standards achieved in the areas of learning, subjects and courses of the curriculum?
3.2 How well are pupils' attitudes, values and other personal qualities developed?
The Quality Of Education Provided By The School
4 How effective are teaching and learning?
5 How well does the curriculum meet pupils' needs?
6 How well are pupils cared for, guided and supported?
7 How well does the school work in partnership with parents, other schools and the community?
The Leaderships And Management Of The School
8 How well is the school led and managed?
9 How good is the quality of education in areas of learning, subjects and courses?
10 What is the quality of other specified features?

The latter depends on Ofsted's decision based on a combination of the following four factors:

- a favorable quality of education reported in the previous inspection;
- a tendency of improvement in test/GCSE performance;
- relative standards achieved in test/public examination compared to all schools/similar schools; and
- good overall performance in relation to national averages (Ofsted, 1999c).

In other words less effective schools are inspected more frequently than more effective ones. Differentiation may also mean focusing more on some year groups than others, on

particular groups of pupils or on particular aspects of the school (Ofsted, 2003). The major principles of all inspection activities are to contribute to school improvement, to promote inclusion, to conduct the inspection process openly with those being inspected, and to ensure valid, reliable and consistent finding are reported (Ofsted, 2003).

In the following sections the methodology of school inspection is detailed in terms of the data and evidence that is collected and the procedures followed.

Procedures Prior to Inspection

Initially schools are required to provide key information to Ofsted within one week after receiving inspection notification (Ofsted, 1999a, 1999b). The information is collected via a form entitled 'consultation about the inspection and information about the school' (Form S1, for further details see Appendix A). Subsequently, Ofsted sends the inspection contractor the completed Form S1 in addition to the school's Performance and Assessment report (PANDA) and the previous inspection report(s). The inspection contractor then sets up an inspection team that consists of the registered inspector, team inspectors and lay inspectors (Ofsted, 1999b).

The role of the registered inspector is to choose and develop the inspection team, lead the inspection process and to provide the inspection report to Ofsted. Team inspectors role is to inspect particular aspects of a school's work, such as National Curriculum (NC) subjects and contribute their findings to the report. Lay inspectors, who have no significant personal school management experience (except as a governor or acting in any other voluntary capacity), take a wide-ranging view of the school from the perspective of users' levels of satisfaction with the school. Each inspection team has to include at least one lay inspector (Ofsted, 2003).

Once the inspection team is established, the inspection contractor must inform the school about the members of the team and arranges the inspection date with the school. From this point onwards, the registered inspector liases with the headteacher regarding all inspection matters. This includes establishing the head and school staffs' and views regarding the forthcoming inspection; discussing and agreeing dates for visiting the school, meeting up with parents, pupils and other staff before the inspection; discussing the arrangements for analyzing samples' of pupils' work and for providing feedback to staff; and introducing himself/herself and members of the team with their CVs. In the meantime, the school is required to provide additional data and information to the registered inspector before the initial visit to the school (Ofsted, 1999b, 1999c, 2003). This information is collected via various proformas (Form S2—information about school; Form S3—school self-audit; and Form S4—self-evaluation report) along with other related and specified documents (e.g. previous inspection reports, the current school development/management plan, school prospectus, most recent LEA monitoring report, the school's timetable, and a plan of the school). For further details about the information and data collected for the purpose of the inspection see Appendix A. The inspected school is also required to inform parents about the forthcoming Ofsted inspection and send out a parents' questionnaire. After the initial visit and before the formal inspection, the registered inspector produces a pre-inspection commentary on the school based on the full range of pre-inspection evidence. There are two main purposes for the registered inspector to complete the pre-inspection commentary. Firstly, it allows all members of

the inspection to have a clear picture of the inspected school's characteristics before the inspection starts. Secondly, the inspection team and headteacher can share early interpretations of the pre-inspection evidence. The pre-inspection commentary includes hypotheses related to all significant strengths and apparent weaknesses of the school based on analysis of evidence, particularly the performance data in the PANDA report and Forms S1 to S4. In particular, the self-evaluation exercise undertaken by the school (Form S4) is used to focus inspection effort where it matters most and to respond to any specific issues that the inspection can usefully include. The school's summary of its self-evaluation is used as the basis for discussion between the registered inspector and the headteacher and, where possible, governors of the school, when the inspection is being planned. Evidence of how effectively schools undertake self-evaluation and the use they make of it helps inspectors to evaluate the quality of management in the school and the capacity of the school to improve. The headteacher's statement gives the headteacher an opportunity to draw the attention of the inspection team to the specific context of the school and aspects of pupils' progress since the last inspection, particularly details of the school's monitoring of its own performance and progress (Ofsted, 1999c). In addition the registered inspector discusses the accuracy and interpretations of the data included in the pre-inspection commentary with the headteacher and the chair of governors and also briefs the members of the inspection team in the light of these discussions (Ofsted, 2003).

Procedures During the Inspection

The inspection activities during the inspection cover a wide range of data collection approaches and techniques including: observing lessons and extra-curricular activities, sampling pupils' work[4], talking with pupils, analyzing pupils' work, analyzing records of pupils with special education needs, analyzing documents provided by the school, discussion with staff including appropriate local authority staff, discussion of the findings with teachers and other stakeholders, tracking school processes, and joining and observing meetings (e.g. school council or management meetings) (Ofsted, 1999b, 2003). Overall Inspectors are required to ensure that sufficient first hand evidence is collected about the requirements listed in the Evaluation Schedule (see Table 14.1) and to record them accurately on 'Evidence Forms' using pre-specified evidence form codes. High standards are expected from the way the inspection is conducted to facilitate strong professional relationships and respect for inspectors' work. The aim is that teachers and those with leadership and management responsibilities in the school receive well-informed and helpful feedback. Evidence forms and inspectors' records and any briefings, plans or instructions prepared by the registered inspector, contribute to the evidence base for the inspection. The registered inspector is responsible for compiling and assuring the quality of the evidence base (Ofsted, 2003).

[4] It is required that the inspection team should allocate at least 60 percent of inspection time for observing lessons and sampling pupils' work while the inspected school is in session (Ofsted, 1999b).

Procedures After Inspection

There are three main tasks involved in this stage. The first is for the inspection team to reach the corporate final judgement about the quality of education in the inspected school and the reasons underlying this judgement. The registered inspector then completes the 'Record of Corporate Judgements' on the basis of the discussion between team members (Ofsted, 1999b). Secondly, after the inspection team has reached its conclusion, the registered inspector holds a meeting to orally present interim feedback to the head teacher and members of the senior management team (Ofsted, 2003). Additionally, a separate confidential meeting is also held by the registered inspector to give a debriefing to the governing body where the head teacher is also present. The third task is the preparation of inspection report and

summary report. It is the registered inspector's duty to follow the structure of Evaluation Schedule in completing the inspection report, which is unique to the school (Ofsted, 1999b). Thus the inspection report is written to a prescribed format and includes a summary of the school's effectiveness, its strengths and weaknesses, what it must do to improve, and the parents' and pupils' views of the school; reports on each curriculum area inspected, together with more detailed evaluations of subjects and courses as relevant for the type of inspection, and evaluations stemming from the inspection of any issues specified by HMCI. Also, the registered inspector should produce a summary report for parents to understand 'how the school is doing' and 'what the school should do to improve further' (Ofsted, 1999b).

The school is given a copy of the final draft of the inspection report to ensure that judgements made about the school are appropriate and fair before Ofsted publishes the report on its website. Subsequently, the school is required to propose an action plan to indicate how the recommendations suggested in the inspection report will be implemented and arrangements for a follow-up inspection are made where necessary (for example if the school is judged to require 'special measures').

Outcomes of the inspection process

The published inspection report and summary inform parents, the school and the wider community about the quality of education at the school and whether pupils achieve as much as they can. The inspection team's findings provide a measure of accountability and aim to help the school to manage improvement. The inspection process aims to help the school by providing an overall judgement on the effectiveness of the school, and identifying its main strengths and weaknesses and the most important points for improvement. (Ofsted, 2003).

Other outcomes of the inspection process include the HMCI's Annual report to parliament on the quality and standards of education in England, which is based on all the inspections conducted in the previous academic year, including thematic inspection exercises conducted by HMI and additional inspectors. Ofsted holds data from all inspections electronically and in addition to contributing to the HMCI Annual Report the data is analyzed to provide the basis for surveys by HMI and to contribute to the advice they provide to ministers and the education system. Importantly after each inspection, the

school is invited to evaluate the quality of the inspection and report and in the case of dissatisfaction a formal complaints procedure can be followed.

In summary this section described and illustrated the processes of inspection in the context of one education system. The chapter was concluded with a critique of inspection systems as a form of monitoring and as a means of improving educational standards, particularly in relation to Ofsted.

14.4 Conclusion

Critics of external evaluation systems have noted that the impact of inspection systems can distract schools and teachers away the major task of pupil learning in schools. For example, Standaert (2000) pointed out that the impact of an inspection on schools can span the whole continuum of the inspection process lasting two or more years. Indeed, in a UK study carried out by Ouston and Davies (1998) over a three-year period from 1994 to 1997, it was found that the impact on schools increased during the preparation stage and the period when the inspection took place, then decreased after about a year to 18 months. In this study, the impact varied according to the school staffs' attitude towards inspection, the perception of possible 'failure', the perceived quality of inspection and the feedback from the inspectors. In conclusion the study generally reports a positive influence on many secondary schools from Ofsted inspection. However, the researchers also noted that questions remain about whether there could be other, more effective and less costly, ways of helping schools to improve their practice and outcomes.

The question of whether the Ofsted model of inspection is appropriate for all types of schools was also addressed in a recent study by Chapman (2002). The research focused on schools staffs' perception of Ofsted's contribution to school improvement in schools identified by the Department for Education and Skills (DfES) as 'facing challenging circumstances'. The preliminary findings suggest that heads and other senior leaders adopt a more autocratic approach to leadership than they would prefer during the preparation for an inspection. This is due to the central aim of schools facing challenging circumstances being to avoid being judged as requiring 'Special Measures' or worse as an outcome of the inspection process. In contrast, when schools in challenging circumstances where exposed to HMI monitoring visits, teacher relationship with these inspectors appeared to be more positive than their Ofsted counterparts. This may be due to the fact that HMIs in a position to understand the complexities of context more readily, due to multiple site visits to one school over a period of time. Arguably, if relationships are more positive and there is a greater understanding of contexts, then the likelihood of teachers changing their practice is higher, and therefore the possibilities for improvement greater. Teachers in this study perceive high stress levels, workload, and lack of job satisfaction as important factors associated with Ofsted inspection.

Interestingly, the findings indicate that while changes in practice appear limited in the context of inspection of schools in challenging circumstances, the nature of these changes follow particular patterns. For example, it appears that Ofsted is a more effective tool for changing management or non-classroom practices than classroom practice in schools facing challenging contexts. Also the changes made to practice appear to be changes that could be generated without the expense and pressure of an Ofsted inspection. Thus,

Ofsted as a lever for change at the classroom level appears to be limited. The researchers conclude that a more productive and sustainable model for generating classroom improvement needs to be developed and that variation in the quality and quantity of feedback received must be minimized in order to harness Ofsted's potential for improvement at the classroom level.

In conclusion there seems to be evidence that inspection systems can have a positive impact on school improvement but not necessarily for all schools, particularly those in challenging or disadvantaged circumstances. Moreover, in terms of improving classroom practice, Ofsted inspection in its present form (and presumably also other similar inspection systems in other countries), may have only a marginal capacity to improve schools. Further research is needed relating to the number and quality of innovations being developed within classrooms, and whether the existing nature of monitoring and inspection can support successfully the experimentation and artistry necessary to engage pupils in meaningful learning.

Overall current research findings suggest that the appropriate balance between the roles of internal and external evaluations are crucial as well as the balance of the application of pressure and support. As a final point, Chapman (2002) argues that the emerging findings from his study suggest that future frameworks or inspection systems must consider:

1. context specificity—the inspection process must be flexible enough to support improvement in schools at different stages of development, exhibiting diverse cultural typologies, structures and perhaps most importantly differential capacities for change.
2. change at all levels. The inspection process must identify meaningful areas for change at all levels within schools. Appropriate levers must then be used to facilitate the changes with the aid of specialized local knowledge.
3. post inspection relationships. In order to generate sustainable improvements the inspection process must provide post-inspection support to facilitate the change process.

Appendix A— Information and Data Collected to Inform the Inspection Process

A variety of information is collected prior to the inspection including basic information about the school, pre-entered where possible (Form S1). Form S2, which includes more detailed information about the school and its pupils. Form S3, which is completed by the governing body and includes its assessment of how far statutory arrangements and policies are in place. Form S4, which provides the school with an opportunity to summarize its own perceptions of its quality and standards, gained through monitoring and self-evaluation (Ofsted, 2003). Specific details of the data collected on Forms S1-S4 are detailed below.

Pupils including the number of pupils categorized by year group, gender and according to different background factors such as ethnicity, refugee status, English as an additional language, entitlement to free school meals (FSM—a measure of low family income), and special educational needs (SEN)). The section on organization and staffing **(1)**

Consultation about the inspection and information about the school proforma (Form S1) provides four kinds of information[5] (Ofsted, 2000a). The Section A covers basic information about the School including data such as type of school, age range of pupils, gender of pupils, and contact details. Section B covers information about (section E) is related to the number of teachers, support teachers and unqualified teachers. The section on 'Further Information to Help Set Up the Inspection' (section G) indicates figures regarding the school's site, classes taught for each subject (including GNVQ and other courses taught) by year group. Whether the school is currently subject to reorganization proposals and factors which the appropriate local authority wishes the inspection team to take into account and significant changes which will take place before the inspection are also stated in Form S1.

(2) Information about the school proforma (Form S2) consists of six sections. Section A (Information that Identifies Your School with the Data You Enter) contains any change in the information provided in Section A of Form S1. Section B (Information about Pupils) is information additional to Section B in Form S1 including numbers related to fixed period/permanent exclusions on each gender by ethnicity, pupil mobility, admission, attendance, routes taken by pupils at age 15+. Section C (Standards of attainment) concerns the results of National Curriculum (NC) assessments, pupils' attainment on entry for each core subject by level, and percentages of statutory targets achieved. Section D (the Curriculum) covers information regarding hours of teaching time at each key stage, intended percentage of total teaching time on each subject by year group, curriculum description (e.g. number of teaching groups, size of each group, methods of allocating pupils, and an indication of where and how much

support teaching is provided), numbers of boys/girls studying foreign languages, number of pupils for whom the NC is disapplied, and withdraw from religious education and from collective worship. Furthermore, section E (Organization and Staffing) includes extra information in addition to section E in Form S1. This covers, for example, teaching staff lists and their details; indications about periods taught in each year group, teacher mobility during the last two school years, temporary teachers, ITT students, educational support staffs by SEN by total hours/week, support staffs for minority ethnic/traveller pupils, administrative staffs, premise staffs, average group size by year by key stage; and pupil:teacher ratio. Section F (Finance) states school's financial information including income and expenditure in support of pupils with SEN, expenditure on information technology and on books, pupil:computer and pupil: book ratios.

(3) School self-audit Proforma (Form S3) completed by the governing body provides indications about whether the school's statutory requirements in terms of the Evaluation Schedule (see Table 14.1), are fully in place, partly in place, not in place or do not apply with relevant explanations. There are a total of at least 30 items related to statutory requirements and 10 items related to other areas(Ofsted, 2000c).

[5] The details in Form S1 and Form S2 are slightly different subject to the use for primary/nursery, secondary and special schools (Ofsted 2000b).

(4) Self-evaluation report (Form S4) is a school self-evaluation report and includes a statement from the headteacher. In addition the school has an opportunity to state its distinctive features. The school is asked to evaluate rather than describe on a variety of items regarding the requirements listed in the Evaluation Schedule (Table 14.1) in terms of impact and outcomes for pupils against seven grades (excellent to very poor). On each item, the school can also report the actions being taken to improve each particular aspect (Ofsted, 2002).

In addition to Forms S1-S4, a school's **PANDA** report prepared by Ofsted is made available to the school and the inspection team (Ofsted/SEU, 2001). It is worth noting that since Autumn 2001 the PANDA replaces the previous Pre-Inspection Context and Statistical Indicators (PICSI) used for the purpose of informing the inspection process. There are four main types of information included in the 2001 format of PANDA report (Ofsted, 2002). The first section contains basic characteristics of the school in relation to the numbers on roll and percentages regarding pupils entitled FSM, speaking English as an additional language, SEN status and ethnicity—at both the school and national levels based on the Pupil Level Annual School Census (PLASC). In the second section, a one-page summary (Inspection Judgements) of the last inspection judgement about the school in four broad aspects[6] against a four-point scale[7] is provided. Also the percentages of schools rated at each point of the scale in terms of each of the four aspects presented nationally and separately for schools with similar contexts. Further, a one-page of Attainment Summary provides the school's attainment grade based on a seven-grade scale[8] in comparison with that of all schools nationally and of schools in similar contexts. Furthermore, a section named 'Attainment Statistics for the Inspection Report' provide the numbers and percentages of pupils in the school reaching each statutory required NC level on each of core subjects. The latter is also provided at national level. Section three (Attainment Section) provides the details of the school's attainment in comparison with national averages/national benchmarks and value added scores. The latter is a measure of the progress pupils make between two key stages and was published in each secondary school's PANDA report for the first time in 2002 (Ofsted, 2002). The last section is Additional Information concerning pupils' attendance and school's context.

[6]The four aspects are standards achieved by pupils, quality of education, the school's climate and management and efficiency.

[7] very good, good, some improvement required, and substantial improvement required).

[8] The seven-grade scale is defined as A* (very high), A (well above), B (above), C (broadly in line), D (below), E (well below) and E* (very low) which is not compatible with judgements made by inspectors (Ofsted, panda, 2002).

In addition to the pre-inspection data collection and the PANDA report, further evidence relevant to the inspection process includes the school's previous inspection report(s) including the judgements with specific details about how well the school achieved and why in relation to the requirements listed in the Evaluation Schedule. Information collected at parental meetings and via the standard Ofsted parental questionnaire is designed to find out parents' view about the inspected school regarding the content of the Evaluation Schedule (see Table 14.1), their expectation of the school and any other issues that parents may wish to raise. Similarly, information is gathered at the meeting with school staff and pupil representatives in relation to the Evaluation schedule. Finally each school's development or management plan[9], prospectus and any other relevant policy document concerned with where the school is now and where the school intends to be in the future feeds into the inspection process

[9] A good development plan will link with the school's assessment policy in using the results of pupil assessment to target improvement in pupils' achievement. The assessment policy should describe how this achievement is to be monitored and evaluated..

15
School Evaluation: Basic Concepts

15.1 Introduction

In this chapter school evaluation will be defined, on the basis of an analysis of the evaluation concept and the structural context of the school within educational systems. Since evaluation is closely related to the issue of quality in education and school evaluation approaches are close to all kinds of systems for "quality care" and quality control within organizations, some attention will also be paid to the concept of "quality" of schooling. The chapter sums up to a taxonomy of school evaluation types, in which various methods of school evaluation are related to the specific information providers and the specific audiences of each.

15.2 Definitions

15.2.1 Evaluation

As stated before, evaluating means judging the value of an object, and evaluation in the sense of a particular type of disciplined inquiry emphasizes that this "judging" and "valuing" is based on some kind of systematic information gathering approach.

In the case where this systematic information gathering is formalized according to the criteria for social scientific inquiry the term *evaluation research* is appropriate. A third major component of evaluation, next to the valuing aspect and the systematic approach to information gathering, is the applied context: evaluation results are expected to be used by relevant audiences. Again there is a prototype situation, often related to policy-evaluation, where evaluation results are expected to shape, or at least have a certain impact on, policy *decisions*.

In the evaluation literature authors vary in their emphasis of each of these three basic components: valuing, systematic inquiry and use for decision making. In all types of definitions where goal attainment is placed central, the value aspect is prominent (since whether or not program goals are attained provides the basis for judging it as either successful or unsuccessful). Thus Tyler defines evaluation as "*The process of determining to what extent educational objectives are actually being realized'* (Tyler, 1950, p. 69, cited by Nevo, 1995, p. 10).

Also Provus' "Discrepancy Evaluation Model" (Provus, 1971) depends heavily on pre-established goals which serve as a basis for judging the success of a program. Scriven's reaction, namely his idea of "Goal Free Evaluation" (Scriven, 1967), also

emphasizes the valuing aspect, although he denounces program goals as providing the basic orientation for making judgements. Instead of goals and objectives the demands and needs of clients or relevant audiences of the program that is to be evaluated are seen as the basis for choosing evaluation standards (i.e. the norms used to determine "success" or "failure" of a program).

In making other definitions in the literature both elements of "valuing" and "systematic inquiry" are present, like for example in the definition presented by the "Joint Committee on Standards for Evaluation", led by Daniel Stufflebeam: *"evaluation is the systematic investigation of the worth or merit of some object"* (*Joint* Committee, 1981, p. 12, cited by Nevo, 1995, p. 10).

Finally there is a category of authors who seem to altogether leave out the judgmental component from their definitions of evaluation and define evaluation in terms of providing information for decision making.

Stufflebeam's earlier CIPP-model (Madaus & Stufflebeam, 1983) is an example of this as are authors who speak of "utilization focused evaluation" (Alkin et al., 1979; Patton, 1978). It could be argued that in these approaches the judgmental component is merely left implicit, since valuing is always there whenever information is interpreted as favoring or disfavoring a particular decision alternative.

Apart from these mainstream distinctions in defining (educational) evaluation, there are examples in the literature where still other aspects of the "evaluation endeavor" are placed central. Cronbach and his associates, for example, depict the evaluator as an "educator", who enters a dialogue with the professionals in the object situation of the evaluation. Their view is also seen as an example of suppressing the role of the evaluator as a "judge" (Nevo, 1995, p. 10). This view where the component of descriptive information gathering is placed central upon which "illumination" or "education" of the evaluees and clients is supposed to follow is also predominant in perspectives on evaluation where qualitative description and naturalistic methods are propagated (i.e. Stake, 1975; Guba & Lincoln, 1982). In what is called "stakeholder-based evaluation" the fact that it is often the case that different parties have an (often divergent) interest in program outcomes, is used in shaping the evaluation. The idea is that giving these various parties more proprietary feeling for the evaluation process and its outcomes will increase the chances of the evaluation results being used (cf. Scheerens, 1990, p. 38).

In advocacy oriented or "judicial" evaluation varying value positions of relevant parties are also used, but more in the final stage of interpreting, and weighing and judging the information that has been gathered (Wolf, 1990, pp. 79–81). During a public presentation of the data a hearing is organized according to the format of the functioning of a court of law. Witnesses are called to provide evidence before or against the case (i.e. the success or failure of a program) and juries decide.

In summary, it seems wise to contain all three elements: systematic inquiry, judgement, and use in decision-making settings in our definition of educational evaluation. Therefore our working definition of educational evaluation is: *Judging the value of educational objects on the basis of systematic information gathering in order to support decision making and learning.*

From the brief overview of views on the evaluation phenomena in the relevant literature it has also become clear that there are some important "contextual conditions" at stake when we deal with educational evaluation. The most important dimension on

which these conditions manifest themselves is the variation in positions and interests in the evaluation process and outcomes of relevant parties. This realization gives cause to paying considerable attention to the political and organizational contexts, throughout this book.

15.2.2 School evaluation

In our definition of educational evaluation in the preceding paragraph we spoke of "educational objects". When "schools" are the educational objects to be evaluated—instead of programs in which many schools take part, teachers or individual students—one can speak of "school evaluation".

The fact that schools are the objects which—on the basis of systematic information gathering—are being judged, leaves open the possibility that data on "objects" or "units" within the school are the focus of data collection. However, information on these within-schools units (classrooms, teachers, departments or pupils) will then be aggregated to the school level in order to allow for judgements on the individual school. As we shall see further on, such judgement often requires information on other schools, as a basis for comparison.

15.2.3 Internal and external school evaluation

There are four main categories of actors in all types of evaluation, including school evaluation:

A the contractors, funders and initiators of the evaluation;

B the (professional) staff that carry out the evaluation;

C the persons in the object-situation which provide data;

D the clients or users or audiences of the evaluation results.

Mostly categories A and D will partly overlap, in the sense that contractors will almost always be "users" as well, although they may not be the only category of users. For example, a particular department at the Ministry of Education may be contractor and user of a particular program evaluation, although other important parties, such as Members of Parliament and the tax-payers may also be considered as relevant audiences.

> *If all of these audiences are situated within the organizational unit which is the object of evaluation we speak of internal evaluation. Even if a special unit or team is composed for the evaluation within the organizational unit, but which is not part of the "production/service "part" of the project (Nevo, 1995, p. 48), the classification of "internal" evaluation would still apply.*

Next, a distinction can be made between two types of external evaluations:

a. when contractors, evaluators and clients are external to the unit that is being evaluated;

b. when the unit that is evaluated initiates and contracts the evaluation to external evaluators and users may be either exclusively internal of both internal *and* external to the evaluation object.

Note that the distinction between internal evaluation with a specialized internal evaluation unit and external evaluation where the unit (school) initiates the evaluation is solely dependent on the institutional setting of the evaluator.

15.2.4 School self-evaluation

After the preliminaries in the preceding section it is now simple to define *school self-evaluation*, namely as the type of internal school evaluation where the professionals that carry out the program or core-service of the organization (i.e. teachers and head teachers) carry out the evaluation on their own organization (i.e. the school).

This definition would also apply if school teams would make use of external advisors to provide them with advice on evaluation methods etc., because the school teams would still take the responsibility for carrying out the evaluation.

The definition of school self-evaluation is analogue to the following definition of "self-report: "Self-report refers to the result produced by any measurement technique in which an individual is instructed to serve both as assessor or observer and as the object of the assessment or observation".

15.3 Evaluation of School Quality

15.3.1 What is quality?

School evaluation activities have the function of monitoring quality. Having said this, one is faced with the task of clarifying what is meant by quality in the sense of "the quality of schools" and "the quality of education". First, within the context of school evaluation, the interest in quality refers to the school as a whole and not just to the quality of certain aspects or elements like: teaching methods, teachers or school management.

Next, when it is established that school evaluation ideally should look at the sum of all aspects and elements of school functioning, one is faced with the need to make certain selective choices and set priorities, if only for practical purposes. But, in order to make these choices, one needs frameworks and/or analytical schemes to capture the "whole" of school functioning. Two of these conceptual frameworks will be used to elucidate choices with respect to quality: (i) a basic model from systems theory and (ii) perspectives on organizational effectiveness.

15.3.2 A basic model from systems theory

An abstract way to picture the functioning of an organization is the construct of the organization as a black box into which inputs flow and by which outputs are "somehow" produced (see Fig. 15.1). This model is in fact a more basic description of the systems model that was introduced in Chapter 2.

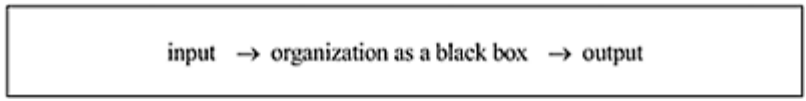


Figure 15.1 The organization as a black box.

Even this rather simple model can be used to make the construct of "quality" more concrete. The economic construct of efficiency is defined as the ratio of outputs to inputs, where output (in the case of schools) can be defined in terms of attainment level averaged over pupils, whereas, from the economic perspective, inputs can best be thought of in terms of financial and material resources. Economic efficiency, as a particular operationalization of organizational quality, is focused on the highest possible level of outputs for the lowest cost level of inputs.

In Figure 15.1 it is assumed that within the black box *processes* take place that transform inputs into outputs. When it is attempted to further describe these processes in terms of which process characteristics are most effective in obtaining desired levels of outputs, the model of Figure 15.1 becomes more elaborate. In addition, a further distinction of the "input" category is usually made by separating direct inputs into the system and characteristics of the larger *context* from which these inputs originate. In this way a Context-Input-Process-Output-model is obtained. This model is often used as a conceptual framework to summarize the results of school effectiveness research. Compare the figure based on Scheerens, 1990, that was presented in Chapter 11.

The notion of quality inherent in integrated school effectiveness models like the one in Figure 11.2 is that:

a. outputs are the basic criteria to judge educational quality;
b. in order to be able to properly evaluate output, achievement or attainment measures should be adjusted for prior achievement and other pupil intake characteristics; in this way the value added by schooling can be assessed;
c. in selecting variables and indicators to assess processes and context one should look for those factors that have been shown to be correlated with relatively high "addedvalue" factors.

As was also made clear in Chapter 10 it should be noted that school effectiveness models do not prescribe the types of outputs that should be used to assess quality. In principle all types of outputs, cognitive or non-cognitive could be inserted. In the actual practice of school effectiveness research cognitive outcomes, mostly in terms of achievement in core-subjects like reading, arithmetic, and language, have predominated.

To the degree that educational effectiveness models provide an acceptable operational definition of quality, they can also be used as a guideline in the design of instruments for school evaluation. Points *a* (focus on outcomes), *b* (proper adjustment of outcomes) and *c* (measure process characteristics associated with high added value) mentioned in the above can be read as as many guidelines to make choices with respect to instrumentation. The reasoning here is exactly similar to that in Part 4 of this book, where the school effectiveness research literature was used as a basis for the selection of education indicators.

However, a broader perspective on quality can be considered. Such a broader perspective can be obtained from multiple orientations towards organizational effectiveness that will be discussed in the next section.

15.3.3 Multiple criteria to assess organizational effectiveness

In organizational theory models like the school effectiveness model are seen as belonging to just one of several effectiveness perspectives. The effectiveness perspective in which the school effectiveness model fits is referred to as the *rational goal model*, where productivity and efficiency are the central criteria.

Alternative models are: the *open systems model*, with growth and resource acquisition as effectiveness criteria; the *human relations model* with human resource development as a central criterion and the *internal process model* in which stability and control are the main issues. Quinn and Rohrbaugh (1983) depict these four models as determined by two dimensions; one that has flexibility and control as its extremes and one that represents an internal versus an external orientation (see below).

From this framework additional process indicators of school functioning may be generated.

As far as the rational goal model is concerned it should be noted that this model does not specify *which* educational objectives are relevant. Next to knowledge and skills in basic school subjects other educational aims may be recognized. Two important additional categories of educational objectives are *social, emotional and moral development* on the one hand and the development of *general cognitive skills* on the other. For our purposes these categories of educational aims (next to the basic cognitive skills that have been the focus in empirical school effectiveness research) are

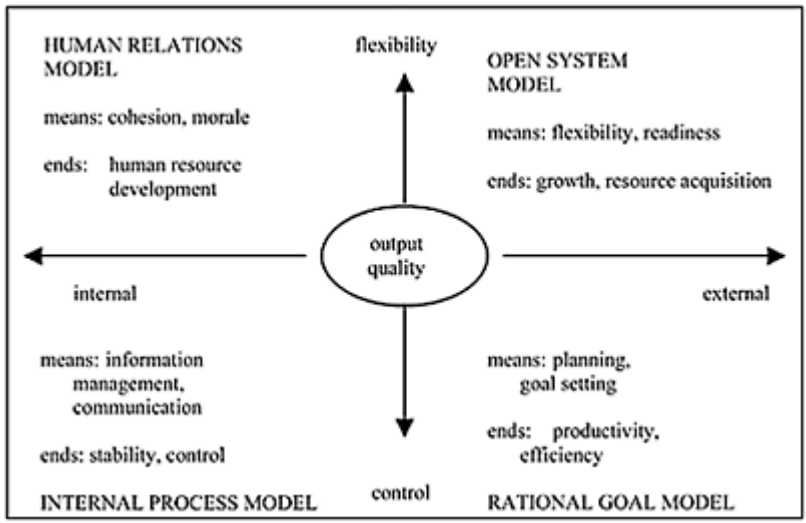


Figure 15.2 Typology of effectiveness models. Source: Quinn & Rohrbaugh (1983).

relevant to the degree that they may require somewhat different teaching approaches and different school organizational arrangements than the process variables that have been shown to matter in the traditional school effectiveness models (Scheerens, 1994).

According to Goodlad and Anderson (1987) *multiage* and *interage grouping* have the advantage of fostering social and emotional development apart from being effective in realizing traditional educational goals. The disadvantages of a competitive achievement-oriented atmosphere are supposed to be modified by these organizational arrangements, while the motivational disadvantages of both promoting and nonpromoting as in graded system are prevented. *Non-gradedness* and *team-teaching* are seen as measures to realize differentiated adaptive teaching and an integrated, continuous learning route. Such approaches are thought to contribute to the degree that students are comfortable and happy in the school.

Educational psychologists increasingly emphasize the importance of self-regulated learning and meta-cognition. "Subject-free" cognitive skills can be acquired in programs in which learning how knowledge is acquired ("learning to learn") is taught.

The *human relations model* is strongly concerned with the work satisfaction of teachers. Louis and Smith (1990) identified seven "quality of work life indicators":

- *respect from relevant adults,* such as the administrators in the school and district, parents, and the community at large;
- *participation in decision-making* that augments the teachers' sense of influence or control over their work setting;
- *frequent and stimulating professional interaction* among peers (e.g. collaborative work/collegial relationships) within the school;
- structures and procedures that contribute to a high sense of efficacy (e.g. *mechanisms permitting teachers to obtain frequent and accurate feedback about their performance* and the specific effects of their performance on student learning;
- *opportunity to make full use of existing skills and knowledge,* and to acquire new skills and knowledge (self-development); the opportunity to experiment;
- adequate *resources to carry out the job;* a pleasant, orderly physical working environment;
- a sense of *congruence between personal goals and the school's goals* (low alienation).

Other factors that may contribute to teachers' satisfaction are task differentiation and possibilities to make promotion (though these are usually limited) and financial incentives, though this approach, according to some authors might prove counterproductive—McLaughlin and Mei-ling Yu (1988).

The *open system model* emphasizes the responsiveness of schools with respect to environmental requirements. This means on the one hand that school organizations can create effective buffers against external threats and on the other hand that schools can manipulate their environments to the degree that their own functioning is not only safeguarded but also improved. In some countries (The Netherlands for instance) external regulations for schools are relaxed and school autonomy is enhanced. This state of affairs

offers new possibilities, but also confronts the school with new requirements, for instance to conduct their own financial policy.

Demographic developments (less pupils) may force schools to be active in stimulating student enrolment and "school marketing". Developments in educational technology, initiatives for educational innovations from higher administrative levels as well as accountability requirements can be seen as additional external forces that challenge the school's readiness to change.

In a Dutch study, Gooren (1989) found evidence for a dichotomy of schools that could either cope or not cope with these new external requirements. The schools that could cope more frequently had strong leadership or a collegial structure in contrast to non-coping schools which answered the image of the loosely-coupled, segmented school organization.

Capacities of schools to deal with an increasingly demanding and dynamic environment are described in terms like *"the policy-making potential of school"* and *"the self-renewing capacity of schools"*. School organizational characteristics that are thought to contribute to these capacities are:

- leadership (also in sense of entrepreneurship);
- collegiality;
- capacity for self-evaluation and learning (see for instance Morgan" image of the learning organization—Morgan, 1986, Ch. 4);
- overt school marketing activities;
- strong parental involvement;
- boundary-spanning positions;
- support of external change agents.

Proxy-indicators concerning the success of responsiveness are enrolment figures and characteristics of buildings and equipment.

Whereas the *human relations model* is concerned with social and cultural aspects of "what keeps organizations together", the *internal process model* reflects a preoccupation with formalization and structure. From this perspective the following factors are of interest:

- explicit planning documents (such as school curricula, school development plans);
- clear rules regarding discipline;
- formalization of positions;
- continuity in leadership and staffing;
- integrated curricula (coordination over grades).

Proxy-indicators for the stability of school organizations are attendance rates, the number of teaching periods not given, and figures about the continuity in staffing.

15.3.4 Quality indicators

The ideas for additional process indicators that come from this more comprehensive treatment of organizational effectiveness are summarized in Figure 15.3. (Process indicators induced from the narrower model of school effectiveness research are also included.)

15.4 A Taxonomy of Basic Types of School Evaluation Approaches

In this section basic types of school evaluation will be discussed. By "basic" we mean that these approaches, which are mainly distinguished on the basis of *evaluation methodology*, have a certain tradition and are rooted in specific socialscientific disciplines.

Stating that these approaches are "basic" and perhaps more "classic" than some of the methods to be discussed further on does not mean that their application should be considered "less innovative". Even for these approaches there exists no widespread practice and application, particularly from an international perspective. Indeed, pupil monitoring systems and school management information systems often require very sophisticated instruments and tools and are therefore potentially innovative from a more technical point of view. School-based review approaches are likely to be innovative from a different perspective, namely in the challenge they provide with respect to the social functioning of school teams, the discussion of norms and values and what is sometimes referred to as "organizational learning" by teachers as "reflective practitioners" (Argyris & Schön, 1974).

Human relations model	Open system model
Quality of work life indicators	– entrepreneurship
– respect	– collegiality
– participation in decision-making	– capacity for self-evaluation and learning
– professional interaction	– overt school marketing activities
– performance feedback	– parental involvement
– opportunity to use skills	– boundary-spanning positions
– resources	– external change agents
– congruence personal/organizational goals	– student enrolment figures
	– resources (buildings, equipment)
Internal process model	**Rational goal model**
– planning documents	*(school effectiveness research)*
– disciplinary rules	– educational leadership
– management information systems	– success-oriented ethos
– formalization of positions	– monitoring of student's progress
– continuity in staffing and leadership	– time on task
– integrated curricula	– content-covered (opportunity to learn)
– attendance rates	
– lessons "not given"	*(broader set of educational goals)*
	– non-gradedness

– team teaching

– individualization, differentiation

– continuous learning route

– time spent on social, emotional, creative and
- moral development

– "learning to learn" activities

– diagnostic testing

Figure 15.3 Additional factors for process indicators generated form the Quinn and Rohrbaugh framework.

15.4.1 Basic types of school self-evaluation approaches

Currently, several approaches to school self-evaluation are being used. Each has a specific disciplinary background and a specific context in which the approach was originally employed, as is shown in Table 15.1.

Each of these approaches will be sketched briefly and strong and weak points will be discussed.

Table 15.1 Different Origins of School Self-Evaluation Approaches.

approach	disciplinary background	context
school-based review	social psychology, education	schools
management information systems	business administration, operations research	private industry
educational indicators	economics, educational statistics	macro-level applications
organizational diagnosis	management consultancy	private industry, public-sector organizations
pupil monitoring systems	educational measurement	(remedial) teaching

School-based review

School-based review depends heavily on opinions of school personnel on discrepancies between the actual and an ideal state of affairs in schools. In this way a broad perspective, in which all the main aspects of school functioning can be scrutinized, is possible. Usually, respondents are also asked to indicate whether a certain discrepancy should be actively resolved. This approach to school self-evaluation seeks to gear improvement-oriented action to appraisal. The context of application is usually school improvement,

which means that a school-based review is carried out when there is a prevailing commitment to educational innovation.

Advantages of this approach are: a broad scope, a user-friendly technology, an explicit linkage between evaluation and action, and a high degree of participation (all school personnel take part in the review). A definite weakness of school-based review is its dependence on subjective opinions and its (usual) neglect of "hard" factual data on school functioning, most notably output data.

Examples of procedures for school-based review are the GRID and GILS-systems (see Hopkins, 1987) and the SAS-system (Voogt, 1989).

School management information systems

School management information systems have been inspired by similar systems in private industry. Generally they consist of a careful modeling of information streams and information needs within a company, deciding which data should be available for purpose on a more or less permanent basis, followed by design and implementation of a computer configuration and software. Bluhm and Visscher (1990) describe a management information system as an information system based on one or several computers, consisting of a data-bank and one or several software applications, which enable computer-based data storage, data analysis and data distribution. A question that could be answered by means of such a school management information system would be: "to which degree has absenteeism decreased after the implementation of specific measures to fight absenteeism?"

Management information systems have a great potential for supplying important information on a routine basis. At present practical barriers: one needs to have sufficient and adequate computer hardware and even when professionally developed software packages become available, quite a few specific maintenance functions must be carried out, while new routines and perhaps even functions to guarantee adequate data-entry should be developed.

Educational indicators

Although educational indicator systems are usually employed at the macro level (the level of national educational systems), for instance to describe the "state of education" of a country on a yearly basis, some authors have suggested applications at the school level (Taeuber, 1987; Oakes, 1987; Scheerens, 1990). When applied at the school level, educational indicator systems typically will include "process" or "throughput" information, next to input, school-context and output data.

Results of school effectiveness research studies are usually employed to select process indicators. The general idea of indicators is to provide an at-a-glance profile of certain important characteristics of an educational system.

This means that there is no aspiration to "dig deep", while employing easily measured characteristics and so-called proxy measures. This feature is at the same time a definite limitation of the approach. Another "danger" is the use of process or throughput data as evaluation criteria, instead of explanatory conditions of educational outputs. This could easily lead to goal displacement, where the "means" in education are treated as "goals" in

themselves. A technical limitation which might encourage this improper use of process indicators is the fact that the question of relating process and output indicators by means of formal statistical analysis has hardly been tackled for applied purposes. This problem will be addressed in other sections of this article.

Organizational diagnosis

As educational institutes (schools and universities) are made to function more autonomously, it is quite likely that they will become more like private companies in their managerial and organizational characteristics. An example of this would be a stronger emphasis on strategic planning and on scanning the external environment of the school. It is therefore not surprising that approaches used in management consultancy are introduced in schools. Although these approaches, generally labeled as "organizational diagnosis" or "management audit", usually depend on an external organizational consultant—they are also available for school self-diagnosis. In contrast to school-based review these approaches tend to be exclusively based on information provided by the management of the organization. So, when they are used without an external consultant they would appear to be somewhat like "management introspection". A strong point of this approach is that it is likely to pay attention to issues that were kept largely unnoticed by the educational province, such as external contacts, anticipation of developments in the relevant environment, and flexibility in offering new types of services. The most important disadvantage remains, however, that this approach is not so easy to transform to a school-based application, without an external consultant.

Pupil monitoring systems

Pupil monitoring systems operate at the micro level (class level) of educational systems. In the ensuing sections of this article it will be shown how this class of techniques can also be used for self-evaluation at the school level.

Basically pupil monitoring systems are sets of educational achievement tests that are used for purposes of formative didactic evaluation. An important function is to identify those pupils who fall behind and where they experience difficulties.

Pupil monitoring systems have one asset which, in our opinion, is essential for all efforts to make school functioning more effective: the centrality of output data at the level of the individual pupils measured by means of achievement tests. If approaches to school self-evaluation neglect these type of data there is a risk that the information basis they supply for educational or administrative decision-making is faulty (see the earlier reference to the phenomenon of goal displacement).

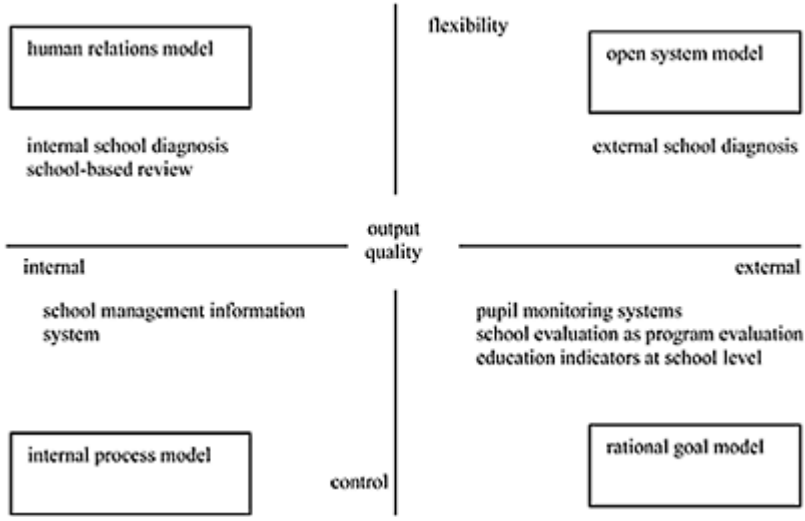


Figure 15.5 Association of organizational effectiveness perspective and basic school selfevaluation approaches.

15.4.2 Basic types of school self-evaluation and perspectives on educational quality

The four types of perspectives on organizational quality, distinguished by Quinn and Rohrbaugh (1983) and presented in section 1.2, can be related to the basic types of school evaluation as distinguished in the preceding section.

The association of organizational quality perspective and self-evaluation approach is based on the similarity in disciplinary orientation and correspondence in the criteria that are likely to be central in each of the evaluation approaches, see Figure 15.5.

15.4.3 A more extensive taxonomy of school evaluation methods

When school evaluation at large—not exclusively school self-evaluation—is considered and when methods are distinguished on the basis of actors and objects of the evaluation a more extensive set of approaches can be distinguished (cf. Van Amelsvoort et al., 1998):

Evaluation methods, when pupils are the object

- informal procedures of evaluating learning tasks, marking [teachers];
- curriculum-tied progress tests for different subjects (i.e. unstandardized tests) [teachers];
- semi-formal presentations of completed learning tasks such as portfolios [teachers];

- authentic assessment, i.e. when pupils' progress is evaluated in natural circumstances [teachers, schools];
- pupil monitoring systems of standardized tests and assignments [schools];
- certifications (not necessarily with diploma) [central government];
- assessment tests initiated by [local, regional or national authorities].

Evaluation methods when teachers are the object

- formal methods of teacher appraisal [school boards, school leaders, inspectors].
- informal methods of teacher appraisal [school boards, school leaders]
- monitoring teachers by means of a form on the quality of instruction [senior school management]

Evaluation methods when the school (or department within a school) is the object

- school diagnosis in the form of so-called "GRIDS" depending on opinions and self-appraisal of school staff [school leader, department];
- school management information systems, e.g. a computerized registration of absenteeism [school management and other administrative levels];
- integrated school self-evaluation systems in which assessment of school processes is combined with assessment of pupils' achievement [school management, head of department];
- so-called "visitation committees", whereby peers (e.g. colleagues from other schools) screen and evaluate a school [unions of schools];
- accreditation, whereby an external private company screens aspects of school functioning using a formal set of standards [private agency];
- inspection, qualitative or semi-qualitative assessment by inspectors of school [Inspectorate];
- school level indicators or key data (school monitoring) [school management and other administrative levels];
- assessment and market research of the school in its relevant environments, e.g. with respect to expectations on future enrolments [external research institute].
- external school review by [private consultancy institutes]

Evaluation methods when the system of schools is the object

- national assessment [national government];
- program evaluation [national government];
- inspection [national government];
- educational indicator projects [national government].

16
Issues and Dilemmas in School Self-Evaluation

16.1 Introduction

In Chapter 5 various models of the school as an organization were discussed. The more traditional perspective of the school, as a 'professional bureaucracy' depicted the school as a relatively horizontal structure, a lot of autonomy of the teachers and a minimal need for hierarchical management. Ideas on school leadership, to some extent inspired by the concept of "educational" or "instructional leadership" from school effectiveness research modified this image, in the sense that the primary process of teaching was now seen as "coordinated" and at least marginally controlled by others, notably the school director. The discussion in Chapter 5 ended by considering the reality and feasibility of the school as a "learning organization", in which the evaluation and feedback function was seen as being of central importance. All these considerations are extremely important as it comes to estimating the possibilities for school self-evaluation. If the more traditional view is still the most valid description of reality, school self-evaluation would be deemed to be a marginal phenomenon in schools.

Experience from actual work in school self-evaluation projects and results from studies conducted in the European Union underline that a lot is possible when school self-evaluation is conducted in integrated projects of Networks of schools. In the final two chapters of this book descriptive case-studies of two of these projects will be presented. At the same time, there are also quite a few experiences that seem to indicate that, without specific external support, school self-evaluation is likely to remain a "Fremdkoerper" (English "alien body") in the school. Schools have difficulty in interpreting quantitative data presentations, are not used to systematic record keeping and have specific problems in making the link between "diagnosis" and "therapy", if it comes to interpreting and applying the results from school selfevaluations. The introduction of school self-evaluations in a school should be interpreted as an educational innovation in its own right. Theoretical perspectives inspired by the theory of autopoietic systems, underlining the importance of "selfreference" in organizations are relevant in the sense that they would predict a "real" incorporation of school self-evaluations as dependent on becoming incorporated in the organizations patterns of self-reference (Scheerens, 2002).

16.2 Interpretation and Use of Results: How Helpful is the School Effectiveness Perspective for this Issue?

When schools are confronted with school self-evaluation activities the setting of evaluation priorities and the collection of data draw most of the attention. The moment when "masses" of data, sometimes in the form of tables and graphs become available a whole new set of issues arises. There is a strong risk that, when interpretative frameworks are missing, and communication between practitioners and evaluation technicians is complicated, evaluation results will be under-utilized.

The use of the school effectiveness knowledge-base to identify relevant factors to be included in indicator systems and school self-evaluation instruments has been discussed at length in earlier chapters. An additional application rests on the assumption that the model also implies various "logical" possibilities for interpreting information.

In the first place the distinction between inputs, processes and outcomes, possibly including contextual factors, offers a helpful, very basic, classification of types of information.

Secondly, information could be used either in a disjointed, descriptive way, whereby each indicator stands more or less on its own, and is evaluated against certain norms or standards, or the information could be used by combining certain types of variables with others.

When the information is used descriptively, measures of central tendency, like the mean or average of a set of scores, are often used as a summary statistic. For example all positive responses on a particular sub-scale of an instrument that measures parents' perceptions of the school, can be added for each parent and, based on the total of all positives scores of all parents, the average is computed as this total divided by the number of parents. If the school would have stated in advance that it would be happy with parents' perceptions if the average would have been, for example 70% positive responses, then 70% is the norm or standard to decide on a positive or not-positive judgement.

It may also be informative to consider disparities in the data. For example one could look at the difference between the minimum and maximum score; this value is known as the range. So called interquartiles indicate the values on the scale that is marked by the 25% lowest scoring respondents, the next 25% higher up in the scale and so on until the 25% highest on the scale. From the position of these four points it can be seen whether most respondents are at the high or low end of the distribution. The educational relevance of measures of variation or disparity is equity. From the perspective of equity, it might be an explicit objective to keep the differences in scores of the pupils between certain limits, for example.

When it comes to relating certain variables to others, the school effectiveness model emphasizes two major types of associations. The first is the construction of so called value added output or outcome indicators (see Chapter 13). The second refers to the key-objective of school effectiveness research: finding out which malleable school or classroom factors "work" in the sense that they are positively associated with achievement.

Both types of associations require statistical modeling and analyses, as is discussed in Chapter 13 as far as the issue of value added school performance is concerned. It is

beyond the scope of this presentation to explain the principles of the techniques that are used to investigate whether certain school or classroom characteristics are positively or negatively associated with educational achievement. The basic logic of it, might also be applied in less formal types of analyses, better referred to as explorations, within the framework of school self-evaluation. The key word in finding out in educational practice whether a certain approach works better than another is *comparison*. So, for example if parallel classes of pupils at the same grade level, follow different teaching methods and the results are much better in the one than in the other, this gives reason to suppose that it was due to the difference in teaching method. This approach can, of course, be applied to other educational aspects as well, like different teachers, use of computers, or different forms of grouping pupils.

So far two ways of making use of the school effectiveness knowledge base for school self-evaluation have been referred to. In the firsts place the factors identified by school effectiveness research can be used as a source of inspiration in selecting relevant phenomena to be included in school self evaluation and for the identification of indicators. In the second place does the "logic" of the school effectiveness model point at certain ways in which information might be looked at, as disjoint, descriptive information, as a way to conceptualize "value-added" outcome indicators and as a general logic of evaluative comparison in educational settings. There is still a third important way of using the school effectiveness knowledge base.

To the degree that school effectiveness models, as the one depicted in Figure 11.2 in Chapter 11, are validly established and strongly supported by empirical evidence, they can also be used in a prescriptive way and point at directions for change and improvement. So, if for example a school shows educational leadership at a relatively low level, and achievement is below standard, strengthening the educational leadership could be a relevant "therapy" for this diagnosis. Although there is still a lot of uncertainty about the validity of school effectiveness models, and they may differ somewhat between national educational contexts, such a functioning, if prudently applied, can still be considered. This should definitely not be seen as following a cooking-book with recipes, but be strongly embedded in the practical know-how and knowledge of the particular situation of the school staff.

A final aspect of the use of the integrated school effectiveness model is the assumption that conditions at school level are somehow facilitating conditions at classroom level. So, for example, an orderly atmosphere during classroom work can be seen as supported by an orderly school environment and clear disciplinary rules at school level. This notion could be used for discussing in school teams the extent to which classroom level problems might be partially resolved by means of school level solutions.

16.3 Organizational and Communicative Aspects of Information Use

What was presented in the previous section can be described as providing a substantive educational framework for interpreting and using the information that is yielded by means of school self-evaluation. However, there are different issues at play as well. One of these issues refers to the organizational and communicative aspects of information use

and is discussed in this section, the other refers to the applied "policy"-context of school self-evaluations and is presented in a subsequent section.

There exists an interesting research literature on the use of the results of evaluation research results by policy-makers. The assumption when applying evaluation research is that the results will play an important role in policy decisionmaking. According to the rational ideal (see Chapter 5), evaluation research provides evidence on the attainment of policy goals. If the evidence would point out that these goals are insufficiently attained, the program that is the object of evaluation should be modified or even terminated. The research in question, however, pointed out that in many cases the evaluative conclusions of evaluation researchers had *no* implication for decision-making (see the introduction of this issue in Chapter 4). Information was disregarded, evaluation reports disappeared in drawers for ever, politicians distorted the information, used it selectively (only those aspects that were judged favorably given publicity) or ignored to the degree that programs indicated as successful by evaluation research were terminated and unsuccessful programs continued. Another finding of the evaluation utilization research was that the degree to which the technical quality of evaluations was scrutinized depended strongly on the degree to which the results supported the point of view or political stance of the users.

Interesting theoretical interpretations of these findings were given. For example, it was stated that the rational, also termed "linear" or "instrumental" concept of research use depended on an invalid model of public policy-making. It was maintained that decision-making should be seen as incremental (small steps at a time, a lot of negotiation and compromise, unclear goals and shaped by conflicting interests of relevant actors). It was also proposed that in such a policy-context research results could only penetrate slowly and use should be seen as conceptual, gradually shaping the frames of reference and perspectives of key actors, rather than instrumental.

The "political economy" of evaluations can also become visible at school level. In evaluation studies that were carried out in the Netherlands during the seventies there were frequent examples of teachers boycotting the data collection procedures set up by external researchers, because they wanted to prevent that pet-programs were criticized. Thus making evident the power data providers have in social research in which the stakes are considered to be high.

As a reaction some evaluation theorists proclaimed a type of evaluation, indicated as *utilization focused evaluation* which distinguished itself in propagating a dialogue between researchers and practitioners or policy-makers, seeking commitment from decision-makers with the evaluation and by trying to present the evaluation result in a "user friendly" way. Often, but not always, did utilization focused evaluation theorists vouch for the application of qualitative methods, because they were less "authoritarian" and offered opportunity to provide information in a way closer to the narrative of practitioners.

Huberman (1987) developed a framework in which utilization is described as depending on structural organizational background conditions of the key actors in providing and using research and on communicative aspects.

Although Huberman's conceptual framework was developed with regards to the use of research based information by policy-makers, it can, to some extent be generalized to the situation of school evaluation.

His framework distinguishes three partial models, an organizational model of the researchers (or evaluators), an organizational model of the users and a model which shows the efforts made to stimulate dissemination and use of the research or evaluation results.

In the *organizational model of the researcher* several factors referring to the setting and status of the researchers (evaluators) are included. The experience they have in policy- or practice-oriented research is one example, and the incentives or "disincentives" there would be for them to invest a lot in the dissemination and facilitation of use of the research results is another. A disincentive for academic researchers could be that the time spent on dissemination to practitioners or policymakers would be lost to activities, like preparing journal articles that provide academic status. When schools employ external technicians to help with school selfevaluation they may also encounter cases where the technicians in question simply have no experience in feeding back information to a less technically trained audience.

Apart from the characteristics of the researcher/evaluator or the unit he or she works in, is the linkage, or structural interdependence, with the organization that is using the results. In the case of school self-evaluations this structural interdependence can be thought of as optimal, since the initiator and controller of the self-evaluation activities belongs to the same organization, i.e. the school where the evaluation activities take place. Apart from structural ties of the organization of the researcher/evaluator with the user's organization, there are also procedural aspects involved. In this respect strengths of personal relationships between researcher and users is one factor and the role of intermediaries is another.

The interesting tension in these relationships is that on the one hand a certain distance between user and evaluator is presupposed as part of his or her perceived expertise and professional credibility. On the other hand there is also a certain demand for closeness, in the sense that a certain involvement and commitment from the part of the researcher/evaluator is seen as important as it comes to using the evaluation's results. This inherent tension is perhaps the core of what makes evaluations difficult in an organizational and political sense. When, in the case of school self-evaluations colleagues are taking this role, this tension between intellectual distance and "natural" closeness may be particularly difficult to handle. Colleagues' communicational skills in either role (evaluator and user) and clarity about the mutual roles and functions in self-evaluation are important means to overcome these difficulties. In school self-evaluation external facilitators could act as intermediaries. On the one hand they may enhance the technical credibility of the self-evaluation activities, on the other hand they could work as external changeagents and help in optimizing communication processes.

The organizational conditions of the researcher and the structural and procedural aspects of linkage functions result in actual dissemination activities and actual dissemination activities that can be qualified in terms of *intensity* and *quality*. Investment is evident from the amount of time and level of expertise involvement specifically in dissemination activities. Quality of the dissemination efforts is characterized in terms of the smoothness of the execution of the activities, user specificity of products, multiple channels to convey the information, a personal touch in the transmission of the results to the user, and the quality of written products.

Quality aspects of written products are: readability, specificity and operationality, focus on malleable variables, incorporation of user context, realism of recommendations, sensitivity to local susceptibilities and by attractiveness of products (humor, packaging, graphics) (ibid, p. 602).

The model of the evaluator finally includes costs and benefits on the part of the researcher/evaluator concerning the efforts to optimize dissemination. On the benefit side could be increased understanding and skills to operate in practice oriented research settings. On the cost sides one might think in terms of trade-off between investment in quality of the technical execution of the evaluation and efforts invested in dissemination and use. A questionable aspect for academic researchers/evaluators is the degree to which they perceive to be rewarded for successes in practice or policy oriented research as compared to more fundamental academic research.

The main components of the organizational model of the researcher/evaluator are depicted in the upper three blocks shown in Figure 16.1.

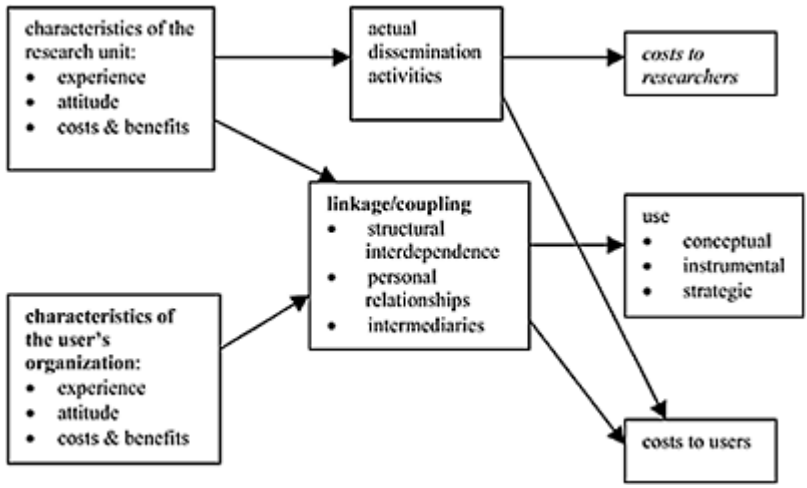


Figure 16.1 Modelling the use of policy-oriented research. Adapted from Huberman, 1987, p. 597.

In the *organizational model of the user* relevant entrance characteristics like earlier experience with research and evaluation, know-how concerning research and a positive climate for the use of research results are distinguished. The linkage factors between the researchers' organization and the organization that is to use the results are similar to those in the model of the research organization (see the central block in Fig. 16.1).

More specifically, as part of this block, seen from the perspective of the user organization, are *predictors of local use.* These are:

• the users' understanding of the main findings;
• amount of organizational time and resources devoted to the uses of findings;

• compatibility of findings with users' opinions;
• perceived quality/validity of the study;
• compatability of findings with the organization's objectives.

In the organizational model of the users the results of the dissemination activities are described as different types of use. A major distinction that is made is between *conceptual use* and *instrumental* use. In conceptual use the evaluation results do not lead to immediate actions or decisions, but to a gradual and incremental re-shaping of frames of reference of the users. For example exposure to the logic of educational evaluations may make school teams more sensitive and focused on measurable outcomes of education, or to variation in teaching methods. Instrumental use is the more classic idea of decision-oriented evaluation, where there is an immediate and concrete action following the interpretation of evaluation outcomes. For example the school-wide adaptation of a set of textbooks, after evaluation in a few pilotclassrooms has shown positive results.

A third type of use that Huberman distinguishes is *strategic use.* In the case of strategic use evaluation results are used in a selective way, in order to defend vested positions and interests.

Use of research may be further qualified in terms of the extent of use and the scope of the impact.

A final set of factors that is part of the organizational model of the user refers to the costs of use. This involves the distinction of negative impacts on the user like confusion and increased uncertainty, delay of actions, and intra-organizational tensions and conflict.

Huberman's model is quite rich in making explicit the complexity of organizational, motivational and "political" factors that are at play with respect to the implementation, interpretation and use of evaluations. Awareness of these factors in planning and implementing school self-evaluations may help in bringing about adequate use of the results. Among the (positive) secondary or side-effects of implementing school self-evaluation the organizational learning aspects that result from the very process of preparing and implementing school self-evaluation activities should be stressed. West and Hopkins (1997) refer to this phenomenon as school self evaluation *as* school improvement.

16.4 Contexts of Use

Huberman's model, presented in the previous section, sensitizes the student of school self-evaluation to the organizational and political dimensions of evaluations. These political dimensions become even clearer when different contexts of use and application of school self-evaluations are considered.

External accountability

As case-studies on school self-evaluation in European countries indicate, it is a quite common phenomenon that school self-evaluations arise as "spin-offs" of external evaluations. In such cases all kinds of combinations between external and internal functions of the school evaluations may occur, varying from tailor-made selfevaluations

of individual schools to school self-evaluations that are spin-offs of national or district-level assessment programs, where school results are fed back to individual schools (Van Amelsvoort & Scheerens, 1997).

a) School self-evaluations that serve internal and external purposes and are subject to meta-evaluation by inspectorates.
b) School self-evaluations that are explicitly aimed at providing information to external constituencies as well as aimed at use of the information for school improvement processes.
c) Self-evaluations that are part of improvement programs that involve a number of schools (evaluations may have the additional purpose of assessing the effects of the school improvement project as a whole).
d) Tailor-made self-evaluations of individual schools.

If the results of school self-evaluations are made available to administrative units above the school, which may use their discretion to "hold schools accountable" for the results, schools may become more cautious. Negative aspects of feeling judged may arise, like staff feeling threatened by the evaluation and tendencies to strategic application of information gathering and use.

Consumer-orientation

School self-evaluations and their results may also be used as part of an overall policy of schools making themselves more responsive to local constituencies and "consumers" (e.g. parents, local organizations). Such a strategy may result from the needs felt by the school to "market" itself in a local competition of obtaining a sufficient influx of new students. It may also be externally induced in situations where the results of school functioning are made public in general or local media, like the practice of publishing "league tables" in the United Kingdom and the formal duty of schools in the Netherlands to publish an annual "school guide".

Particularly when the publishing and making available of evaluation outcomes is externally induced, and less the active choice of the school itself, the consumer orientation may lead to the same cautiousness as in the case of accountability oriented evaluation, and to similar less positive side-effects.

Organizational learning

When the results of school self-evaluations are exclusively used by the school itself, for making a diagnosis and possibly trying to improve its own functioning, at first sight the political stakes may seem less high. But still, even within-school use of self-evaluations may lead to teachers feeling threatened, particularly when they would have the impression that evaluations are used as appraisal by the school's management.

One should be careful in not overstating the negative implications leading to political and strategic use of school self-evaluations, however. Particularly when there is broad participation in planning and execution of evaluation activities and when absolute clarity about the objectives and means of the evaluation is provided, there is less likelihood that

evaluation apprehension will get the upper hand. Of course this is easier for the organizational learning context than for the two other contexts of use.

If the process of interpretation and actual use of the information goes well it is likely to function as an incentive for carrying on with the school self-evaluation activities.

16.5 The Confidentiality of the Results from School Self-Evaluation

Systematic evaluation is full of inherent tensions and contradictions. At the same time evaluations are expected to be "objective" and "engaging", they are about "facts" and "judgements", they often have an "external" element and are expected to be used "internally". Actors are sometimes expected to play the rather passive role of information providers, but then they are also expected to be active partners in the shaping of evaluation questions and the interpretations of results.

In the case of school self-evaluation these tensions are partially avoided because it seems to be evident that school self-evaluation takes a clear position on what side of these pairs of opposites it stands:

- it is internal rather than external;
- it is improvement rather than accountability oriented
- it uses methods that are transparent to the practitioners
- all actors in the school are expected to play an active rather than a passive role

Maybe one could conceive of a prototype of school self-evaluation that has surmounted these tensions and is totally at the non-judgmental, "learning" side. The concept of the "reflective practitioner" developed by Schön (1983) comes close to this ideal-type situation. However, in many examples in actual practice school selfevaluation has also external, objectifying and judgmental aspects. Some might even argue that some degree of judgement and objectifying is necessary to evoke learning. Evaluation needs to have "an edge" or even "a bite".

In this section the confidentiality of evaluation results will be discussed in the larger context of these inherent tensions in systematic evaluation. For simplicities sake the terms "internal" and "external" will be used to indicate the polar sides of the various continua that are sketched in the above. The key question is not to defend a particular choice between the poles of the continuum, but rather how to deal with "external" elements in school self-evaluation in the most appropriate and acceptable way. In order to provide solutions some of the "standards" that are available in the evaluation literature will be discussed. The question of "ownership" of school self-evaluation will be addressed as well and a proposal to link decisional discretion to "circles of confidentiality" will be discussed.

16.5.1 Evaluation standards

In August 1980 the Standards for the Evaluation of Educational Projects and Materials were published (Joint Committee for Educational Evaluation, N.Y.: MacGraw-Hill, 1981)

These standards have the function to regulate professional practice in the field of educational evaluation. In the standards of the Joint Committee 29 standards are distinguished, divided over four main areas:

- accuracy standards (which concentrate on research-technological criteria, like objectivity, reliability and validity of procedures),
- utility standards (relevance for policy and educational practice),
- propriety standards (ethical issues), and
- feasibility standards (organizational and technical aspects).

In 1982 another American committee published a set of standards that were not exclusively formulated for education, but pertained to all societal domains where program evaluation takes place, the "Evaluation Research Society Standards for Program Evaluation" (San Francisco: Jossey-Bass) These standards have been categorized in a different way. Examples of specific standards that are related to what the Joint Committee would call "propriety standards" and which relate to the ethical quality of agreements, arrangements and relationships between the main actors in evaluations (initiators, technicians, data-providers and users) are presented below.

(7) "Restrictions, if any, on access to the data and results form an evaluation should be clearly established and agreed to between the evaluator and the client at the outset".

(8) "Potential conflicts of interest should be identified, and steps should be taken to avoid compromising the evaluation process and results"

(9) "Respect for and protection of the rights and welfare of all parties to the evaluation should be a central consideration in the negotiation process"

(11) "All agreements reached in the negotiation phase should be specified in writing, including schedule, obligations and involvement of all parties to the evaluation, and policies and procedures on access to the data. When plans or conditions change, these, too, should be specified."

(18) "The necessary cooperation of program staff, affected institutions, and members of the community, as well as those directly involved in the evaluation, should be planned and assurances of cooperation obtained (see standard 11)."

(21) "Evaluation staff should be selected, trained and supervised to ensure competence, consistency, impartiality, and ethical practice"

(22) "All data collection activities should be conducted so that the rights, welfare, dignity, and worth of individuals are respected and protected."

(25) "The data collection and preparation procedures should provide safeguards so that the findings and reports are not distorted by any biases of data collectors."

(28) "Data should be handled and stored so that release to unauthorized persons is prevented and access to individual identifying data is limited to those with a need to know (see standard 7)"

(39) "Findings should be reported in a manner that distinguishes among objective findings, opinions, judgements and speculation"

(40) "Findings should be presented clearly, completely and fairly (See standard 39)"

(41) "Findings should be organized and stated in language understandable by decision makers and other audiences, and any recommendations should be clearly related to the findings."

(46) “Persons, groups and organizations who have contributed to the evaluation should receive feedback appropriate to their needs.”

(47) “Disclosure should follow the legal and proprietary understandings agreed upon in advance (standard 7), with the evaluator serving as a proponent for the fullest, most open disclosure appropriate”.

(50) “Evaluation results should be made available to appropriate users before relevant decisions must be made”

(51) “Evaluators should try to anticipate and prevent misinterpretations and misuses of evaluative information. (The evaluator, of course, cannot be held responsible for misuses of evaluative information. Nevertheless…promotion of an open exchange of information should be a part of the evaluator’s responsibility.)”

(53) “Evaluators should distinguish clearly between the findings of the evaluation and any policy recommendations based on them.”

(55) “Evaluators should be aware of the apparent conflict between their role as an evaluator and any advocacy role the choose to adopt.”

A few remarks should serve to “contextualize” these standards. First of all they pertain to program evaluations, usually carried out by a specialized evaluation research institute on the bases of specific requests by a contractor. More recently evaluation forms that are more like “monitoring” ongoing practice, as compared to evaluation as specific innovatory programs, have gained in importance. School selfevaluations are mostly of the “monitoring” as compared to the “program evaluation” type. Nevertheless in school self-evaluation there are comparable actors, playing the roles of initiator, technician, data-provider and user, even if the same persons may sometimes have more than one of these roles.

Secondly, these standards date from two decades ago. Since then a shift in priority among the major categories of standards seem to have taken place, in the sense that one is less fanatic about the accuracy standards, while propriety, utility and feasibility standards have gained in importance. In a symposium at the annual meeting of the American Educational Research Association in New Orleans (2000), David Nevo spoke of three shifts:

- changing role of the evaluator from expert to coach, or “critical friend”;
- a shift from dependency of practitioners to self-determination and capacitybuilding;
- a shift form independent judgement by an evaluator to collaboration.

In the third place the standards express a preoccupation with planning in advance and committing plans and agreements to paper. In this sense they appear quite legalistic and formalistic as a kind of modern witch-craft to lure the complexities of real-life into the clear-cut categories of blue-prints. Yet, there is definitely some value in thinking in advance about the issues expressed in the standards that were cited in the above, and in drawing up written agreements.

What is most directly relevant to the issue of confidentiality is that there is a clear commitment to being open, and informing all persons that have played a role in the evaluation.

16.5.2 Objectivity and ownership

In school self-evaluations participants are at the same time "objects" that are studied and "judged" and owners playing an active role in the design and interpretation of the objectified facts and explicit judgements. How is this possible and how can the inherent tensions and ambiguities of such arrangements be resolved? There are a few handles to approach this difficult problem:

- the hierarchical structure of schools as organizations;
- overall perspectives and attitudes;
- being explicit about purposes, audiences and the rights of respondents.

Hierarchy and subsidiarity

Schools have at least a degree of hierarchical structure. Images of school as organizations, in the sense of loosely coupled systems and professional bureaucracies underline that school hierarchy is not very tight. There is supposed to be a lot of freedom and autonomy at the level of individual teachers. To the extend that regularities are imposed on teachers from outside they do not only, and probably not primarily, come from heads, but from professional standards acquired during training, from national curriculum guidelines and textbooks. In this type of organizational framework the principle of "subsidiarity" provides a rationale for the partial nature of hierarchical control: all things that can be accomplished at a lower level shouldn't be done by a higher level. All these characteristics call for a type of school leadership that is restricted, particularly in the instructional and pedagogical domain. But this is not what is meant when speaking about subsidiarity; in the case of subsidiarity and functional decentralization within schools, there is control, although it is supposed to be minimal control.

Evaluation can be seen as a regulatory mechanism in its own right. As a category it is "more minimal" than proactive planning, where, for example, school heads would hold teachers responsible for carrying out a school work plan to the letter. Still evaluation and school self-evaluation can function as an instrument of internal accountability, where the head checks the performance of the individual teachers. In this way school self-evaluation could re-instate authoritative hierarchy. Within a context of accountability certain evaluation types are more imposing than others. Output evaluation is less imposing and "more minimal" than process evaluation. When only outcomes are monitored professionals can still have their autonomy in choosing, designing and implementing the means and methods to reach the objectives. A technical problem is to attribute merit of individual teachers to high output in the sense of student achievement. This can only be done fairly if output is expressed as progress, or in other terms as "value-added". Given proper instrumentation to monitor pupils' progress evaluative control by the head can be seen as minimal control and as still respecting subsidiarity. In such a context teachers would be required to help in collecting data that will be used to judge them, which would also, of course, imply that the information is disclosed to the head. The information could remain confidential between the head and individual teachers.

In such a context teachers would also know the achievement of individual students, and possibly use this to adapt their teaching. Information of student performance would also be disclosed to the parents of a particular student. It could be a matter of debate whether school heads should have the achievement results of each and every pupil. Under strict assumptions of subsidiarity this would not be required, unless the head would need this information for special measures he or she would have to take him/herself, e.g. hiring a remedial teacher.

Perspectives and attitudes

Schools can choose to avoid the language of hierarchy and control. Maybe it is more acceptable to speak in terms of the head as a "coach", or at best a "leader" instead of a manager. Whether this is completely in line with reality or more like a "euphemism" is an open question. In some sense the idea of the head as a coach is more imposing and paternalistic than the one of a "minimal manager". Within a context of school improvement the head teacher as a coach might be inclined to emphasize process rather than output evaluation, since he/she might want to give advice on teaching methods. (Ideally, advice on teaching methods should also be based on information or knowledge about process output relationships, however) In such a situation process information would also be likely to be shared with the rest of the teaching staff. This is the ideal of participatory school development, professional consultation and team work. It asks for a considerable amount of being open about each teacher's teaching methods and classroom management. Particularly when the whole staff has a say in the priorities of evaluative activities and the development and choice of instruments such conditions op being open may be strengthened. Under the perspective of using evaluation for participatory school development an overall attitude of mutual trust and openness may be evoked that takes away the threats of accountability based evaluation. In such a context there could be less need for confidentiality of evaluation outcomes.

Next to the ideology of participatory planning, the head as a coach and "empowerment" of teachers, there is another ideal-type image, namely that of the school as a learning organization; which was discussed at length in other chapters.

The overriding attitude here would be investigative: teachers as researchers. Here too, collective involvement of head teacher and staff would be likely to be important. Learning form evaluations and readiness to adapt would override feelings of being controlled and manipulated by evaluations. Confidentiality would be less of an issue in a power-game and just be determined by the propriety standards as described in the previous section.

Being explicit about the purposes, the audiences and the rights of respondents of the evaluation.

Apart from a managerial, a counseling or an investigative orientation, confidentiality depends on the concrete purposes of school (self)-evaluation.

As was also made explicit in some of the standards that were cited it is important to be as clear as possible about the concrete objectives of the self-evaluation in advance. This makes it possible to address the question of confidentiality of results at an early stage.

Examples of concrete purposes are:

- to assess the progress of individual students in order to adapt instruction, to select for difficulty levels of further courses, to determine whether they have reached standards required in examination and to inform the parents (teachers, pupils, parents, head teachers);
- to assess the success of a sub-unit at school, a location, a department, a teachers or a classroom on the basis of either outcomes, processes, or process outcome combinations (head teacher, staff as a collectivity, individual teacher, inspectorate);
- to determine the image of the school in the local community and the satisfaction of the parents (municipality, parents, head, staff);
- to determine the well-being of teachers and students (head teacher, staff, individual teachers, parents, students);
- to assess the functioning of the organization in terms of processes or outcomes (head teacher, staff, municipality, province);
- to assess the functioning of the head and the coordination structure as such (the staff as a collectivity, municipality or school board, province).

The rights of respondents are specified in the evaluation standards. Most importantly respondents are always entitled to some kind of feedback on the basis of the information they have provided. Not unusually this is done by showing the score or standard a respondent has attained in relationship to the average and dispersion of the collectivity to which the respondent belongs.

Circles of confidentiality

Basic principles that bear on the issue of confidentiality are:

- legal requirements;
- whether openness or confidentiality may have harmful side-effects;
- whether openness or confidentiality serve utilitarian principles.

Professional standards can be seen as semi-legal principles that should ideally also have covered the second criterion. Nevertheless it can be assumed to be the responsibility of initiators of evaluations to ask this question (about possible harmful effects) anyway.

The last and more pragmatic principles can be tackled by referring to the subsidiarity principle once again. A paraphrasing of this principle with respect to the issue of confidentiality would be that evaluative results should only be identifiable to users and audiences to the degree that these users need the identification for the interventions that are within their discretion.

16.6 Remaining Dilemmas in School (Self-)Evaluation

There are some inherent tensions in the evaluation endeavor as such, and with respect to the many choices among different approaches. Three of these “major dilemmas” will be discussed in this section:

- can improvement and accountability purposes of school evaluations be combined?
- the qualitative/quantitative debate;

• commitment and objectivity in school evaluations.

Is an effective mixture of accountability and improvement perspectives feasible?

The context of application of school evaluation is clearly different depending on the question whether the context is "accountability" or "improvement". In the case of accountability the audience, and probably also the initiative of the evaluation is external to the school. The conclusions of the evaluation may be used by these external parties, either administrative levels or consumers of education, for decisions that affect the school in ways it might not have chosen on its own accord. Accountability is associated with external control.

When improvement is the context of application the evaluation will most likely be initiated and, maybe also, conducted internally. Evaluative conclusions remain within the school and are supposed to be used in processes of school improvement. The general term for the application of improvement-oriented evaluation is "learning" rather than control.

In the case of accountability, the evaluation procedures will usually take place on a larger scale, involving many schools, a situation that will make it more likely that standardized and rigorously quantitative methods are prepared and used. School evaluation could, in principle, be limited to just one school, which makes it less likely that resources and expertise would be available to design and implement sophisticated qualitative approaches. In this sense school self-evaluation is likely to be "softer" than accountability-oriented school evaluation.

There is likely to be a different attitude among school staff when evaluation is external, accountability-oriented, as compared to internal, improvement-oriented. In the first case resistances against evaluation procedures are more probable, because staff may feel threatened by the evaluation and possibly ensuing decisions. If matters are put like this, it appears that accountability and improvement-oriented evaluation are far apart, and not likely to be integrated. In actual practice, however, one can observe many combination forms and "in between" types of evaluations. In a study of school evaluation procedures in four European countries, Van Amelsvoort and Scheerens (1997) concluded that all cases of school evaluation studied appeared to be both "self"-oriented and accountability-oriented. They propose five categories of school self-evaluation "which show an increasing degree of combination with external accountability-oriented motives:

a. Tailor-made self-evaluations of individual schools;
b. Self-evaluations that are part of improvement programs that involve a number of schools (evaluations may have the additional purpose of assessing the effects of the school improvement project as a whole);
c. School self-evaluations which are explicitly aimed at providing information to external constituencies as well as aimed as use of the information for school improvement processes;
d. School self-evaluations that serve internal and external purposes and are subjected to meta-evaluation by inspectorates;
e. School self-evaluations that are spin-offs of national or district-level assessment programs, where school results are fed back to individual schools.

Reconciliation between accountability-oriented and improvement-oriented evaluation is more likely when the external control element, most notably the taking of sanctions, is less severe. And this may be the case in many educational settings, particularly when schools have a rather large degree of autonomy. In a study on the use of School Performance Reporting (SPR) in the USA, Cibulka and Derlin (1995) conclude that very few instances of actual policy use of SPR-results could be observed.

A more pragmatic argument for integration is the fact that evaluation is an investment, takes time and uses scarce resources, and that it is therefore efficient to try and use evaluation information for more than one purpose. Both internal school self-evaluation and accountability-oriented evaluation benefit from a proper, possibly "value-added" assessment of learning outcomes.

The quantitative/qualitative debate

In the history of evaluation research, there was a certain period (the seventies) when there was an ongoing debate, mostly stimulated by scholars who propagated "qualitative methods", against the main stream of quantitative evaluation research (cf. Patton, 1978; Guba, 1978; Stake, 1975; Parlett & Hamilton, 1972; Eisner, 1979).

Qualitative approaches have the following characteristics:

- use of "open" research formats, such as "open" interview questions, and "free" observation;
- a strong dependence on the views of persons that are part of the "evaluandum" (the evaluation object);
- narrative, and sometimes so-called "thick description" of the object situation rather than quantitative output (tables, graphs); • smaller aspirations towards generalizability of findings because of the fact that fewer units or codes are studied "in depth".

Authors that have published about qualitative evaluation methods differ among themselves with respect to the application of methodological criteria, like objectivity and reliability. Some authors seem to take the position that well-documented elaborate descriptions which are supported by "participants" in the object situation are sufficiently convincing. Others (e.g. Denzin, 1978; Webb et al., 1966, Yin, 1981) propose methods and methodological checks which enable examination of the trustworthiness of qualitative approaches. Triangulation is the best known of these methods. In triangulation the same object is observed or described on the basis of different data collection procedures, for example, describing a teacher's approach on the basis of self-reports, evaluators of pupils and direct observation by a colleague. When the results of these different procedures converge, this will be seen as proof of the credibility of the description.

Presently, there is a more common understanding that qualitative and quantitative approaches each have their strong and weak points, and that, sometimes, a combination is the best solution. The strong points of qualitative approaches are elaborated, illuminative descriptions which are close to the world of the persons in the object situation, while the strong points of quantitative methods lies in a better position with respect to generalizability and a straightforward possibility to verify reliability and objectivity. in

subsequent chapters many examples will be provided of qualitative and quantitative methods in school evaluation.

Commitment and objectivity in school evaluations

Perhaps the major inherent tension in evaluation is that it requires both "distance" and "participation", both objectivity and commitment. Theoretically this tension can be resolved by referring to the various phases or stages of evaluation. Objectively then could be seen as an important requirement for systematic information gathering, while commitment would become a major principle in the stage of the application and use of the evaluation results. In actual practice, questions of commitment and objectivity are likely to play a role in each phase, starting with the design of the evaluation plan. Particularly when school evaluation has the characteristics of internal, improvement-oriented self-evaluation, commitment appears to be the most important desideratum.

Yet, there can be no evaluation, without at least a certain distance, and without at least the possibility of an evaluation result that is negative, or critical.

In earlier sections we referred to political aspects in program evaluation. In school evaluations such aspects are likely to be most prominent when "the stakes" of external accountability are considered high. In those cases attempts to manipulate evaluation results, e.g. by training test items, are not unlikely. The best approach to prevent political biases in school evaluations would be to reach agreement on the aims and methods of the evaluation and to create a non-threatening atmosphere with respect to the use of the findings. Part of such agreements would have to be a certain acceptance of "the rules of the game", including the possibility of critical outcomes. When solely qualitative methods are employed in school self-evaluations such agreements would be particularly necessary, since "open" approaches are particularly vulnerable to distortions if participants would see a reason to bring them about. In subsequent sections the role of an external advisor, sometimes referred to as a "critical friend", in school self-evaluations will be discussed and illustrated.

16.7 Implementation Issues; Applicability in Developing Countries

In this final section feasibility of implementation of school self-evaluation approaches will be considered. Again the evidence is based on experiences in Europe. In particular the results of three research projects funded by the European Commission will be used, these are the EEDS-project (Evaluation of Educational Establishments—Van Amelsvoort et al., 1998); the INAP project (Innovative Approaches to School Self-Evaluation—Tiana et al., 1999) and the EVA-project (Quality Evaluation in School Education—e.g. MacBeath et al., 1999). All three projects provide extensive information on case-studies of school self-evaluation activities in European countries.

Reconsideration of the internal/external dimension

The EEDS and INAP projects found that in practically all cases that were studied in five countries (Scotland, England & Wales, Spain, Italy and the Netherlands) there was a

strongly *external* impetus to the school evaluation projects that were studied. The projects that were studied were usually hybrid forms in which external and internal elements were both present. In all cases networks of schools collaborated in school (self-) evaluation activities. Mostly initiatives came from above school units, municipalities, local education authorities or regional support agencies. In all cases schools obtained external support and mostly used externally developed instruments. In a minority of cases schools adapted externally developed instruments or developed their own instruments with the help of external experts.

The evidence from the EVA-project illustrates genuine school-based initiatives more frequently, although external support is usually present in these cases as well.*)

The reality of school self-evaluation, particularly in countries where this practice is a very recent phenomenon, is *"external evaluation with an increasing degree of school participation"* rather than genuine school self-evaluation. So far, the most common initiation and implementation strategy in Europe seems to be "spin-off' from externally initiated types of school evaluation.

Nevertheless, there are other examples that are more genuinely school-based as well. The example of Dutch primary schools that buy their own pupil monitoring system, and which was referred to earlier, is a case in point. There are also some very positive experiences where schools work with external experts on setting

priorities and standards for school self-evaluation (MacBeath, 1999; Scheerens, 1999). These latter examples are tending towards what West & Hopkins describe as evaluation *as* school improvement.

The relevance of these experiences for developing countries is twofold. Firstly, school self-evaluation can be initiated very well by exploiting the spin-off of external evaluations, like national monitoring systems or evaluations of development projects. Prerequisites for such practice are that information is available at lower levels of aggregation (schools, classrooms) and that specific measures are taken to feed this information back to schools in a comprehensible way.

Secondly, the introduction of basic and simple forms of school self-evaluation in schools in developing countries can be used as a feasible and practical way to bring about a process of self-reflection and school improvement. This latter practice, however, would require a local cadre of support staff, e.g. an inspectorate.

External support

In all cases described in the EU-studies there was some kind of external support for the schools that participated in school self-evaluation projects. The type of required support, as a matter of course, depends on the type of school self-evaluation that is chosen. There are two main areas of support: technical support and management support in creating and maintaining the organizational conditions required for an effective use of self-evaluation. In cases where self-evaluation is largely a spin-off of external evaluations, involving many schools, data will be processed and analyzed externally. Special efforts will need to be made to feed back data to individual schools in an accessible and comprehensible way. In these situations schools would also require some guidance in the interpretation of results, application of standards and benchmarks. *) These outcomes reflect, to some extent, the focus, or sampling bias of these studies, where EEDS and INAP sampled self-evaluation projects, whereas EVA sampled individual schools in each EU country.

When the choice and development of evaluation methods is more of a bottom-up process, schools would require some technical guidance in providing a range of possible approaches, methods and instruments and in the technology of instrument development. As stated before, such collaborative activities, to some extent, are school improvement activities in their own right as they urge school teams to collaboratively reflect on major goals and methods of schooling.

Management support is needed to create and maintain organizational conditions necessary to conduct school self-evaluations. In fact the implementation of school self-evaluation is to be seen as an innovatory process, to which all principles of good practice apply. One of these principles is the essential role of the principal. Other aspects are seeking the involvement of all staff and external constituencies. A basic organizational requirement for good practice of school self-evaluation is the institutionalization of some kind of forum where staff can meet to plan evaluation activities and discuss results.

Apart from technical and managerial support, in many situations, schools would also require more substantive educational support in interpreting results and designing remediation and corrective actions to improve the school's functioning in weak areas. There is definitely the danger of creating an overload of evaluative information that is not fully exploited for its action potential. To put it differently, self-evaluation should not end in diagnosis but be actively used for "therapy". The required individual support to schools in interpretation of data, participation in the development of instruments and procedures and information use, appear to be conditions that are not easily fulfilled in developing countries.

Cost aspects

The need for external support and guidance is the more expensive to the degree that each and every school would develop its own "tailor made" approach to school selfevaluation.

Economies of scale, in working with networks of schools and projects involving many schools, are to be considered, when resources are scarce. School selfevaluation on the basis of data feedback from existing national assessment or monitoring projects exploits this principle even further.

Local support staff to guide schools in school self-evaluation seems to be an unrealistic pre-condition for many developing countries. There would be a lot of potential in small-scale pilot projects, however, where the use of school selfevaluation could be implemented and studied in the specific local context. Among other applications, such experiences could be used in the design of training courses as part of regular training of teachers and head teachers.

Experiments with in-service teacher training activities in school self-evaluation could also be seen as long-term investments in the building of local capacity in the directly practical and foundational skills that are at stake in creating schools that can handle autonomy and self-improvement.

The micro-politics of evaluation

Since evaluations—even school self-evaluations—ultimately lead to judgements and "valuing"—some categories of actors, particularly teachers, are likely to feel threatened. Traditionally schools have functioned according to the principles of the "professional bureaucracy" (Mintzberg, 1979), where enculturation and training in the profession is the key control mechanism and autonomous professionals are described as opposing rational techniques of planning and monitoring.

School evaluation activities have the potential of stimulating managerial control in areas which were traditionally safeguarded under the umbrella of the professional autonomy of teachers. The subsequent greater transparency of the primary process of schooling to external parties, e.g. the principal and the school board, has implications for the balance of power within schools. In the early literature on program evaluation clashes between evaluation experts and practitioners have been documented as the confrontation of "two worlds" (Caplan, 1982); and such tensions cannot be ruled out even when evaluation is internal and improvement-oriented. Several authors have therefore emphasized the creation of non-threatening conditions for school evaluation (Nevo, 1995; MacBeath, 1999). The role of the external expert should become more like an advisor and a "critical friend" to the school.

School evaluation can be perceived in a context of accountability and a context of improvement. Theoretically one would expect that evaluation apprehension would be stronger in an accountability as compared to an improvement context. In actual practice, at least in Europe, school self-evaluation often arises as consequence, spin-off or counter-balance to accountability-oriented assessments. Reconciliation and integration of accountability and improvement orientations is the more likely when the external control element, most notably the taking of sanctions, is less severe. In Europe there are examples where external accountability-oriented assessments, like the production of league-tables, actually function as the main incentive for schools to embark upon school self-evaluation which considers a broader spectrum of aspects of school functioning.

But even when there is no accountability at stake, and school self-evaluations are designed bottom-up, the issue of teachers feeling threatened arises. It is therefore important that school self-evaluation is clearly and explicitly introduced to all stakeholders and participants and that initial activities are experienced as intrinsically, professionally rewarding. Ultimately the relevance and use of data and application of standards for all school staff should function as the main incentive to sustained school self-evaluation.

The micro-politics of school evaluation are likely to differ according to the structure and educational culture of a country. Therefore, no generally applicable guidelines can be given for applications in developing countries other than the strong recommendation not to overlook the political aspects and all the repercussions they may have for issues of reliable data-collection, anonymity of results, facilitation of coupling databases and good professional cooperation between teachers, principals and support staff.

16.8 Conclusion

In this chapter school self-evaluation has been defined as a type of school evaluation where the school takes responsibility for its own evaluation. From an extensive overview of categorizations it appears that there are many forms where the school taking responsibility does not preclude the shaping of evaluation methods by external parties. Case-studies from Europe indicate that in many instances school self-evaluation occurs as a spin-off, consequence or counter-balance to external evaluation. Many of these case-studies show an orientation to accountability *and* self-improvement rather than an exclusive preoccupation with one of each. When it comes to applying school self-evaluation in developing countries the European experience of hybrid forms of external and internal school evaluation is seen as an advantage rather than a handicap. Similarly, from a methodological perspective, integration and combination of different "pure" types of school self-evaluation appears to be the rule rather than the exception.

Given the costs, the required expertise *and* the fact that in many developing countries system-level assessment and monitoring are already implemented or in a stage of development, school self-evaluation could get off the ground in the wake of these large-scale programs. Bottom-up developments, where schools design their own self-evaluation, should also get a chance, however. For these, small-scale pilot projects could be set up to explore the possibilities of school self-evaluation as a form of reflection and school improvement in its own right. Results of such pilots could have an important function in the shaping of initial and in-service teacher training programs.

A final observation—also for the application in developing countries—was that the micro-politics of evaluation should be an important focus of consideration in the way school self-evaluation is introduced and designed. Tackling this potential problem area well can avoid a lot of loss of energy in dealing with resistance, distortions and even corruption of evaluation.

School self-evaluation contains the possibility to bridge the distance between evaluation and school improvement, particularly when it is tackled as a joint learning experience from internal and external actors, like administrators, school leaders, teachers and external researchers. It is therefore to be seen as an important lever to educational change and improvement with considerable potential, also for developing countries.

17
A Practical Example of Developing and Using Value Added Indicators: The Lancashire LEA Value Added Project

17.1 Introduction

The theory and rationale of value added approaches was described in Chapter 13, as well as the methodology of creating value added indicators. This chapter aims to illuminate the theory and methodology by providing a practical example of a value added indicator system developed and implemented by one English Local Education Authority (LEA)—Lancashire LEA. Thus the case study focuses on the Lancashire LEA Value Added Project (VAP) and demonstrates how the systematic feedback of value added and other performance measures to schools can assist teachers and LEA advisors in evaluating and monitoring school performance and educational quality. The overall rationale, development and structure of the Lancashire VAP will be described as well as the implementation and impact of the project in one secondary school.

The Lancashire VAP was set up in 1992 to provide an innovative system of secondary school evaluation and self-evaluation via the feedback of student outcome and performance data. This information is intended to inform the improvement processes of state funded schools within the Lancashire LEA region. The collection of evidence for the case study was conducted via interviews with two key LEA advisers and the staff, pupils and a governor of one school, and also by analysis of key documents. The case study research was originally carried out as part of the EU Socrates funded project *'Innovative Approaches in School Evaluation'* and this chapter draws on that work (see also Tiana et al. 1999; Smees & Thomas, 1999).

17.2 The Development of the Lancashire Value Added Project

The original impetus for Lancashire LEA to begin to look at new ways of evaluating schools came from the external pressure of the UK conservative government's education policy in the late 1980's to increase the public accountability of schools. League tables of secondary school performance in England and Wales, based on the national system of General Certificate of Secondary Examination (GCSE), were planned to be published nationally for the first time in 1992 by the Department for Education (DFE) and a few poorly achieving Lancashire schools had been told they were going to be highlighted in the press. As a result of this climate a collaboration between LEA advisers and a group of

secondary head teachers was set up to discuss a fairer way of assessing school performance than the raw examination league tables. An additional pressure from central government came from a report by Her Majesty's Inspectorate (HMI) which criticized the authority for not using data enough. One LEA adviser reported:

> *"it was a challenge to us to start to improve the way in which we used data"*
>
> (LEA adviser and project manager)

The initial development work for the project involved a simple statistical approach, looking at the aggregated school means for attainment entry score at age 11 (Year 7) using the NFER Cognitive Abilities Test, compared to mean GCSE score at age 16 (Year 11) to examine the progress made by secondary pupils over a five year period. There were a number of methodological flaws with such an approach, and many schools soon began to question the appropriateness of a method that did not look also at the effect of pupil background factors such as social class and gender on attainment. It was at this stage that the LEA decided to seek help from external consultants and university academics.

From this new collaboration the Lancashire Value Added Research Project was set up in 1992. The aim was to create a system that could contextualize raw GCSE results, taking into account both prior attainment and pupil background factors. After an initial pilot analysis in 1992 based on 11 secondary schools, the main Value Added Research Project was set up in 1993 involving 87 schools. Since 1994 all 98 Lancashire secondary schools have been involved in the project.

17.2.1 The rationale of the evaluation system: accountability versus improvement

The LEA wanted primarily, a robust, 'hard' quantitative method of assessing schools' attainment in a fairer context. They were particularly fortunate in that they already had in place an excellent pupil tracking system, whereby all pupils in the LEA mainstream schools took the NFER Cognitive Abilities Test at entry to secondary school. It was also possible to collect a number of other pupil background details as well as the GCSE outcome results in order to carry out a value added analysis of the relative progress of pupils during their time at secondary school. As one of the LEA managers of the project points out:

> *"It's very powerful feedback to a school on their performance for them to have to face up to the fact that they may not have been doing as well as they could have done and they would have to do something about it. For quite a number of school it still is, never mind was, a real eye-opener for them"*
>
> (LEA adviser and project manager)

But the evaluation process was not intended for external accountability purposes, rather a tool for internal accountability and school improvement, in terms of assessing the

performance of different subjects and groups of pupils as well as the whole school. The LEA tries extremely hard to encourage schools to use the Value Added data confidentially for 'internal purposes only', not to disclose such information to parents or the press, to prevent any of the negative aspects of the raw league tables:

> "*A key element within it [the project] has been the integrity of the data because what we have never wanted to do is to publish an alternative league table"*
>
> (LEA adviser and project manager)

17.2.2 Management of the project

The management structure of the Value Added Project within the LEA is shown in Table 17.1. A key issue for the Value Added Project was to continue to include schools in the decision making process, and as a consequence the Value Added Working group was set up comprising of a group of secondary heads and the LEA project manager. The group meets two or three times a year to discuss new additions to the project or useful changes.

Table 17.1 The Management Structure of the Value Added Project Within the LEA.

INSPECTION and SPECIAL SUPPORT TEAM
ASSESSMENT SUPPORT GROUP
(Senior advisor)
VALUE ADDED PROJECT
(Overall responsibility Project Manager)
(2.5 data analysis staff)
(Association of secondary heads Value Added Working Group)
SCHOOLS
(Head teacher formal link)

The formal link with each school in the project is always through the head teacher, but it is usual practice for her or him to work with deputies, assessment coordinators or examination officers to analyze and disseminate the evaluation data to classroom teachers and other staff.

17.2.3 Evaluation instruments and feedback to schools

Since 1992 the Lancashire VAP it has developed substantially to incorporate a number of different types of evaluation feedback generated from the original Value Added research project[1], the National Consortium of Examination Results (NCER[2]) and the LEA. For example, the information now received annually by schools comprises four main types of feedback:

(i) Lancashire Value Added Research project—GCSE Value Added Analysis

The annual results from the GCSE Value Added analysis comprise a total of 44 separate Value Added scores involving six outcome measures. The six outcomes[3] employed are the Total GCSE Score Total GCSE/GNVQ Score, Best 5 GCSE Score and GCSE scores for the core individual subjects: English, Maths and Science. For each outcome, overall value added scores are fed back or all pupils in the school, but also value added scores for three ability groups: high achievers, average achievers and low achievers at entry to secondary school[4]. In addition to the scores for a single year (e.g. 2000 results only), a 'rolling average' score was introduced in 1995, creating an additional set of value added scores from three years of data rather than one (e.g. combined results for 1998–2000).

Multilevel modeling is the method of statistical analysis employed to calculate the value added scores (See Goldstein 1995 for a detailed explanation of the methodology). This technique is regarded as the most powerful and appropriate methodology to adopt. It deals with pupil level attainment rather than aggregated school level data, and gives a measure of 'relative' pupil progress in comparison with other schools—after controlling for intake in terms of previous attainment and other background factors such as gender and entitlement to free school meals (a measure of low family income). This means that schools are able to examine the 'relative' progress of their pupils in terms of value added scores, which in some cases may be very different to raw attainment.

[1] Originally located at the London Institute of Education but from January 2001 located at the Graduate School of Education, University of Bristol.

[2] National Consortium Examination Results (NCER) is an organization responsible for the collection of GCSE, GCE and other examination results from almost all the different examination boards.

[3] Grades for GCSE examinations were re-coded to numerical scores: A*=8, A=7, B=6, C=5, D=4, E=3, F=2, G=1, U=0, X=0, Q=0. General National Vocation Qualification (GNVQ) standard equivalencies to GCSE scores are employed. Total GCSE Score includes all GCSE grades summed. Total GCSE/GNVQ Score includes all GCSE grades and GNVQ awards summed. Best 5 GCSE Score includes only the best 5 GCSE grades of each pupil summed.

[4] The grouping of pupils into achievement bands was based on the UK distribution of individual sub-tests for the NFER Cognitive Abilities Test (CAT) scores. Using an average CAT score across all three sub-tests (verbal, quantitative, non-verbal), Band 1 pupils represents approximately the top 25% of the UK population, Band 2 the middle 50%, and Band 3 the bottom 25%.

Table 17.2 displays all the intake and background variables that are taken into account when calculating the value added scores. This approach was developed in the first year of the project after considerable exploration of the data to identify the best value added model for the purpose of informing school improvement processes (see Thomas & Mortimore, 1996). However, the relevance and statistical significance of the model and each explanatory variable is checked on an annual basis before the school feedback results are prepared.

Table 17.2 Variables Accounted for When Creating the Value Added Scores.

Prior attainment variables	Pupil background variables
Verbal CAT score	Gender
Quantitative CAT score	Age
Non Verbal CAT score	Entitlement to Free School Meals
	Ethnicity
	Mobility (1) Years in UK education
	Mobility (2) Number of secondary schools attended

The actual format and presentation of feedback is prepared by the LEA including the value added scores and raw scores, along with the score ranks (shown in Appendix 1). Value added scores that are statistically significantly (at 0.05 level) below or above expectation are denoted by an asterix symbol. All other scores are not statistically significant and indicate that the school is performing as expected. At the request of Lancashire LEA an additional set of scores for boys and girls was also produced for each school in 1996 and 1997. However, in these cases there was little evidence of differential school effects according to gender.

The value added data is provided annually in the autumn term in draft form (and subsequently ratified after extensive checks and feedback from schools). Support in the process of using the value added scores is ongoing. A series of LEA seminars on Value Added are held for schools every year, as well as separate sessions for school governors. In addition, LEA in service training often uses the value added project for data analysis training.

(ii) NCER GCSE Subject Differences Analysis

This feedback involves a technique which allows for comparisons between subjects by making an assumption of differences in subject difficulty. By looking at all GCSE subjects for all pupils in England, the NCER have identified a level of difficulty for each GCSE subject. This is represented by a ‘residual’; positive representing a subject easier

than the average subject, negative representing a more difficult subject. The level of subject difficulty for the whole of England is compared to the residual for a single school, to create a final 'net residual', the difference between schools residual and the England residual (See Appendix 2 for an example of the feed back received).

The technique is designed to identify particular departments or subjects within a school that are over or underachieving, and claims to 'compensate for variations in pupil ability or different attainment of pupils resulting from socioeconomic factors' (NCER 1996), as a schools subject residuals are created by comparing with the average for that school only. However, this approach is sometimes viewed as problematic because of the fact that not all pupils are entered for all subjects either nationally or within a single school. This data is provided to schools annually in the autumn term.

(iii) GCSE Subject Analysis Aids

Schools are also provided with annual feedback information in the autumn term based on previous years data. For example, in 1998 schools received the following feed back for GCSE outcomes: 1) the average GCSE points gained by different prior attainment groups (prior attainment split to approximately 30 separate groups), for *all* GCSE subjects examined. This information is prepared in-house by the LEA and given in the form of tables for all pupils jointly, and also for boys and girls separately, 2) schools are provided with distribution tables showing the percentage of pupils gaining each grade, for each level of prior attainment (split by stanine) for boys and girls separately, 3) graphs for all GCSE subjects showing a prior attainment to GCSE grade line for girls and boys separately, 4) projected grades for each prior attainment (split by stanine and full 30 separate groups) for each GCSE subject. Feed back for National Curriculum Key stage 3 outcomes (at age 14 years) follows a similar structure.

(iv) Pupil and Teacher Attitude Questionnaires

Annually, from 1996, all Lancashire secondary pupils aged 14 years (Year 9) and 16 years (Year 11) are asked to complete a 42 item questionnaire in the spring term covering various aspects of school life. This questionnaire was originally developed for the Improving School Effectiveness Project in Scottish schools (see Thomas 1998, 2001, Smees & Thomas 1998). The individual item results of the questionnaire are fed back to schools by the LEA in the autumn term, split by gender, ethnicity and ability bands for each year as well as the overall results for each year.

In addition, as part of the Lancashire Value Added Research project a LISREL factor analysis[5] of the data is carried out, from which 5 factor scales are created: Engagement with school, Pupil Culture, Self Efficacy, Behaviour and Teacher Support (see Appendix C). Thus, as well as the individual questionnaire item results, mean school scores for each attitude scale are calculated and fed back to schools annually, along with county averages for the scales.

[5] LISREL follows the same principles as factor analysis, in that it attempts to pull underlying factors from the data. However this approach does use different methods and can cope with an assumption of non-normality of the data.

From 2000, the project will be able to track pupil attitudes from age 14 (Year 9) to age 16 (Year 11) to examine the change in pupil attitudes over a two year period. If appropriate, value added scores will subsequently be developed to reflect the relative influence of schools on pupil attitudes (i.e. using the same approach as with GCSE value added outcomes).

In Spring 2001 a teacher questionnaire—originally developed for the Scottish Improving School Effectiveness Project (see MacBeath & Mortimore, 2001 for details)—was also be administered in Lancashire secondary schools to provide an additional source of feedback information to schools and teachers.

17.2.4 Use of the evaluation information within schools

Lancashire schools are still in the process of learning in terms of self evaluation activities, and although a lot of the schools use the data, there is a small minority that do not understand fully what the scores mean. However, the LEA project manager reported:

> *"the vast majority of schools can have an intelligent conversation on this data now"*
>
> (LEA adviser and project manager)

The introduction of Value Added scores for different ability bands has led to a closer attention to differing needs of different pupil groups within the school. Many schools have adopted able pupil policies, streaming of pupils, and changes in the curriculum to fit with the different ability groups. The profile of systematic individual pupil monitoring has also increased as a consequence of the Value Added Project. Assessment systems, setting targets for pupils, and whole school approaches such as calling in school books have been set up to:

> *"…monitor progress and attainment, to monitor the way in which the assessment of pupils is being carried out and how that assessment information is being used to plan future learning strategies for individual improvement"*
>
> (LEA adviser and project manager)

The project also seems to be having an effect on schools reflecting on the quality of teaching and learning within the school. Advisers are working with heads of

department, heads of faculty, and senior managers to look at strategies to develop monitoring of teaching and learning quality, an application that the LEA hope to develop much more in the future, when they hope to network similar departments with differing attainment success:

> *"It's all about opening up schools for a much more detailed and critical analysis about what is the best practice we can find, what is it that is*

successful and how can we actually improve and disseminate improvement strategies and ideas"
(LEA adviser and project manager)

In terms of evaluating improvement, schools use both the value added and raw GCSE results to carefully examine trends in pupils academic performance over the five years of the project. The *'3 year rolling average'* results are a particularly useful aid to valid interpretations of improvement as year to year fluctuations in results are smoothed out. Schools have also employed the pupil questionnaire results related to different aspects of school culture, such as bullying and behavior, to investigate the impact of current policy and practice as well as new initiatives:

"Things like anti-bullying policies, behavior policy, they love things like that [i.e. as reflected in the questionnaire item results] to see whether or not they have got that right in school"
(LEA adviser and project manager)

In 1997 the LEA carried out their own evaluation of how schools intended to use all the different types of information provided by the Lancashire VAP and from this information an on-going database was created for use by LEA advisers. The responses from teachers to this evaluation largely support the comments reported above and a summary of their responses can be found in Appendix D.

17.2.5 The impact of the value added project

The beginning of the Value Added Project corresponded with the time when the local education authorities were vulnerable to the pressure of schools leaving the authority to go grant maintained. Lancashire LEA are sure that one of the reasons they managed to keep schools within the authority was the Value Added Project[6]. The effect of the project on schools and increasing awareness of school effectiveness and improvement is quite fundamental, the focus of improvement being on all pupils in the school not just selective groups:

"There is no Lancashire secondary school now where we can't go in and ask a teacher how are you improving the quality of learning for the pupils in your class and what information are you collecting and using and analyzing to help

you in that process. There is no teacher in Lancashire that wouldn't understand that was a legitimate source of inquiry, even if they were at very different levels of understanding of what they would mean and that is very much down to this project…"

(LEA Adviser) [6] A survey carried out by Riley et al. (1998) Of secondary heads found that the value added project was one of the main reason they stayed with the authority.

17.2.6 Relationship of the project to national education policy

When asked how the Lancashire LEA Value Added Project fits in with national policy, LEA managers were keen to point out the level of sophistication of the Lancashire value added methods in comparison to any initiatives at the national level. In terms of the current national policy of setting pupil attainment targets for each LEA, the Lancashire Value Added Project feeds into the whole process by enabling fairer school targets to be set:

> *"What we are able to do is to take into account measures of prior attainment, we are able to take into account trends when all background factors using multilevel analysis, take those factors into account and as a results of that start to set quite reasonable targets for schools"*
> (LEA adviser and project manager)

The school targets set by the LEA are based on intake (prior attainment and background factors) and previous performance in terms of both raw and value added scores, to give schools the space to gradually improve their results within a realistic context:

> "*we take into account past performance and if someone is a high performing school we want them to maintain that. For those who haven't obtained at the same level in the past we said 'look we want you to start to get up to this higher level'"*
> (LEA adviser and project manager)

LEA advisers also see the proposals for providing value added data on a national basis as far too simplistic in that:

> "*it doesn 'take into account all the variables you want it to take into account and it doesn 'take into account all the outcomes that you would want, so it won't necessarily give a true picture of what is actually happening in the school"*
> (LEA adviser and project manager)

Due to these pitfalls, the LEA has misgivings about the national value added policy developments, as it may threaten the confidentiality of the Lancashire Value Added scores if some schools feel they need to defend themselves by disclosing the results to the public.

17.3 The Case Study of Self Evaluation Activities in One School

For the case study one school involved in the Lancashire LEA value added project was selected, with the advice of LEA managers, on the basis of illustrating good practice in school self evaluation activities. The case study evidence was collected over a two day period (Summer term 1998) and involved separate interviews with the head teacher, deputy head (also responsible for disseminating the LEA value added results and other data analysis in the school), heads of department, classroom teachers, pupils, parents and governors.

17.3.1 The context and history of the school

The school has recently changed status from a 11–16 comprehensive school to become a Technology College[7], and is situated within a mixed economic setting of state funded council housing accommodation and middle class suburb. The school has 58 teachers, and seven form groups per year. Years 7–9 are split into nine separate classes for the core subjects of English, Maths and Science, but the number of classes fluctuates for years 10 and 11 when key stage 4 courses begin. In total 16.7% of pupils are entitled to Free School Meals, slightly lower than the national average (national average 18.2 %), and 0.4% have English as a second language. The number of pupils with special needs (including statements) is slightly above the national average at 18.7% (national average 16.6%). Table 17.3 displays additional contextual information drawn from the 1991 national census data for the local area surrounding the school and national averages for comparison[8].

By the late 1980s the school was heading towards closure due largely to disappointing examination results, and pupil numbers had began to decline. Within this context, the schools senior management team (including a newly appointed head teacher) decided that a new emphasis on achievement was required to turn the school around, so that they would be able to compete with other secondary schools in the new climate of parental choice. This commitment led to the head, and in particular the deputy head, engaging quickly and positively with the developments in value added analysis initiated by the LEA.

Table 17.3 School Census Data.

	National Average	School Local Area
Adults with higher education	13.5	8.1
Children in high social class households	31.0	23.1
Minority ethnic children	10.1	0.5
Children in overcrowded households	10.5	10.7

[7] Technology Colleges receive additional funding for Science.

[8] This data is drawn from the data provided in the 1997 Performance AND Assessment report (PANDA).

17.3.2 Development of self evaluation activities in the school

The self evaluation program was originally conceptualized as a tool to facilitate the turn around of the school:

> *"We saw it very much in the beginning as a way of addressing the sorts of problems that we face and I think we learnt very quickly that there were some solutions in there for us"*
> (Head teacher)

Evaluation data helped the school to understand the problems they faced, and most importantly to begin to grapple with issues in a completely new way. It was this self empowerment that was the major push towards improvement:

> *"I think it was really fairly quickly that we began to realize what a powerful tool that we had got. For first of all, I suppose selfishly, I was their head and for me as a management tool it was the one thing I would never give up, because it has enabled me to understand so many things better than I have understood before, and to deal with them, I think, in a more dynamic and positive way"*
> (Head teacher)

However, the program did not materialize overnight, and it was soon realized that the initial work they had embarked on, looking at the results of previous years of pupils, although useful to the school, did not directly help the pupils presently attending:

> *"After two or three years it told us how we were doing, but it didn't say [how] we could do anything because they [the pupils] have gone. So we had to look at what we could do in the school before we get to this point"* [9]
> (Deputy head and project manager)

This led in time to the re-conceptualization of the program within the school to include systematic individual monitoring of all pupils in the school as part of the self evaluation process, via predicted grades, on-going assessments and the use of interim reports (see later section on evaluation instruments).

[9] This comment refers to the fact that the value added data provided by the LEA project is retrospective in nature and relates to previous pupil cohorts attending the school.

17.3.3 Management of the program with the school

The overall management and dissemination of the LEA value added information and further data analysis is led by the deputy head, a member of the senior management team (SMT).

There is a strict top down structure to the self evaluation program, whereby evaluation data is only fed down the management structure that is felt to be

appropriate or necessary. The SMT were aware that the evaluation data would only be used by teachers when it was directly linked to classroom practice.

The structure and use of evaluation data has also helped focus the dialogue between the SMT and 'middle managers', to concentrate on a common goal:

> *"It has enabled us to create a very useful dynamic between middle managers and the new management team because we talk the same language. You haven't attended any of our meetings, but you would find that there is very little discussion about anything other than the most effective way of using data, to use systems better, and to take steps forward that way. I can't remember who said it, but someone said it, that you know how bad things are when there is only talk of control measures and discipline. And things like that are not really part of our conservation, our conservation is about children, evaluating what we are doing, and taking it forward that way. So I think you would find the quality of discussion at this school is very high and it is due to the project and how we have been helped by it"*
>
> (Head teacher)

17.3.4 School self evaluation instruments and feedback to teachers

The school receives all the data produced by the overall LEA value added project. In addition to this a vast amount of material and information to use within the self evaluation system is also prepared internally by the school. This additional information includes:

(i) Graphical Feedback

The school produces a number of graphs tracking the pupils in their own school compared to all pupils in the LEA, which they split by gender and ability groups, tracking the GCSE results for different subjects over a number of years.

(ii) Pupil Predicted Grades

The school uses the predicted grade information from the LEA to produce a residual for each class on the basis of the previous years GCSE performance. This consists of a mean residual taken from the individual pupil residuals in the class (the distance the pupils real grade was from the expected grade).

(iii) School Specific Questionnaires

The views of both pupils and parents are collected via the regular design and administration of questionnaires for the two groups, covering current issues and ongoing themes such as pupil satisfaction.

(iv) Pupil Attendance

The school also monitors attendance at school by both teachers and pupils and attendance of parents at parents evening. Targets are set for all of these.

17.3.5 How information is used and fed back within the schools

(i) Pupil level evaluation

The last HM inspection team that visited the school stated that there was very few children that could get *'through the net'* at the school, highlighting the heart of the evaluation program: systematic pupil monitoring. We have already mentioned the types of data that the school utilizes for pupil monitoring, but not the extent to which it is used within classroom practice.

The most valuable tools the school feels they have are the CAT score information and the predicted grades reported in each student's interim reports, used mainly for pupil target setting for GCSE (also some use of CAT to look at KS3 performance). The school first began using a CAT score to GCSE graph. In that first year 49% of pupils were falling below the 'expected score' line, compared to approximately 20% in 1998.

An integral part of the evaluation and monitoring process is the interim reporting system. The interim report is an additional report to the traditional end of term report, designed to inform both teaching staff and pupils of the progress pupils are making, especially to identify pupils who are falling behind their expected progress (see Appendix E for example of the interim report). Shortly after the interim report is sent to parents, pupils and their form tutors have an 'individual review', where they discuss their results.

In the final year of statutory schooling any students aged 16 (Year 11) who are identified as seriously under-achieving are put on a mentoring scheme. Approximately 5–10% of the year group are mentored each year, where mainly pastoral staff are allocated a pupil whom they see on a regular basis. It is their responsibility to monitor the pupil's

progress, check the course work and generally make sure that are keeping up to date with their GCSE course.

(ii) Classroom evaluation

Departments and class teachers also look at pupil residuals from previous years to see which pupils were under or over achieving. By looking at the 'average class residuals' teachers assess whether there are any classes that have been particularly successful, and to attempt to pinpoint a reason for this:

> *"We can see which ones slipped up and we can also summarize it together. I did something like this for the head and we looked at average residual for each class as well. The averages for all of those, so we could see which classes had done well and, for instance, sets for 1998, set X who happened to be an all boys set… they did really well and so we started to think now maybe we should be teaching in single sex groups in certain cases…there has been a lot of talk about it at the moment and this get's us thinking"*
>
> (Department Head)

The pupil level evaluation data also influences the planning of classroom practices. The data they have on entry year attainment allows the teachers to tailor the teaching approach to the particular children they have in the class:

> *"Well I would make decisions about the kind of work that I am doing and the way I'm going about revision with the KS4 class based on what I know of them. I've got a very weak set at the moment with a CAT score in the low 90s, so my whole approach to teaching them is a very different approach from the approach that I took last year with a history group whose average CAT score was 104, a very bright able group. So I'm looking at this in different ways with my current Year 11 group and that's based on the data that I have here"*
>
> (Department Head)

Such data also gives the teachers a guide to the reasons a child might not be achieving, leading to a different teaching strategies for different pupils within the class. Such knowledge can help teachers manage the classroom more effectively, and enable them to keep a closer eye on pupils who may need more intensive help, as in the following example about particular pupils:

> *"For instance I've got him now at the front of my classroom; him and a couple of others, that are weak. They're right at the very front, so whenever the rest of the class are working I can check that they actually*

understand what I am doing. They don't realize that but it just helps me out, you know who the weak ones are"
(Class teacher)

(iii) School level evaluation

Evaluation of whole school and departmental results is also an integral part of the self evaluation program. Information from the Value Added Project produced by the LEA and the Institute of Education is discussed and assessed in individual meetings between heads of faculty, head and deputy head. The heads of faculty get to see value added data for their own subject and the overall value added data for the whole school only. This includes the graphs and the value added scores. They are only shown the value added scores during the meeting but not permitted to take the data away with them for confidentiality reasons.

The staff at the school are very much focused on the pupils in their own classes and looking internally for answers, rather than outside the school. They felt the school, any school or class is unique, so the best place to begin looking for solutions is within the school itself. Part of the coming years agenda is for the heads of faculty to look at how they can 'pick out' best practice from the different departments in the coming year, which would work alongside the existing system of classroom observation. At present each teacher is allowed covered time each term to observe a class of their choice, often pairing with either a highly effective subject or one closely related to their own. This helps towards both dissemination of best practice:

"We are all doing little bits of a good job but if you can sort of mould that and pick out the best bits and all do that, then that would give overall exam success"
(Department Head)

Teachers were clearly willing to take responsibility for the welfare of their own students and their performance. As one teacher asked:

"How I can make a difference? How can I put things right if I've made a mistake or done something wrong?"
(Class teacher)

17.3.6 Internal and external support for the program

Support is given by the LEA, who regularly visit the school to work through the data they receive. The school felt they took a proactive approach to the support they received by actively seeking out this help. Internally, the school has a very strong training ethos. All new staff are given training in data analysis, the deputy head regularly puts aside time in

meetings to explain particular data, and In-Service (INSET) time is also regularly used for evaluation training.

17.3.7 Attitude of the school to self evaluation activities: a multiple perspective on school self evaluation activities—the voices of the teacher, governor, parent and pupil

(i) The teacher's perspective

From the comments noted above there is an overwhelmingly positive attitude among teachers towards self evaluation within the school. This is due in some part to the recruitment of staff that have congruent views towards self evaluation, but also to the benefits the staff are experiencing of using the data.

(ii) The governor 's perspective

The governors' main focus of school evaluation was academic progress from Year 7 (in terms of the 'Cognitive Abilities Test', the year of entry to secondary school, to Year 11 (in terms of Total GCSE attainment). The chart plotting the two scores for each individual pupil was the key piece of evaluation data used by the governors to appraise school quality. He stressed the benefits of this data in terms of accountability and was keen to point out the goal of the school was not to achieve in line with expectation (which would be a pupil lying on the regression line), but to achieve above expectation for each pupil, and evidence of this would be clearly visible on the plot.

> *"It does several things. It effectively judges the quality of teaching at a top level, because if the teachers are just sitting back, letting them go along, they'd be on the line, and any monitoring you do, checking the quality of the teaching, is giving them as much as they can. The whole system within the school is to add as much as they can to everybody, and it's because of the assessments and the way it's published, it's there for everyone to see: the parents, the pupils, the teachers, right across the board"*
>
> (School Governor)

Clearly the data used by governors concentrated on GCSE academic attainment, in terms of progress from Year 7 to Year 11, but the governor didn't feel that focusing on a single academic outcome for accountability purposes was problematic because the school did strongly value non-academic areas that were not covered in any explicit self-evaluation exercise such as vocational qualifications, community work and extra-curricular activities, and was keen to point out:

> *"your not just feeding the academic side, your feeding the whole of a student".*
>
> (School Governor)

In fact there was a concern about how such areas could be accurately measured. He did feel that attendance was a good quasi indicator of children's enjoyment of school that was already being monitored:

> *"If they don't like school, if they're bored or whatever, they won't come. So if attendance rates are higher than the other schools around you can't be getting it all wrong. And that applies to teachers as well, they keep figures on levels of attendance of teachers as well as pupils"*
> (school Governor)

(iii) The parent's perspective[10]

The parent's main focus was also on Year 7 (age 11) entry attainment but was more finely focused on how the school was using this data to push children of all abilities to reach their academic potential via individual monitoring of pupil progress (e.g. annual feedback meeting class tutor) and interim reports. Contextualizing this in terms of her own children:

> *"My youngest sons very bright but he has a problem getting things down on paper…. I can only say that they've been very good with him. He doesn't find things as big a problem now as he used to, his work has improved a lot. I don't know whether it is through all theses evaluations and value added, I couldn't prove that one way or another. But for him, whatever it is it's working and I can only speak as a parent on that, and the approach is very positive"*
> (Parent)

Additionally, it was pointed out that parents were kept well informed of the progress of their children, especially through the two reports a year sent home. A key element to the process of pupil monitoring was the positive attitude the school and staff possessed, so even if a child was struggling, the way they addressed the problem was to stress the positive aspects.

[10] The parent interviewed also worked in the school as part of the ancillary office staff and as an consequence saw school evaluation information as part of her work, therefore she is not a 'typical' parent.

(iv) The pupils' perspective[11]

The pupils were very aware of the monitoring and evaluation of their own work that goes on in the school through CAT tests, classroom tests, Key stage and GCSE module assessments and homework. They knew that CAT scores from Year 7 (age 11) were used as a guide[12], both to stream them at the beginning of school and to check their progress was at the level expected. Improvement to them could be measured through grade increase and written comments from teachers.

> *"Because they know how many levels you should be moving up they can see if you are improving, by seeing the grades from your CATs and grades from your SATs[13]"*
> (Year 10 Pupil)

One of the key elements of the evaluation process for pupils was the interim report, as this gave them an opportunity to do something about the results if they weren't as good as expected, or were a confidence boost if they were doing well.

> *"Because it used to be you just had one [a report] at the end so if you 're not done very well you can't do anything about it during the year, so if you have one halfway through the year, which is just grades not a written report, you know how your getting on"*
> (Year 10 Pupil)

It also gave them a chance to voice their own opinions about their work via the personal statement, although they felt they were unlikely to disagree with the teachers assessment. Older pupils were keen for the individual reviews to be with individual subject teachers rather than their form tutor.

A strong theme that came out of the interviews with pupils was the feeling that they all felt it was possible to be successful. Flexibility within the system allowed pupils at least twice yearly opportunities to move up to a higher set if progress was above expectation, so pupils did not appear to feel constricted to a particular attainment target. Although pupils weren't directly aware of the self evaluation that the school engages in, older pupils were aware of some comparisons between their own schools attainment results and the national picture. For example, Year 10 students (age 15) had been told by many of their subject teachers at the beginning of Key Stage 4 that the schools was performing above the national average, and had seen charts displaying the same information outside their classrooms, which had been strategically placed for pupils to look at while waiting for classes to start.

[11]Pupils from years 8–10 (aged 12–15) were interviewed.
[12] Pupils were aware that test scores alone were not a valid indicator of what a pupil could achieve. They felt classroom performance and attitudes to work were also important, and that test scores can be an imperfect measure as some people may not be as good at tests as class work, or people may have had different amounts of preparation for the test.
[13] SAT stands for Standard Assessment Task, which are part of the National Curriculum Key stage assessments. For example at Key Stage 3 these would be the assessment tests.

17.3.8 Wider impact of the school self evaluation activities: measuring what we value or valuing what we measure?

Most staff interviewed recognized a concern with school effectiveness indicators that the outcomes measured within education concentrates heavily on academic outcomes, at the expense of other equally valuable educational indicators. However, it was felt that a wide range of outcomes including citizenship and sports achievement are celebrated at the school and that it was inappropriate and unnecessary to measure them quantitatively:

> *"There are so many other things that are not easily measurable that we know we do because we work here, we live here we know the children and only people in the school can do that. You can't always measure it, it comes from experience, it comes from feelings"*
> (Deputy Head and project manager)

School staff are beginning to use the student questionnaire data to check on certain key issues such as truancy, homework and student-teacher relationships. The results from these items has led to direct changes in school policy such as calling in school books and closer checks on absence notes. At present work on student questionnaire data is carried out mainly by the pastoral staff and the SMT, who pass onto faculty heads results that they feel are particularly relevant. The faculty heads are then asked to assess the results and report back their conclusions, including what areas they feel needs to be addressed from the results. They feel the questionnaire helps to keep certain important issues in the forefront of peoples minds:

> *"After things have been going along for a while you can become relaxed about that, you need to check up on them. It serves as a reminder"*
> (Department Head)

However, classroom teachers have yet to have any real contact with the questionnaire data, and are a little sceptical of the validity of the answers students give. Their channeled focus on classroom practice has also led to such external data not to be utilized.

17.4 A Summary of Good Practice

The selected case study school illustrates current good practice in self evaluation activities where evaluation data is incorporated into all levels of school practice from classroom teaching to staff appraisal. Pulling together the experiences of teaching staff, SMT, parents, governors and students, the following self evaluation strategies seem to be particularly effective at the school:

- An intensive monitoring system, that tracks students throughout their school career, and has within it a program of intervention if students are falling behind.
- A clear, shared focus on high expectations for all students in the school, achieved through positive reinforcement.
- A realistic awareness that evaluation data is only an aid to teaching and learning, and as such is only useful when linked to other information available about students, classes and whole school structure.
- The importance of including multiple perspectives in the school self evaluation process.
- Being realistic about their own capacity for using the data. This includes providing adequate technical support for staff and knowing the boundaries of their capacity in terms of resources.

The school is still engaged in the process of learning how to interpret the data and grapple with the issues that arise from it, and each year builds upon the structures and skills they

have already developed. The next step for them is to find a way of pulling together all the data they receive, produced internally and use, to make it more user friendly and accessible to all the staff in the school.

Appendix A— School VA results table (fictional school)

VALUE ADDED PROJECT 1997
Draft School Value Added and Raw GCSE Results

	1997 TOTAL GCSE BASIC SCORE (N=98 schools)				
	Basic Score	Raw Total GCSE Score	Band 1	Band 2	Band 3
	-6.23*	12.03	-4.24*	-6.43*	-5.22*
Rank	95	93	92	96	95

	TOTAL GCSE BASIC SCORE OVER3 YEARS (N=98 schools)				
	Basic Score	n/a	Band 1	Band 2	Band 3
	-6.72*		-4.24*	-6.43*	-5.22*
Rank	95		92	96	95

	1997 BEST 5 GCSE BASIC SCORE (N=98 schools)				
	Basic Score	Raw Best 5 GCSE Score	Band 1	Band 2	Band 3
	-2.65*	10.56	-3.45*	-2.43*	-4.26*
Rank	92	93	92	89	95

	BEST 5 GCSE RESULTS OVER3 YEARS (N=98 schools)				
	Basic Score	n/a	Band 1	Band 2	Band 3
	-2.57		-4.24*	-2.45*	-3.56*
Rank	93		95	87	95

	1997 ENGLISH GCSE SCORE (N=98 schools)				
	Basic Score	Raw English GCSE Score	Band 1	Band 2	Band 3
	0.21	3.45	0.45*	0.12	-0.05
Rank	45	56	23	53	60

	ENGLISH GCSE BASIC SCORE OVER 3 YEARS (N=98 schools)				
	Basic Score	n/a	Band 1	Band 2	Band 3
	0.34*		0.46*	0.22	0.10
Rank	28		19	34	39

	1997 MATHS GCSE SCORE (N=98 schools)				
	Basic Score	Raw Maths GCSE Score	Band 1	Band 2	Band 3
	-0.33*	1.55	-0.34*	-0.55*	-0.22
Rank nk	82	98	84	91	73

	MATHS GCSE BASIC SCORE OVER 3 YEARS (N=98 schools)				
	Basic Score	n/a	Band 1	Band 2	Band 3
	-0.55*		-0.51*	-0.67*	-0.34*
Rank	97		95	98	91

	1997 SCIENCE GCSE SCORE (N=98 schools)				
	Basic Score	Raw Science GCSE Score	Band 1	Band 2	Band 3
	-0.12	3.22	-0.21	-0.15	-0.08
Rank	59	38	60	62	55

	SCIENCE GCSE BASIC SCORE OVER 3 YEARS (N=98 schools)				
	Basic Score	n/a	Band 1	Band 2	Band 3
	-0.02		0.01	-0.04	-0.09
Rank	51		49	55	56

Notes: 3 years = 1995 – 1997; * = significant score at 0.05 level, # = less than 5 pupils; Band 1 - Highest ability (NFER Stanines 9-7) Band 2 (Average ability (NFER Stanines 6-4) Band 3 - Below average ability (NFER Stanines 3-1).

Appendix B— NCER subject differences example (example from 1996)

A School Subject Residual is the average of the individual pupil's residuals for the subject. This is arrived at by adding together the pupil residuals for the subject and by dividing by the number of pupils entered for the subject by the school. The same process has been completed to arrive at the LEA Area Subject residual and the England residual for the subject.

The example below shows the overall England subject residuals.

1996 SUBJECT DIFFERENCE ANALYSIS	YEAR 11
ENGLAND	ALL PUPILS

30 MOST POPULAR SUBJECTS

Subject	Pupils	Residual
CDT:DES&COMM	22199	-0.47
D/T/ELECTRONICS	10989	-0.42
SPANISH	33942	-0.38
GERMAN	126377	-0.33
FRENCH	315981	-0.29
MATHEMATICS	539256	-0.28
RELIG.STUD	99443	-0.27
INFORM.SYSTMS	45508	-0.25
BUS.STUDS.SNGLE	102150	-0.22
HUMANITIES:SGL	12746	-0.21
HISTORY	212643	-0.20
CHEMISTRY	36361	-0.12
GEOGRAPHY	269840	-0.11
SCI:SINGLE	63028	-0.11
DES&TECH	159481	-0.08
BIOLOGY	37207	-0.08
PHYSICS	35838	-0.06
MUSIC	38144	-0.01
CDT:DES&RES	43330	0.00
HE:FOOD	41022	0.04

SCI:DOUBLE	435037	0.06
MEDIA/FILM/TV	17283	0.11
SOCIOLOGY	11481	0.11
INT.HUM.SNGLE	31412	0.17
HE:CHILD DEV	38259	0.19
ENG.LIT.	455463	0.20
EXPRESSIVE ARTS	14134	0.23
SPORT/PE STUDS	71512	0.24
ENGLISH	540138	0.27
ART&DESIGN	146104	0.45

Appendix C— Pupil attitude scale details

Table A-C. 1 Items on each attitude scale

Factor 1:	**ENGAGEMENT WITH SCHOOL**
Item 1	I always like school
Item 3	I always get on well with teachers
Item 5	Teachers are always fair
Item 6	School work is always interesting
Item 31	Teachers are nearly always friendly towards pupils
Factor 2:	**PUPIL CULTURE**
Item 2	I always get on well with others in my year
Item 30	I never feel left out of things
Item 33	I never get bullied
Item 36	I find it easy to make friends
Factor 3:	**SELF EFFICACY**
Item 26	My work in class is very good
Item 28	I think I'm very clever
Item 29	All my teachers think my work in class is good
Factor 4:	**BEHAVIOUR**
Item 3	I always get on well with teachers
Item 37	How would you describe your behavior in class? Good
Item 38	How do you think teachers would describe your behavior? Good
Factor 5:	**TEACHER SUPPORT**
Item 8	Teachers always help me to understand my work
Item 11	Teachers always tell me I can do well
Item 14	Teachers always tell me how I am getting on with work
Item 16	Teachers always praise me when I have worked hard
Item 31	Teachers are nearly always friendly towards pupils

Creating Pupil Attitude Scales using LISREL: Methodological and Technical Details

Pupil responses to items on the secondary pupil attitude questionnaire were fed into a LISREL analysis (Linear Structural Equation Model for Latent variables, see Joreskog & Sorbom, 1989) to identify and, if possible, measure any underlying attitudinal dimensions. LISREL is a relatively new and highly sophisticated approach to the analysis of categorical data and the creation of latent variables (for example, attitude scales). This methodology has several advantages in comparison to other methods such as Principal

Component Factor Analysis. First, LISREL handles correctly skewed (or non-normal) distributions of subject responses for particular items and second it provides the most accurate estimates of scale reliability (see Rowe, 1996).

The reliability of the scales and the individual items have been tested on the Lancashire sample schools and an equivalent secondary sample from the Improving School Effectiveness Project. The composite scale reliability co-efficient[14] of each scale was calculated in the LISREL analysis on two consecutive cohorts in 1996 and 1997 for the Lancashire data (including both Year 9 and Year 11 pupils), producing test-retest reliability estimates ranging from 0.73–0.92 for the Lancashire sample and 0.76–0.93 for the ISEP secondary sample (see Thomas et al 2001). In addition, an actual test-retest analysis was carried out with 153 Lancashire Year 9 pupils in 1996, yielding correlations of 0.42–0.73 for the secondary pupil attitude scales[15], and 0.12–0.69 for individual items (the average test-retest correlations across all items was 0.47). This evidence supports the results from the LISREL analyses, as none of the items with low test-retest correlations were subsequently found to be robust enough to include in the LISREL scales.

[14] Rowe (1996) has argued that this statistic is an improved estimate of test-retest reliability in comparison to the internal consistency estimate provided by Cronbach's (1951) standardized item *alpha*.

[15] The actual test-retest correlations for the Lancashire secondary scales were: Engagement 0.73, Pupil Culture 0.42, Self Efficacy 0.64, Behaviour 0.63 and Teacher Support 0.60. The equivalent 1996 LISREL estimates of test-retest reliability (employing a combined Year 9/11 sample n=17000) were: Engagement 0.75, Pupil Culture 0.79, Self Efficacy 0.77, Behaviour 0.92, Teacher Support 0.77.

Appendix D— Summary of LEA database of how schools intend to use the data

Each school is asked to summarize, on up to 2 pages, how the following assessment and related data is being analyzed in the school: Multi-level—GCSE, Analysis Aids: GCSE subject tables and graphs, Key Stage 3 subject tables and graphs, NCER pupil-referenced data, Pupil questionnaire raw and adjusted data, anything else.

The Multilevel GCSE results are used mainly by the Senior Management Team, who pass them onto the individual departments concerned, although a few schools provide all the results to all staff. The results are discussed in curriculum and head of department meetings, and in quite a few cases feed them back to governors. The results are used for more generally for target setting and future planning of school improvement initiatives. Issues looked at include differential effects across departments and different ability groups, looking at three year trends and identifying weaknesses.

The GCSE subject tables and graphs are also used by the Senior Management Teams who pass them onto the heads of department. The biggest use of this data is to set targets through prediction of grades, a system used by most schools to some extent. More sophisticated analyses of the data include looking at gender differences, trends over years, yearly improvement in results, the performance of particular sets/classes and comparing school results to the county results. These results have also been used as the basis of INSET days.

At this stage many of the schools reported honestly that they have not yet started to use the Key Stage 3 data, but many state they will be looking at it in future years. There is some strong concerns still for the validity and the reliability of the Key Stage 3 assessment, especially in English. Where the data is used it is passed onto Heads of Department by the Senior Management Team, to use as they wish. Some schools have begun relatively crude target setting and prediction to GCSE, and also use it as a tool to identify underachievement, by looking at progress from Key Stage 2.

The NCER data isn't used by all the schools but those that do again pass it onto the Heads of Department and sometimes to other staff. A few schools stated they found this data confusing. Where the data is used it aids the identification of particular subject strengths and weaknesses, by giving what schools feel is a more realistic view of subject results. Trends over time have been used as the basis to inform individual subject initiatives such as revision courses.

The pupil questionnaire data is found very interesting by a lot of schools, although quite a few still haven't got round to utilizing the data fully or comprehensively yet. The data is fed back to different people in different schools such as Heads of department, year heads, pastoral staff and in some cases all staff and governors. Most schools use the data in a more general sense, for general interest, although a few have special meetings solely to discuss what areas require action or are cause for concern. Issues addressed cover identifying strengths and weaknesses, looking at the differences between LEA results and schools' own results, the differences between girls and boys results and ability groups and looking at differences between year groups. Other uses include looking at the link

with GCSE attainment (additional information produced by the Institute of Education) and looking at specific areas such as ethos and bullying.

Other data most frequently reportedly used by nearly all schools was the individual pupil CAT scores that are fed back to schools. In nearly all cases all staff in the school received the individual pupils scores. Schools use these scores in the early years for setting classes in years 7 and 8. Mainly in later years they are used for target setting, predicting grades and identifying underachievement. Where underachievement exists schools report often having mentoring schemes which may involve giving those pupils their test results as an incentive. Other data used includes GCSE/GCE/GNVQ data and attendance data being fed back to parents, and SIMS GCSE results for different teaching groups.

Appendix E— Case study school: example of individual student monitoring interim report

(Key Stage four example—Fictitious student)

School X - Interim Report - Key Stage Four

Name:	Form: X	Date: December 1997

This report is an indication of your child's progress at the current time, and is intended to keep you informed before you receive the full profile later in the year.

Subject	Examination target	Current Performance	Effort	Homework/ Coursework	Behavior
English	1	2	2	1	1
Mathematics	2	3	2	3	1
Science	1	2	2	2	2
Religious Education	1	1	1	1	1
Physical Education	1	1	1	1	1
Painting & Drawing	3	3	2	2	2
Spanish	2	2	1	1	1
Textiles	2	2	1	1	1

Attendance	Actual:	Possible:	Percentage:

GCSE BOUNDARY GRADES	VOCATIONAL COURSE
Grade 1 = A*- B	Distinction = D
Grade 2 = B - D	Merit = M
Grade 3 = C - D	Pass = P
Grade 4 = C - F	Fail = F
0 = Not applicable to this subject	

Mathematics only - Boundary Grade 4 includes D - F Grades

Examination Target = refers to the grade that your son/daughter should aim to achieve in relation to his/her ability in this subject.

Current Performance = refers to the grade that your son/daughter could be likely to achieve based upon their present standard of work.

Effort/Behavior =

1 = Examplary - Maintaining a very high standard and deserves commendation

2 = Good - Attaining pleasing standards and deserves praise

3 = Could be better - With more effort might attain a better standard; there is room for improvement

4 = Unacceptable - This is below the standard which we expect from our pupils

---

I/We have this report and agree to monitor and support my/our child in partnership with the school.

Signed:---------- (Parent(s))

References

Achilles, C.M. (1996). Students achieve more in smaller classes. *Educational Leadership, 53*(5), 76–77.

Ackerman, T.A. (1996a). Developments in multidimensional item response theory. *Applied Psychological Measurement 20,* 309–310.

Ackerman, T.A. (1996b). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement 20,* 311–329.

Ackoff, R.L. (1981). *Creating the corporate future*New York: J.Wiley.

ACT, Inc. (1993). *Compass.* (Computer Software). Iowa City, IA: American College Testing.

ACT, Inc. (1998). *Assessing listening comprehension: A review of recent literature relevant to an LSAT Listening Component* [unpublished manuscript]. Newton, PA: Law School Admission Council.

ACT, Inc. (1999). *Technical Manual for the ESL Exam.* Iowa City, IA:Author.

Adams, R.J., & Wilson, M.R. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In G.Engelhard and M. Wilson, (Eds.), *Objective measurement: Theory into practice, Vol. 3* (pp. 143–166). Nordwood, NJ: Ablex Publishing Corporation.

Adams, R.J., Wilson, M.R., & Wang, W.C. (1997). The random coefficients multinomial logit. *Applied Psychological Measurement, 21,* 1–25.

Adams, R.J., Wilson, M.R., & Wu, M. (1997). Multilevel item response theory models: an approach to errors in variables of regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Adema, J.J., & Van der Linden, W.J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics, 14,* 279–290.

Aitchison, J., & Silvey, S.D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics 29,* 813–828.

Aitkin, M., & N.Longford (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Association,* Series A (General), *149,* Part 1, 1–43.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical modeling of data on teaching styles. *Journal of the Royal Statistical Society, Series A (General), 149,* 1–43.

Albert, J.H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Alkin, M.C., Daillak, R., & White, P. (1979). *Using evaluations.* Beverly Hills: Sage.

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69–81.

Anderson L.W. & Bourke, S.F. (2000). *Assessing affective characteristics in the schools.* Mahwah, NJ: Lawrence Erlbaum.

Anderson, L.W., Ryan, D.W., & Shapiro, B.J. (1989). *The IEA Classroom Environment Study.* Oxford: Pergamon Press.

APA, AERA, & NCME (1985). *Standards for educational and psychological tests.* Washington DC: American Psychological Association, American Educational Research Association, National Council on Measurement in Education.

Argyris, C. (1976). Single-Loop and Double-Loop Models in Research on Decision Making. *Administrative Science Quarterly, 21*(3), 363–375.

Argyris, C., & Schön, D.A. (1974). *Theory in practice: increasing professional effectiveness.* San Francisco: Jossey-Bass.

Bacharach, S.B, Bamberger, P., Conley, S.C., & Bauer, C. (1990). The Dimensionality of Decision Participation in Educational Organizations: The Value of a Multi-Domain Evaluative Approach. *Educational Administration Quarterly, 26*(2), 126–167.

Bangert, R.L., Kulik, J.A., Kulik, C.C. (1983). Individualized systems of instruction in secondary schools. *Review of Educational Research, 53,* 143–158.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541–562.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N.Frederiksen, R.J.Mislevy, & I.I.Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323–357). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bentler, P. (1992). *EQS.* (Computer Software).

Bereiter, C., & Kurland, M. (1982). A constructive look at follow through results. *Interchange, 12,* 1–22.

Berk, R.A. (1999). Assessment for measuring professional Performance. In: D.P. Ely, L.E.Odenthal, & T.Plomp (Eds.), *Educational Science and Technology: Perspectives for the future*pp. 30–48). Enschede, Twente University Press.

Berryman, S.E., Boyle, N., Golladay, F., Holmes, M., Keefer Ph., & Sigrist, K. (1997). Guidelines for assessing institutional capability. Draft paper, World Bank.

Birnbaum, A. (1968). Some latent trait models. In F.M.Lord & M.R.Novick (Eds.), *Statistical theories of mental test scores,* Reading MA: Addison-Wesley.

Bloom, B. (1976). *Human characteristics and school learning.* New York: McGraw Hill.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H., & Krathwohl, D.R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay.

Bloom, B.S., Hastings, J.T., & Madaus, G.F. (1971). *Handbook of formative and summative evaluation of student learning.* New York: McGraw-Hill.

Bluhm, H.P., & Visscher, A.J. (1990). Administrative Computing in the USA and The Netherlands: Implications for other countries. *School Organizations, 10*(1), 107–117.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443–459.

Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement 12,* 261–280.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika, 1,* 1101–1112.

Boekkooi-Timminga, E. (1989). *Models for computerized test construction.* Unpublished doctoral thesis. University of Twente, Enschede, The Netherlands.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics, 75,* 129–145.

Borich, G.D., & Jemelka, R.P. (1982). *Programs and Systems. An evaluation perspective.* New York: Academic Press.

Bosker, R.J., & Hendriks, M.A. (1997). *Betrouwbaarheid, validiteit en bruikbaarheid van een instrumentarium voor zelfevaluatie door basisscholen.* [Reliability, validity, and usability of instruments for school self evaluation in primary education]. Enschede: University of Twente/OCTO.

Bosker, R.J., & Scheerens, J. (1995). A self-evaluation procedure for schools using multilevel modelling. *Tijdschrift voor Onderwijsresearch, 20*(2), 154–164.

Bosker, R.J., & Scheerens, J. (1999). Openbare prestatiegegevens van scholen; nuttigheid en validiteit. *Pedagogische Studiën, 76*(1), 61–73.

Bosker, R.J., & Witziers, B. (1996). *The magnitude of school effects. Or: Does it really matter which school a student attend.* Paper presented at AERA Annual meeting, New York.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153–168.
Brandsma, H.P. (1993). *Basisschoolkenmerken en de kwaliteit van het onderwijs.* Groningen: RION.
Bray, M. (1994). Centralization/decentralization and privatization/ publicization: conceptual issues and the need for more research. In W.K.Cummings & A. Riddell (Eds.), *Alternative Policies for the Finance, Control, and Delivery of Basic Education.* Special issue of the *International Journal of Educational Research, 21*(8), 817–824.
Brennan, R.L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27–34.
Brennan, R.L., & Johnson, E.G. (1995). Generalizability of performance assessments. *Educational Measurement: Issues and Practice, 14*(4), 9–12.
Breukers, H., De Haan, D., Rikers, J., Glas, C.A.W., Boon, B., Gorissen, V., & Hogen, D. (1992). *Vooronderzoek ITEM: Definitiestudierapport.* Heerlen: Open Universiteit, OTIC.
Brookover, W., Beady, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *School social systems and student achievement: Schools can make a difference.* New York: Bergin.
Brophy, J. (1996). *Classroom management as socializing students into clearly articulated roles.* Paper presented at AERA Annual Meeting, New York.
Brophy, J., & Good, T.L. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Third Handbook of Research on Teaching* (pp. 328–375). New York: MacMillan.
Burnstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In: M.D.Shermis & J.Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
Butcher, J.N. (1987). *Computerized Psychological Assessment.* New York, NJ: Basic Books, Inc.
Cameron, K.S., & Whetten, D.A. (Eds.) (1983). *Organizational Effectiveness. A comparison of multiple models.* New York: Academic Press.
Cameron, K.S., & Whetten, D.A. (1985). Administrative effectiveness in higher education. *The Review of Higher Education, 9,* 35–49.
Camilli, G., & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items.* Thousand Oaks, CA: Sage.
Campbell, D.T. (1969). Reforms as Experiments. *American Psychologist, 24*(4), p. 409.
Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin 56,* 81–105.
Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi-experimental designs for research on teaching. In: N.L.Gage (Ed.), *Handbook of Research on Teaching.* Chicago: Rand McNally Publishing Company.
Caplan, N. (1982). Social research and public policy at the national level. In D.B.P. Kallen (Ed.), *Social science research and public policy-making: a reappraisal.* Windsor, U.K.: NFER-Nelson.
Card, D., & Krueger, A.B. (1992). Social quality and black-white relative earnings: a direct assessment. *The quarterly journal of economics, 107*(1), p. 151.
Carroll, J.B. (1963). A model of school learning. *Teachers College Record, 64,* 722–733.
Carvalho, S., & White, H. (1996). Implementing projects for the poor: what has been learned? Washington D.C.: The World Bank.
Chapman, C. (2002). Ofsted and School Improvement: Teachers' Perceptions of the Inspection Process in Schools Facing Challenging Circumstances. *School Leadership and Management, 22*(3), 257–272.
Cheng, Y.C. (1993). *Conceptualization and measurement of school effectiveness: An organizational perspective.* Paper presented at AERA annual meeting, Atlanta, GA.
Cibulka, J.G., & Derlin, R.L. (1995). State educational performance reporting policies in the U.S.: Accountability's many faces. *International Journal of Educational Research, 23*(6), 479–492.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.
Cohen, M.D., March, J.G., & Olsen, J.P. (1972). A garbage can model or organizational choice. *Administrative Science Quarterly, 17,* 1–25.
Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of educational opportunity* . Washington, DC: U.S. Government Printing Office.
Conley, D.T. (1997). *Roadmap to Restructuring Oregon.* Eric Clearing House.
Cook, Th.D., & Campbell, D.T. (1979). *Quasi Experimentation. Design and Analysis, Issues for Field Settings.* Chicago: Rand McNally Publ. Comp.
Cotton, K. (1995). *Effective schooling practices: A research synthesis.* 1995 Update. School Improvement Research Series. Northwest Regional Educational Laboratory.
Creemers, B.P.M. (1994). *The Effective Classroom.* London: Cassell.
Creemers, B.P.M., Reynolds, D.,& Swint, F.E. (1994). The International School Effectiveness Research Programme ISERP First Results of the Quantitative Study. Paper presented at the British Education Research Association conference, Oxford, September 1994.
Cronbach, L.J. (1971). Test validation. In R.L.Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington DC: American Council on Education.
Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.
Cronbach, L.J. and Associates (1980). *Toward reform of program evaluation. Aims, methods and institutional arrangements.* San Fransisco: Jossey Bass.
Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481.
Davies, J.K. (1972). Style and effectiveness in education and training: a model for organizing, teaching and learning. *Instructional Science,*
De Jong, T., & Van Joolingen, W.R. (1996). Discovery learning with computer simulations of conceptual domains. Enschede: University of Twente.
De Leeuw, A.C.J. (1982). *Organisaties: Management, Analyse, Ontwerp en Verandering.* Een systeemvisie. Assen: Van Gorcum.
Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B, 39,* 1–38.
Denzin, N.K. (1978). *The research art.* New York: Mc Graw-Hill.
DfEE (1996). *Setting Targets to Raise Standards: A Survey of Good Practice.* London: Department for Education and Employment.
DfEE (1998a). *School Evaluation Matters.* London: Department for Education and Employment.
DfEE (1998b). *The Autumn Package: Pupil Performance Information KS1–3 & GCSE/GNVQ.* London: Department for Education and Employment.
Dobart, A. (2001). Development of Schools Inspection in Austria and Its Neighbouring Countries. Inspecting in a New Age—International Meeting on the Occasion of the 200th Anniversary of the Netherlands Inspectorate of Education. N.Troost. Utrecht, Netherlands, the Netherlands Inspectorate of Education, 21–24.
Doolaard, S. (1996). Changes in characteristics and effects of school leadership over time. Paper presented at ECER, Sevilla.
Doyle, W. (1985). Effective secondary classroom practices. In: M.J.Kyle (ed.), *Reaching for excellence. An effective schools sourcebook.* Washington DC: US Government Printing Office.
Drábek, P. (2000). Following Up On Inspection—Report on the SICI Workshop Held in Podebrady, the Czech Republic, 4 to 6 October 2000. Utrecht, SICI.
Dror, Y. (1968). *Public policy-making re-examined.*, Scranton Pens.: Chandler.
Duffy, Th.M., & Jonassen, D.H. (1992). *Constructivism and the Technology of Instruction: A Conversation.* Hillsdale, NJ: Lawrence Erlbaum Ass.

Ebel, R.L. (1951). Writing the test item. In: E.F.Lindquist (Ed.)., *Educational Measurement* (1st ed., pp. 185–249). Washington, DC: American Council on Education.
Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of Educational Measurement.* Englewood Cliffs, N.J.: Prentice Hall.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics, 7,* 1–26.
Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and the cross-validation. *The American Statistician, 37,* 36–49.
Eisner, E.W. (1979). *The educational imagination. On the design and evaluation of school programs.* New York: Macmillan Publ. Co Inc.
Elliot, K., Smees, R., & Thomas, S. (1998). Making the Most of Your Data: School Self-Evaluation Using Value Added Measures. *Improving Schools Journal, 1*(3), 59–67.
Emons, W.H.M. (1998). Nonequivalent Groups IRT Observed Score Equating. Its Applicability and Appropriateness for the Swedish Scholastic Aptitude Test. Twente University.
ETS (1990). *Accuplacer* (Computer Software). Princeton, NJ: Educational Testing Service.
ETS (1994). *Computer-based tests: Can they be fair to everyone?* Princeton, NJ: Educational Testing Service.
ETS (1998). *Computer-Based TOEFL Score User Guide.* Princeton, JJ: Author.
Evers, A., Vliet-Mulder, J.C. van, Resing, W.C.M., Starren, J.C.M.G., van Alphen de Veer, R.J., van Boxtel, H. (2002). *COTAN: Testboek voor het onderwijs.* NDC Boom.
Ezemenari, K., Rudqvist, A., & Subbarao, K. (1998). *Impact evaluation: a note on concepts and methods.* PRMPO World Bank, Washington D.C.
Faerman, S.R., & Quinn, R.E. (1985). Effectiveness: the perspective from organization theory. *Review of Higher Education, 9,* 83–100.
Ferguson, R.F. (1991). Paying for public education: New evidence on how and why money matters. *Harvard Journal on Legislation, 28*(2), 465–498.
Fischer, G.H. (1974). *Einführung in die theorie psychologischer tests: Introduction to the theory of psychological tests.* Bern: Huber.
Fischer, G.H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 459–487.
Fischer, G.H., & Molenaar, I.W. (1995). *Rasch models. Their foundation, recent developments and applications.* New York, NJ: Springer.
Fischer, G.H., & Scheiblechner, H.H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12,* 23–51.
Fitz-Gibbon, C.T. (1991). Multilevel modelling in an indicator system. In S.W. Raudenbush & J.D.Willms (Eds.), *Schools, Classrooms and Pupils International Studies of Schooling from a Multilevel Perspective.* San Diego: Academic Press.
Fleishman, E.A., & Quaintance, M.K. (1984). *Taxonomies of human performance: the description of human tasks.* London: Academic Press.
Fox, J.P., & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika, 66,* 271–288.
Fox, J.P., & Glas, C.A.W. (2002). Modeling measurement error in structural multilevel models. In G.A.Marcoulides and I.Moustaki (Eds.). *Latent Variable and Latent Structure models.* Mahwah, NJ: Laurence Erlbaum.
Fraser, B.J., Walberg, H.J., Welch, W.W., & Hattie, J.A. (1987). Syntheses of educational productivity research. Special Issue of the *International Journal of Educational Research, 11*(2).
Fraser, C. (1988). *NOHARM: A Computer Program for Fitting Both Unidimensional and Multidimensional Normal Ogive Models of Latent Trait Theory.* (Computer Software). NSW: University of New England.
Frisbie, D.A., & Becker, D.F. (1991). An analysis of textbook advice about truefalse tests. *Applied Measurement in Education, 4,* 67–83.
Fuller, B. (1987). What factors raise achievement in the Third World? *Review of Educational Research, 57,* 255–292.

Fuller, B., & Clarke, P. (1994). Raising school effects while ignoring culture? Local conditions and the influence of classroom tools, rules and pedagogy. *Review of Educational Research, 64,* 119–157.
Fuller, W.A. (1987). *Measurement Error Models.* New York, NJ: Wiley.
Gage, N. (1965). Desirable behaviors of teachers. *Urban Education, 1,* 85–95.
Galton, M. (Ed.) (1998). Class Size and Pupil Achievement. *International Journal of Educational Research, 29*(8), 689–818.
Gelfand, A.E., & Smith, A.F.M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association, 85,* 398–409.
Glas, C.A.W. (1988a). The derivation of some tests for the Rasch model from the multinomial distribution. *Psychometrika, 53,* 525–546.
Glas, C.A.W. (1988b). The Rasch model and multi-stage testing. *Journal of Educational Statistics, 13,* 45–52.
Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M.Wilson (Ed.), *Objective Measurement: Theory into practice, Vol. 1* (pp. 236–258), New Jersey, NJ: Ablex Publishing Corporation.
Glas, C.A.W. (1997). Towards an integrated testing service system. *European Journal of Psychological Assessment, 13,* 38–48.
Glas, C.A.W. (1998) Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* 647–667.
Glas, C.A.W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64,* 273–294.
Glas, C.A.W. en Béguin, A.A. (1996). Appropriateness of IRT observed score equating. OMD Research Reports, 96–4, Twente University.
Glas, C.A.W., & Suárez-Falćon, J.C. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27,* 87–106.
Glas, C.A.W., & van der Linden, W.J. (2001). *Modeling Variability in Item Parameters in Educational Measurement.* Twente University, OMD Research Report 01–11.
Glas, C.A.W., & van der Linden, W.J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 256–272.
Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G.H.Fischer & I.W.Molenaar (Eds.), *Rasch models: foundations, recent developments and applications.* (pp. 325–352). New York, NJ: Springer.
Glas, C.A.W., and Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54,* 635–659.
Glas, C.A.W., Wainer, H., & Bradlow (2000). MML and EAP estimates for the testlet response model. In W.J.van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp. 271–287). Boston MA: Kluwer-Nijhoff Publishing.
Glass, G.V. (1972). The many faces of educational accountability. *Phi Delta Kappan, 53,* 636–639.
Godwin, J. (1999, April). *Designing the ACT ESL Listening Test.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal Canada.
Goldstein, H. (1986). Multilevel mixed linear models analysis using iterative generalized least squares. *Biometrika, 73,* 43–56.
Goldstein, H. (1993). Assessment and Accountability. *Education, 183*(3), 33–34.
Goldstein, H. (1995). *Multilevel Statistical Models,* second edition. Kendall's library of statistics 3. London, Sydney, Auckland: Edward Arnold.
Goldstein, H. (1997). Methods in School Effectiveness Research. *School Effectiveness and School Improvement, 8*(4), 369–395.
Goldstein, H., & Healy, M. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, 581*(1), 175–177.

Goldstein, H., & Sammons, P. (1997). The influence of secondary and junior schools on sixteen year examination performance: a cross-classified multilevel analysis. *School Effectiveness and School Improvement, 8*(2), 219–230.

Goldstein, H., & Spiegelhalter, D. (1996). League Tables and Their Limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society A, 159*(3), 385–443.

Goldstein, H., Baxter-Jones, A., Helms, P., & Ware, J.H. (1993). Models for analysis of longitudinal data. *The European Respiratory Journal, 6*(9), p. 1416.

Goodlad, J.F., & Anderson, R.H. (1987). *The nongraded elementary school.* New York: Columbia University, Teachers College.

Gooren, W.A.J. (1989). Kwetsbare en weerbare scholen en het welbevinden van de leraar. In: J.Scheerens & J.C.Verhoeven (Eds.), *Schoolorganisatie, beleid en onderwijskwaliteit.* Lisse: Swets & Zeitlinger.

Gray, J., Jesson, D., Goldstein, H., Hedger, K., & Rasbash, J. (1995). A multilevel analysis of school improvement: changes in schools' performance over time. *School Effectiveness & School Improvement, 6*(2), 97–114.

Gray, J., Goldstein, H., & Jesson, D. (1996). Changes and improvements in schools' effectiveness: trends over five years. *Research Papers in Education, 11*(1), 35–51.

Gray, J., Hopkins, D., & Reynolds, D. (1998). *The Improving Schools Research Project.* A paper presented at The International Congress of School Effectiveness & Improvement, Manchester, January 1998.

Gray, J., Jesson, D., & Sime, N. (1990). Estimating differences in the examination performance of secondary schools in six LEAs: a multilevel approach to school effectiveness. *Oxford Review of Education, 16*(2), 137–158.

Green, S.B., Halpin, G., & Halpin, G.W. (1990). *The emphasis on rote memory items on classroom tests: Why are teachers so interested in hearing their own lectures.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Greeno, J.G. (1980). Some examples of cognitive task analysis with instructional implications. In: R.E.Snow, P-A.Federico, & W.E.Montague (Eds.), *Aptitude, learning, and instruction: Vol. 2. Cognitive process analyses of learning and problem solving.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Grisay, A. (1996). *Evolution des acquis cognitifs et socio-affectifs des eleves au cours des annees de college.* Liège: Université de Liège.

Guba, E.G. (1978). Toward a methodology of naturalistic inquiry in educational evaluation. Los Angeles: CSE Monograph in Evaluation.

Guba, E.G., & Lincoln, S. (1982). Effective evaluation. Improving the responsiveness of educational results through responsive and naturalistic approaches. San Francisco: Jossey Bass.

Gulliksen, H. (1950). *Theory of mental tests.* New York, NJ: Wiley.

Haladyna, T.M. (1992). Context dependent item sets. *Educational Measurement: Issues and Practices, 11,* 21–25.

Haladyna, T.M. (1994). *Developing and validating multiple-choice test items.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Haladyna, T.M. (1997). *Writing test items to evaluate higher order skills.* Boston: Allyn and Bacon.

Hambleton, R.K., & Rogers, H.J. (1989). Detecting potentially biased test items: comparison of IRT area and Mantel-Haenszel methods. *Applied Measurement in Education, 2,* 313–334.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications* (2nd ed.). Boston: Kluwer-Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundaments of item response theory.* Newbury Park, CA: Sage.

Hanushek, E.A. (1979). Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources, 14,* 351–388.

Hanushek, E.A. (1986). The economics of schooling: production and efficiency in public schools. *Journal of Economic Literature, 24,* 1141–1177.
Hanushek, E.A. (1995). Interpreting Recent Research on Schooling in Developing Countries. *The World Bank Research Observer, 10,* 227–246.
Hanushek, E.A. (1997). Assessing the effects of school resources on student performance: an update. *Educational Evaluation and Policy Analysis, 19,* 141–164.
Hauser, R.M., Sewell, W.H., & Alwin, D.F. (1976). High School effects on achievement. In W.H.Sewell, R.M.Hauser & D.L.Featherman (Eds.), *Schooling and achievement in American Society.* New York: Academic Press.
Haywood, H.C. (1982). Compensatory education. *Peabody Journal of Education,* 59, 272–301.
Hedges, L.V., Laine, R.D., & Greenwald, R. (1994). Does money matter? A metaanalysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher, 23*(3), 5–14.
Hendrawan, I., Glas, C.A.W., & Meijer, R.R. (2001). *The Effect of Person Misfit on Classification Decisions.* Research Report 01–05, Faculty of Educational Science and Technology, University of Twente, the Netherlands.
Hendriks, M.A., & Scheerens, J. (1996). Zelfevaluatie in het basisonderwijs. Constructie van een instrumentarium “School-en klaskenmerken”. Enschede: OCTO.
Hill, P., & Goldstein H. (1998). Multilevel Modelling of Educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioural Statistics, 23*(2), 117–128.
Hill, P.W., & Rowe, K.J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement,* 7(1), 1–34.
Hill, P.W., Rowe, K.J., & Holmes-Smith, P. (1995). *Factors Affecting Students’ Educational Progress: Multilevel Modelling of Educational Effectiveness.* Paper presented at the 8th International Congress for School Effectiveness and Improvement, Leeuwarden, the Netherlands, January 3–6, 1995.
Hirsch, D. (1994). *School: A matter of choice.* Paris: OECD/CERI.
Hively, W., Patterson, H.L., & Page, S.H. (1968). A ‘universe-defined’ system of arithmetic achievement items. *Journal of Educational Measurement, 5,* 275–290.
HMSO (Her Majesty’s Stationery Office, 1992). The Education (Schools) Act 1992 (c. 38). London, the Stationery Office Limited.
HMSO (Her Majesty’s Stationery Office, 1996). The School Inspections Act 1996. London, the Stationery Office Limited.
Holland, P.W. and Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H.Wainer and H.I.Braun (Eds.), *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Holland, P.W. en Rubin, D.B. (1982). *Test Equating.* New York: Academic Press.
Holland, P.W., & Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, N.J., Erlbaum.
Hopkins, D. (1987). *Doing school based review.* Leuven, Belgium: Acco.
Huberman, M. (1987). Steps towards an integrated model of research utilization. *Knowledge: Creation, Diffusion, Utilization, 8*(4), 586–611.
Inspectie van het Voortgezet Onderwijs. (1992). *Examens op Punten Getoetst.* ’s Gravenhage: Inspectie van het Voortgezet Onderwijs.
Irwin, C.C. (1986). What research tells the principal about educational leadership. *Scientica Paedagogica Experimentalis, 23,* 124–137.
James, E. (1994). The public-private division of responsibility in education. In W.K. Cummings & A.Riddell *(Eds.), Alternative policies for the finance, control, and delivery of basic education.* Special issue of the *International Journal of Educational Research, 21*(8), 777–783.
Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 285–306.

Jimenez, E., & Paquea, V. (1996). Do Local Contributions Affect the Efficiency of Public Schools? *Economics of Education Review, 15*, 377–386.
Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling.* New York, NJ: Springer.
Joint Committee on Standards for Educational Evaluation (1981). *Standards for Educational Evaluation* New York: Holt, Rhinehart and Winston.
Jöreskog, K.G. & D.Sörbom, (1996). *LISREL.* (Computer Software). Chicago, IL: Scientific Software International, Inc.
Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin,* 112, 527–535.
Kane, M.T., Crooks, T.J., & Cohen, A.S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5–17.
Keefe, J. (1994). CASE/IMS (Computer program) (Nederlandse licentie Roders, R., Van der Wolf, J.C., Amsterdam: Seneca). U.S.A.: NASSP.
Kelderman, H. (1984). Loglinear RM tests. *Psychometrika, 49,* 223–245.
Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika, 54,* 681–697.
Kerr, S. (1977). Substitutes for leadership: Some implications for organizational design. *Organizational and Administrative Sciences, 8,* 135–146.
Kervezee, K. (2001). Speech at the Opening of the International Meeting 'Inspecting In a New Age'. Inspecting in a New Age—International Meeting on the Occasion of the 200th Anniversary of the Netherlands Inspectorate of Education. N.Troost. Utrecht, Netherlands, the Netherlands Inspectorate of Education, 7–9.
Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887–903.
Knuver, J.W.M., & Doolaard, S. (1996). Rekenen/wiskunde en natuuronderwijs op de basisschool. Enschede: OCTO.
Kok, F.G., Mellenbergh, G.J., & van der Flier, H. (1985). Detecting experimentally induced item bias using the iterative logit method. *Journal of Educational Measurement, 22,* 295–303.
Kolen, M.J. en Brennan, R.L. (1995). Test Equating. New York: Springer.
Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273–287.
Lawley, D.N. (1944). The factorial analysis of multiple test items. *Proceedings of the Royal Society of Edinburgh, 62-A,* 74–82.
Leeuw, F.L., Gils, G.H.C. van., & Kreft, C. (1999). Evaluating anti-corruption initiatives. Underlying logic and mid-term impact of a World Bank program. *Evaluation, 5*(2), 194–219.
Leithwood, K.A., & Montgomery, D.J. (1982). The role of the elementary school principal in program improvement. *Review of Educational Research, 52,* 309–399.
Leune, J.M.G. (1987). Besluitvorming en machtsverhoudingen in het Nederlandse onderwijsbestel. (Decision-making and the balance of power in Dutch education). In J.A. van Kemenade et al. (Red.), *Onderwijs, bestel en beleid deel 2.* Groningen: Wolters-Noordhoff.
Leune, J.M.G. (1994). Onderwijskwaliteit en de autonomie van scholen. In B.P.M. Creemers (Red). *Deregulering en de kwaliteit van het onderwijs.* (pp. 27–48). Groningen: RION.
Levine, D.K., & Lezotte, L.W. (1990) *Unusually Effective Schools: A Review and Analysis of Research and Practice.* Madison, Wise: Nat. Centre for Effective Schools Research and Development.
Lockheed, M.E. (1988). The measurement of educational efficiency and effectiveness. AERA paper. New Orleans.
Lockheed, M., & Hanushek, E. (1988). Improving educational efficiency in developing countries: What do we know? *Compare, 18*(1), 21–38.
Lockheed, M., & Verspoor, A. (1991). *Improving Primary Education in Developing Countries.* London: Oxford University Press.
Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random factors. *Biometrika, 74,* 817–827.

Longford, N.T. (1993). *Random Coefficients Models.* Oxford: Clarendon Press.
Lord, F.M. (1952). *A theory of test scores.* Psychometric Monograph 7.
Lord, F.M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57–75.
Lord, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–548.
Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, N.J., Erlbaum.
Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.
Lortie, D.C. (1973). Observations on teaching as work. In R.M.W.Travers (Ed.), *Second Handbook of Research on Teaching.* Chicago: Rand McNally.
Louis, K.S., & Smith, B.A. (1990). Teachers' work: Current issues and prospects for reform. In: P.Reyes (Ed.), *Productivity and Performance in Educational Organizations* [pp. 23–47]. Newbury Park: Sage.
Luyten, H. (1994). Stability of School Effects in Dutch Secondary Education: The impact of variance across subjects and years. *International Journal of Educational Research, 21*(2), 197–216.
MacBeath, J. (1999). *Schools must speak for themselves: the case for school selfevaluation.* London: Routledge.
MacBeath, J., Meuret, D., Schratz, M., & Jakobsen, L.B. (1999). Evaluating quality in school education. A European pilot project. Brussels: European Commission.
MacBeath, J., & Mortimore, P. (1994). *Improving School Effectiveness—A Scottish Approach.* British Educational Research Association 20th Annual Conference, September, St Anne's College, Oxford.
MacBeath, J., & Mortimore, P. (Eds.) (2001). *Improving School Effectiveness.* Open University press.
Macready, G.B. (1983). The use of generalizability theory for assessing relations among items within domains in diagnostic testing. *Applied Psychological Measurement, 7,* 149–157.
Macready, G.B., & Merwin, J.C. (1973). Homogeneity within item forms in domain referenced testing. *Educational and Psychological Measurement, 33,* 351–360.
Madaus, G.F., & Stufflebeam, D.L. (1983) *Evaluation in education and human services.* Dordrecht: Kluwer-Nijhof.
March, J.T., & Olsen, J.P. (1976). *Ambiguity and choice in organizations.* Bergen: Universitetsforlaget.
Maslowski, R., & Visscher, A.J. (1997). *Methoden en technieken voor formatieve evaluatie in sociaal-wetenschappelijke ontwerpsituaties* [Methods and techniques for formative evaluation in design contexts in the social sciences]. Enschede: University of Twente, Department of Educational Organisation and Management.
Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–174.
Matthews, P., & Smith, G. (1995). OFSTED: Inspecting Schools and Improvement Through Inspection. *Cambridge Journal of Education, 25*(1), 23–34.
McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometric monographs, No. 15.*
McDonald, R.P. (1982). Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement, 6,* 379–396.
McLaughlin, M.W. (1990). The RAND change agent study revisited: Macro perspectives and micro realities. *Educational Researcher, 19*(9), 11–16.
McLaughlin, M.W., & S.Mei-ling Yu (1988). School as a place to have a career. In: A.Lieberman, *Building a professional culture in schools.* New York.
Medley, D., & Mitzel, H. (1963). Measuring classroom behavior by systematic observation. In N.L.Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand McNally.

Mehrens, W.A., & Lehmann, I.J. (1975). *Measurement and evaluation in education and psychology.* New York: Holt, Rinehart and Winston.
Meisner, R., Luecht, R.M., Reckase, M.D. (1993). *The comparability of the statistical characteristics of test items generated by computer algorithms* (ACT Research Report Series No. 93–9). Iowa City, IA: ACT, Inc.
Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91–100.
Merril, M.D., & Tennyson, R.D. (1977). *Teaching concepts: An instructional design guide.* Englewood Cliffs, NJ: Educational Technology Publications.
Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30,* 955–966.
Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement, 21,* 215–237.
Messick, S. (1989). Validity. In R.L.Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–104). New York, NJ: American Council on Education and Macmillan.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14,* (4), 5–8.
Meuret, D., & Scheerens, J. (1995). An international comparison of functional and territorial decentralization of public educational systems. Paper presented at AERA 1995, San Francisco.
Millman, J., & Green, J. (1989). The specification and development of tests of achievement and ability. In R.L.Linn (Ed.), *Educational Measurement* (3rd ed., pp. 335–366). New York, NJ: American Council on Education and Macmillan.
Millman, J., & Westman, R.S. (1989). Computer-assisted writing of achievement test items: Toward a future technology. *Journal of Educational Measurement, 26,* 177–190.
Minsky, M. (1975). A framework for representing knowledge. In: P.H.Winston (Ed.), *The psychology of computer vision* (pp. 211–280). New York, NJ: MacGraw-Hill.
Mintzberg, H. (1979). *The structuring of organizations.* Englewood Cliffs, NJ: Prentice-Hall.
Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359–381.
Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177–195.
Mislevy, R.J., & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D.Bock (Ed.), *Multilevel analysis of educational data.* San Diego: Academic Press.
Mislevy, R.J., & Bock, R.D. (1990). *PC-BILOG. Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.
Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement, 1,* 3–62.
Mitchell, D.E., & Tucker, Sh. (1992). Leadership as a way of thinking. *Educational Leadership, 49*(5), p. 30.
Mitchell, D.E., & Tucker, Sh. (1992). Leadership as a way of thinking. *Educational Leadership, 49*(5), p. 30.
Molenaar, I.W. (1995). Estimation of item parameters. In: G.H.Fischer, & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications.* New York, NJ: Springer.
Morgan, G. (1986). *Images of organization.* London: Sage.
Mortimore, P., & Whitty, G. (1997). *Can School Improvement Overcome the Effects of Disadvantage?* London: Institute of Education, University of London.
Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School matters: the junior years.* Somerset: Open books.
Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16,* 159–176.
Muthén, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.
Muthén, B. (1987). *LISCOMP.* (Computer Software).

Nevo, D. (1995). *School-based evaluation: a dialogue for school improvement.* Oxford Pergamon.
Neyman, J., and Scott, E.L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16*, 1–32.
Niskanen, W.A. (1971). *Bureaucracy and representative government.* Chicago: AldineAtherton.
Nissan, S. (1999, April). *Incorporating sound, visuals, and text for TOEFL on computer*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
Novak, J.D., & Gowin, D.B. (1984). *Learning how to learn.* New York, NJ: Cambridge University Press.
Nuttall, D.L. (1989). International Educational Indicators. The conceptual paper. Paper for the meeting of OECD Educational Indicator Project, San Francisco.
O'Donoghue, C., Thomas, S., Goldstein, H., & Knight, T. (1997). *1996 DfEE study of Value Added for 16–18 Year Olds in England.* London: HMSO.
Oakes, J. (1987). *Conceptual and measurement problems in the consruction of school quality.* Paper presented at AERA annual meeting, Washington.
Oakes, J. (1989). What educational indicators?: the case for assessing school context. *Educational Evaluation and Policy Analysis, 11*, 181–199.
OECD (1995). *Education at a Glance.* Paris: OECD.
OECD (1998). *Education at a Glance.* Paris: OECD.
OECD (1999). *Education at a Glance.* Paris: OECD.
Ofsted (1999a). Handbook for Inspecting—Primary and Nursery Schools with Guidance on Self-Evaluation. London, Office for Standards in Education.
Ofsted (1999b). Handbook for Inspecting—Secondary Schools with Guidance on Self-Evaluation. London, Office for Standards in Education.
Ofsted (1999c). Inspecting Schools—the Framework—Effective from January 2000. London, Office for Standards in Education.
Ofsted (2000) Parents' questionnaire. London, the Stationery Office.
Ofsted (2000a). Form S1 (SECONDARY)—Consultation about the Inspection and Information about the School. London, Office for Standards in Education.
Ofsted (2000b). Form S2 (SECONDARY)—Information about the School. London, Office for Standards in Education.
Ofsted (2000c). Form S3 (All Schools)—School Self-Audit. London, Office for Standards in Education.
Ofsted (2002). 2002 PANDA Report for An Anonymous Second School Unique Reference Number (URN): 999999 DfES Number: 9999999 UNVALIDATED DATA. London, Office for Standards in Education.
Ofsted (2002). Form S4 (All Schools)—Self-Evaluation Report. London, Office for Standards in Education.
Ofsted (2002). Improvement Through Inspection—10 Years on. (NR 2002–188C, 11 November 2002). London, Ofsted Press Office.
Ofsted (2002). Ofsted Launches New Web Site. (NR 2002–188A, 11 November 2002). London, Ofsted Press Office.
Ofsted (2002). Speech by David Bell at the QEII Conference Centre, Westminster, London to Mark Ofsted's Tenth Anniversary. (NR 2002–188B, 11 November 2002). London, Ofsted Press Office.
Ofsted (2002). Update 38 (Spring 2002). London, Office for Standards in Education.
Ofsted (2002). Update 40 (Autumn 2002). London, Office for Standards in Education.
Ofsted (2003). Inspecting Schools—the Framework for Inspecting Schools in England from September 2003. London, Office for Standards in Education.
Ofsted (2003). Ofsted Announces New Framework for School Inspection—Chief Inspector Hails Interactive Inspections as Way Forward for Ofsted. (NR 2003–4, 31 Janurary 2003). London, Ofsted Press Office.

Ofsted website (accessed Dec, 2002). Ofsted publications. http://www.ofsted.gov.ukypublications/index.cfm?fuseaction=pubs.summary&id=1005

Ofsted website (accessed Jan. 2003). How we inspect. http://www.ofsted.gov.ukyhowwework/index.cfm?fuseaction=howwework.inspectionsHome

Ofsted website (accessed Jan., 2003). How We Inspect State Schools. www.ofsted.gov.uk/howwework/index.cfm?fuseaction=howwework.inspections&id=10

Ofsted/SEU (2001). Reducing The Burden Of Inspection. London, OFSTED/SEU.

Olson-Buchanan, J.B., Drasgow, F., Moberg, P.J., Mead, A.D., Keenan, P.A., & Donovan, M.A. (1998). Interactive video assessment of conflict resolution ski8lls. *Personnel Psychology, 51,* 1–24.

Orbach, E. (1998). How to assess the capacity of client government to implement an education development project. Washington: World Bank.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Osler, D. (2001). The Value of Inspectorates of Education in the 21st Century. *Inspecting in a New Age—International Meeting on the Occasion of the 200th Anniversary of the Netherlands Inspectorate of Education.* N.Troost (ed.). Utrecht, Netherlands, the Netherlands Inspectorate of Education, 12–19.

Ouston, J. and Davies, J. (1998). OFSTED and Afterwards? Schools' Responses to Inspection. In P.Earley (Ed.), *School Improvement After Inspection? School and LEA Responses.* London: Paul Chapman Publishing Ltd.

Parlett, M., & Hamilton, D. (1972). Evaluation as illumination: a new approach to the study of innovatory programs. Occasional Paper, University of Edinburgh.

Parshall, C.G., Davey, T., & Pashley, P.J. (2000). Innovative item types for computerized testing. In: W.J.van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp. 129–148). Boston: Kluwer-Nijhoff Publishing.

Parshall, C.G., Stewart, R., & Ritter, J. (1996, April). *Innovations: Sound, graphics, and alternative response modes.* Paper presented at the annual meeting of the National Council of Measurement in Education, New York.

Patton, M.Q. (1978). *Utilization-focused evaluation.* Beverly Hills: Sage.

Patton, M.Q. (1980). *Qualitative Evaluation Models.* Beverly Hills: Sage Publications.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple Item Types, Missing Data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342–366.

Pfeffer, J., & Salancik, G.R. (1978). *The external control of organizations. A resource dependence perspective.* New York: Harper & Row.

Picciotto, R. (1996). What is Education Worth? From Production Function to Institutional Capital. World Bank: Human Capital Development Working Paper, no. 75.

Postlethwaite, T.N., & Ross, K.N. (1992). *Effective Schools in Reading. Implications for Educational Planners.* The Hague: IEA.

Pritchett, L., & Filmer, D. (1997). What Educational Production Functions Really Show. A positive theory of education. The World Bank. Policy Research Working Paper of the Development Research Group Poverty and Human Resources.

Provus, M.N. (1971). *Discrepancy evaluation.* Berkeley: McCutcheon.

Purkey, S.C., & Smith, M.S. (1983). Effective schools: a review. *The Elementary School Journal, 83*(4), 427–452.

Quinn, R.E., & Rohrbaugh, J. (1983). "Spatial model of effectiveness criteria towards a competiting values approach to organizational analysis", *Management Science, 29,* 363–377.

Ralph, J.H., & Fennessey, J. (1983). Science or reform: some questions about the effective schools model. *Phi Delta Kappan, 64*(10), 689–695.

Rao, C.R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Raudenbush, S.W., & Bryk, A. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401–412.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 271–286). New York, NJ: Springer.

Reynolds, D., Hopkins, D., & Stoll, L. (1993). Linking school effectiveness knowledge and school improvement practice: towards a synergy. *School Effectiveness and School Improvement, 4*(1), 37–58.

Reynolds, D., Sammons, P., Stoll, L., Barber, M., & Hillman, J. (1996). School Effectiveness and School Improvement in the United Kingdom. *School Effectiveness and School Improvement (special issue of country reports), 7*(2), 133–158.

Riddell, A. (1997). Assessing Designs for School Effectiveness Research and School Improvement in Developing Countries, *Comparative Education Review, 41*(2).

Rideout, W.M., & Ural, P. (1993). Centralized and decentralized models of education: comparative studies. Development Bank of Southern Africa. Centre for Policy Analysis.

Riley, D.D. (1990). Should market forces control educational decision making? *American Political Science Review, 84,* 554–558.

Riley, K., Docking, J., & Rowles, D. (1998). *Project on the Role and Effectiveness of the LEA. Report to Lancashire Education Authority.* Centre for Educational Management, Roehampton Institute.

Robertson, P., Toal, D., MacGilchrist, B. & Stoll, L. (1998) *Quality Counts: Evaluating Evidence for School Improvement,* paper presented at the International Congress for School Effectiveness & Improvement, University of Manchester, January 1998.

Roid, G.H., & Haladyna T.M. (1982). *Toward a technology of test-item writing* . NewYork, N.J.: Academic Press.

Rosenshine, B., & Furst, N. (1973). The use of direct observation to study teaching. In R.M.Travers (Ed.), *Second Handbook of Research on Teaching.* Chicago: Rand McNally.

Rowe, K. (1996). *Multilevel Modelling Course Notes.* Melbourne: University of Melbourne, Centre for Applied Educational Research.

Rubin, D.B. (1976). Inference and missing data. *Biometrika, 63,* 581–592.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of schools effectiveness research.* London: OFSTED.

Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: results from a reanalysis of the Inner London Education Authority's Junior School Project data. *British Educational Research Journal, 19*(4), 381–405.

Sammons, P., Nuttall, D., Cuttance, P., & Thomas, S. (1996). Continuity of school effects: A longitudinal analysis of primary and secondary school effects on GCSE performance. *School Effectiveness and School Improvement, 6*(4) , 285–307.

Sammons, P., Thomas, S., & Mortimore, P. (1997). *Forging Links: Effective Schools and Effective Departments.* London: Paul Chapman.

Sammons, P., Thomas, S., Mortimore, P., Owen, C., & Pennell, H. (1994). *Assessing School Effectiveness: Developing Measures to put School Performance in Context.* London: Office for Standards in Education [OFSTED].

Scheerens, J. (1983). *Evaluatie-onderzoek en beleid. Methodologische en organisatorische aspecten* [Evaluation research and public policy-making. Methodological and organisational aspects]. Harlingen: Flevodruk b.v.

Scheerens, J. (1987). *Enhancing educational opportunities for disadvantaged learners.* Amsterdam: North-Holland Publishing Comp.

Scheerens, J. (1990). Beyond decision-oriented evaluation. In: H.J.Walberg & G.D. Haertel (eds.), *The International Encyclopedia of Educational Evaluation* (pp. 35–40). Oxford: Pergamon Press.

Scheerens, J. (1990). Conceptualisering van Schoolmanagement. *Onderwijskundig Lexicon,* editie II. Alphen aan de Rijn: Samsom, F3100–1—F3100–35.

Scheerens, J. (1990). School effectiveness and the development of process indicators of school functioning. *School Effectiveness and School Improvement, 1,* 61–80. Lisse: Swets & Zeitlinger.

Scheerens, J. (1992). *Effective Schooling: Research, Theory and Practice.* London: Cassell.

Scheerens, J. (1994). The school-level context of instructional effectiveness: a comparison between school effectiveness and restructuring models. *Tijdschrift voor Onderwijsresearch, 19*(1), 26–38.

Scheerens, J. (1999). Recent developments in school effectiveness research in the Netherlands. In T.Townsend, P.Clarke, & M.Ainscow (Eds.), *Third Millenium Schools; a world of difference in Effectiveness and Improvement* (pp. 143–159). Lisse: Swets & Zeitlinger.

Scheerens, J. (1999). School self-evaluation: origins, definition, approaches, methods and implementation issues. Paper prepared for a presentation for the Worldbank Effective Schools and Teacher Group.

Scheerens, J. (2002). Conceptual contributions to analyze School Panel Inspections. Contribution to the presentation of the results of the study on school panel inspections in Jamaica, 7 June 2002.

Scheerens, J. (2002). School self-evaluation: origins, definition, approaches, methods and implementation. In: D.Nevo (Ed.), *School-Based Evaluation: An International Perspective,* vol. 8, pp. 35–69.

Scheerens, J., & Bosker, R.J. (1997). *The Foundations of Educational Effectiveness.* Oxford: Elsevier Science Ltd.

Scheerens, J., & Brummelhuis, A.C.A. ten (1996). Process indicators on the functioning of schools: results from an international survey. Paper presented at AERA 1995, New York.

Scheerens, J., & Creemers, B.P.M. (1989). Conceptualizing school effectiveness. In *Developments in School Effectiveness Research.* Special issue of the *International Journal of Educational Research, 13*(7), 691–706.

Scheerens, J., & Praag, B.M.S. van (Eds.) (1998). *Micro-economic Theory and Educational Effectiveness.* Enschede/Amsterdam: Universiteit Twente, Afdeling Onderwijsorganisatie en -management/ Universiteit van Amsterdam, Stichting voor Economisch Onderzoek, 165 p.

Scheerens, J., Stoel, W.G.R., Vermeulen, C.J.A.J., & Pelgrum, W.J. (1988). De haalbaarheid van een indicatorenstelsel voor het basis-en voortgezet onderwijs. Enschede: University of Twente/OCTO.

Scheerens, J., Tan, J-P., & Shaw, R.S. (1999). Monitoring and evaluation in World Bank-financed education operations: current practice and scope for improvement. Washington D.C.: World Bank.

Scheerens, J., Vermeulen, C., & Pelgrum, W. (1989). Generalizibility of Instructional and School Effectiveness Indicators Across Nations. *International Journal of Educational Research, 13*(7), 789–799.

Schön, D.A. (1983) *The Professional Practitioner" Toward a New Design for Teaching and Learning in the Professions.* San Francisco, Cal.: Jossey-Bass

School Curriculum and Assessment Authority (1997). *The Value Added National Project—Report to the Secretary of State.* London: SCAA.

Scriven, M. (1967). The methodology of evaluation. In: R.W.Tyler, R.M.Gagné & M.Scriven (Eds.), *Perspectives on curriculum evaluation.* Chicago: Rand McNally.

Scriven, M. (1991). *Evaluation thesaurus.* Newbury Park: Sage.

Selltiz, C, Wrightsman, L.S., & Cook, S.W. (1976). *Research methods in social relations.* New York, N.Y.: Holt, Rinehart and Winston.

Shalabi, F. (2002). *Effective schooling in the West Bank.* Twente University: Doctoral thesis.

Shermis, M.D., & Burstein, J. (2003). *Automated essay scoring: A crossdisciplinary perspective.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Shi, J.Q., & Lee, S.Y. (1998). Bayesian sampling based approach for factor analysis models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology, 51,* 233–252.

Simon, H.A. (1964). *Administrative Behavior.* New York: Macmillan.

Sireci, S.G., Wainer, H., & Thissen, D. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Slavin, R.E. (1996). *Success for all.* Lisse: Swets & Zeitlinger.

Smees, R., & Thomas, S. (1998). Valuing Pupils' Views About School. *British Journal of Curriculum & Assessment, 8*(3), 8–11.

Smees, R., & Thomas, S. (1999). *Lancashire Case Study of School Self-Evaluation.* Paper presented at the British Educational Research Association Annual Conference, Sussex University, September 1999.

Smith, M.F. (1989). *Evaluability assessment: a practical approach.* Boston: Kluwer Academic Press.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66,* 331–342.

Snow, R.E., & Lohman, D.F. (1989). Implications of cognitive psychology for educational measurement. In R.L.Linn (Ed.), *Educational Measurement* (3rd ed., 263–332). New York: American Council on Education and Macmillan.

Stake, R.E. (1975). *Evaluation the Arts in Education. A responsive approach.* Columbus, Ohio: Merill Publishing Company.

Stallings, J. (1985). Effective elementary classroom practices. In M.J.Kyle (Ed.), *Reaching for excellence. An effective schools sourcebook.* Washington, DC: US Government Printing Office.

Standaert, R. (2000). Inspectorates of Education in Europe—a Critical Analysis. Utrecht, SICI.

Stebbins, L.B., St. Pierre, R.G., Proper, E.C., Anderson, R.R., & Cerva, T.R. (1977). *Education as experimentation: a planned variation model,* Vol. IV-A. An evaluation of Follow Through. Cambridge, Mass.: Abt Associates Inc.

Stern, J.D. (1986). The educational indicators project at the U.S. Department of Education. Washington: Center for Statistics, U.S. Department of Education.

Stiggins, R.J., Griswold, M.M., & Wikelund, K.R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement, 26,* 233–246.

Stringfield, S., & Teddlie, C. (1990). School improvement efforts: qualitative & quantitative data from four naturally occurring experiments in phases III & IV of Lousiana School Effectiveness Study. *School Effectiveness and School Improvement, 1,* 139–166.

Stringfield, S.C., & Slavin, R.E. (1992). A hierarchical longitudinal model for elementary school effects. In B.P.M.Creemers & G.J.Reezigt (Eds.), *Evaluation of Educational Effectiveness* (pp. 35–39). Groningen: ICO.

Swaminathan, H., & Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361–370.

Taeuber, R.C. (Ed.) (1987). Education Data Redesign Project. Special issue of *the International Journal of Educational Research, 11*(4), 391–511.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.
Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.
Thissen, D. (1991). *MULTILOG. Multiple, categorical item analysis and test scoring using item response theory.* Chicago, IL: Scientific Software International, Inc.
Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of IRT models. In P.W.Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 67–113). Hillsdale, N.J., Erlbaum.
Thomas, S. (1995). Considering primary school effectiveness: an analysis of 1992 Key Stage 1 results. *The Curriculum Journal, 6*(3), 279–295.
Thomas, S. (1998). Value Added Measures in School Effectiveness in the United Kingdom. *Prospects, 28*(1), 91–108.
Thomas, S. (2001). Dimensions of secondary school effectiveness: comparative analyses across regions. *School Effectiveness & School Improvement, 12*(3), 285–322.
Thomas, S., & Mortimore, P. (1996). Comparison of value added models for secondary school effectiveness. *Research Papers in Education, 11*(1), 5–33.
Thomas, S., Madaus, G., Raczek, A., & Smees, R. (1998b). Comparing Teacher Assessment and Standard Task Results in England: The relationship between pupil characteristics and attainment. *Assessment in Education, 5*(2), 213–246.
Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997a). Stability and consistency in secondary schools' effects on students' GCSE outcomes over 3 years. *School Effectiveness and School Improvement, 8*(2), 169–197.
Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997b). Differential secondary school effectiveness: examining the size, extent and consistency of school and departmental effects on GCSE outcomes for different groups of students over three years. *British Educational Research Journal, 23*(4), 451–469.
Thomas, S., Smees, R., MacBeath, J.Robertson & Boyd, B. (2001). Valuing Pupils' Views in Scottish Schools. *Educational Research & Evaluation, 6*(4), 281–316.
Thurstone, L.L. (1947). *Multiple factor analysis.* Chicago, IL: University of Chicago Press.
Tiana, A., MacBeath, J., Pedró, F., Scheerens, J., & Thomas, S. *(*1999*). INAP Innovative Approaches in School Evaluation.* Madrid: Universidad Nacional de Education a Distancia.
Troost, N., (Ed.) (2001). *Inspecting in a New Age.* Inspecting in a New Age International Meeting on the Occasion of the 200th Anniversary of the Netherlands Inspectorate of Education. Utrecht, Netherlands, the Netherlands Inspectorate of Education.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43,* 39–55.
Tyler, R. (1950). *Basic principles of curriculum and instruction.* Chicago, IL: University of Chicago Press.
Universiteit Twente, OCTO (1995a). Derde internationale onderzoek rekenen/wiskunde en natuuronderwijs. Vragenlijst voor de schoolleider, populatie 1. Enschede: Universiteit Twente, OCTO.
Universiteit Twente, OCTO (1995b). Derde internationale onderzoek rekenen/wiskunde en natuuronderwijs. Vragenlijst voor de leerkracht, populatie 1. Enschede: Universiteit Twente, OCTO.
Van Amelsvoort, H.W.C.H., & Scheerens, J. (1997). Policy issues surrounding processes of centralization and decentralization in European education systems. *Educational Research and Evaluation, 3*(4), 340–363.
Van Amelsvoort, H.W.C.H., Barzanò, G., Donoughue, C., Gil, R., Mosca, S., Pedró, F., & Scheerens, J. (1998). *Evaluation of Educational Establishments.* Barcelona: Open University of Catalunya. (rapport EEDS-project)

Van Bruggen, J.C. (2001). The Netherlands Inspectorate of Education: Where Are We Now and What Are Our Challenges in 2005? Inspecting in a New Age International Meeting on the Occasion of the 200th Anniversary of the Netherlands Inspectorate of Education. N.Troost. Utrecht, Netherlands, the Netherlands Inspectorate of Education, 26–33.

Van den Wollenberg, A.L. (1982). Two new tests for the Rasch model. *Psychometrika, 47,* 123–140.

Van der Linden, W.J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. *Applied Psychological Measurement, 12,* 201–209.

Van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 53,* 237–247.

Van der Linden, W.J., Meijer, R.R., & Vos, H.J. (1997). Normeringsmethoden voor inspectie-evaluaties. Enschede: University of Twente, Department of Educational Measurement and Data Analysis.

Van der Werf, G., & Driessen, G. (1993). Het functioneren van het voortgezet onderwijs. Kenmerken van scholen en docenten in het eerste leerjaar. Groningen/Nijmegen: RION/ITS.

Van Herpen, M. (1989). Conceptual models in use for educational indicators. Paper for the Conference on Educational Indicators, San Fransisco.

Van Kesteren, J.H.M. (1996). *Doorlichten en herontwerpen van organisatiecomplexen: de ontwikkeling van een methodiek om non-profit organisaties bij te sturen op organisatorische effectiviteit.* Groningen: Rijksuniversiteit Groningen, dissertatie.

Verhelst, N.D., & Glas, C.A.W. (1995). The generalized one parameter model: OPLM. In G.H.Fischer & I.W.Molenaar (Eds.), *Rasch models: their foundations, recent developments and applications.* New York, NJ: Springer.

Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps model to analyze partial credit. In: W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp. 123–138). New York, NJ: Springer.

Verhelst, N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: computer program and manual* Arnhem: Cito, the National Institute for Educational Measurement in the Netherlands.

Verschoor, A.J., & Straetmans, G.J.J.M. (2000). MATHCAT: A flexible testing system in mathematics education for adults. In: W.J. van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp. 101–116). Boston: Kluwer-Nijhoff Publishing.

Verstegen, D.A., & King, R.A. (1998). The Relationship Between School Spending and Student Achievement: A Review and Analysis of 35 Years of Production Function Research. *Journal of Education Finance, 24,* 243–262.

Voogt, J.C. (1989). *'Scholen doorgelicht: een studie over schooldiagnose'* (school diagnosis). Dissertation, Rijksuniversiteit Utrecht.

Wainer, H, Bradlow, E.T., & Du, Z. (2000). Testlet Response Theory: an Analogue for the 3-PL Useful in Testlet-Based Adaptive Testing. In W.J.van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp. 245–269). Boston: Kluwer-Nijhoff Publishing.

Wainer, H. & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15(1),* 22–29.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example. *Applied Measurement in Education, 8,* 157–187.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185–202.

Walberg, H.J. (1984). Improving the productivity of American schools. *Educational Leadership, 41,* 19–21.

Wang, M.C., Haertel, G.D., & Walberg, H.J. (1993). Toward a knowledge base for school learning. *Review of Educational Research, 63,* 249–294.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450.

Webb, E.J., Campbell, D.T., Schwartz, R.D., & L.Sechrest (1966). *Unobtrusive Measures: nonreactive research in the social sciences.* Rand Mc. Nally.

Weeda, W.C. (1986). Effectiviteitsonderzoek van scholen. In J.C.van der Wolf & J.J.Hox (Red.), *Kwaliteit van het onderwijs in het geding.* (About education quality). Publicaties van het Amsterdams Pedogogische Centrum, nr. 2. Lisse: Swets & Zeitlinger.

Weick, K.E. (1969). *The social psychology of organizing.* Reading, Ma.: AddisonWesley Pub.

Weiss, C.H. (1980). Knowledge Creep and Decision Accretion. *Knowledge, Creation, Diffusion, Utilization, 1*(3), 381–404.

Weiss, C.H. (1982). Policy research in the context of diffuse decision making. In D.B.P.Kallen, G.B.Kosse, H.C.agenaar, J.J.J.Kloprogge & M.Vorbeck, *Social science research and public policy-making.* Windsor: NFER-Nelson.

Weiss, C.H., & Bucuvalas, M.J. (1980). Truth tests and utility tests: decision-makers' frames of reference for social science research. *American Sociological Review, 45,* 303–313.

Werf, M.P.C. van der (1988). *Het Schoolwerkplan in het Basisonderwijs* (School development plans in primary education). Lisse: Swets & Zeitlinger.

Wesman, A.G. (1971). Writing the test item. In R.L.Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 99–111). Washington, DC: American Council on Education.

West, M., & Hopkins, D. (1997). *Using Evaluation Data to Improve the Quality of Schooling.* Frankfurt, Germany: ECER-Conference.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa, 70,* 703–713.

Wiggins, G. (1998). Letter to editor. *Educational Researcher, 27*(6), 20–22.

Willms, J.D., & Raudenbush, S.W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of Educational Measurement, 26*(3), 209–232.

Wilson, D.T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, Item statistics, and Item Factor Analysis.* [Computer Software]. Chicago, IL: Scientific Software International, Inc.

Wilson, M., & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika, 58, 85–99.*

Winkler, D.R. (1989). *Decentralisation in education: an economic perspective.* Washington:The World Bank, Population and Human Resource Development

Wolf, R.L. (1990). Judicial evaluation. In: Walberg, H.J. and Haertel, G.D. *The International Encyclopedia of Educational Evaluation.* [76–79] Oxford: Pergamon.

Wolfe, J.H. (1976). Automatic question generation form text: An aid to independent study. *SIGCSE Bulletin, 8,* 104–108.

Woods, R. (1977). Multiple-choice: A state of the art report. *Evaluation in Education: International Progress, 1,* 191–280.

World Bank (1996). Performance monitoring indicators. A handbook for task managers. Washington D.C.: World Bank, Operations Policy Department.

World Bank (1996). Performance monitoring indicators. A handbook for task managers. Washington D.C.: World Bank, Operations Policy Department.

Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23–48.

Wright, B.D., & Stone, M.H. (1979). *Best Test Design.* Chicago, IL: MESA Press University of Chicago.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1997). *ConQuest: Generalized Item Response Modeling Software*. (Computer Software). Australian Council for Educational Research.

Yen, W. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245–262.

Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213.

Yin, R.K. (1981). The case study as a serious research strategy. *Knowledge, creation, diffusion, utilization,* vol. 3.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago, IL: Scientific Software International, Inc.

Index

CONTEXTS OF LEARNING
Classrooms, Schools and Society
ISSN 1384–1181

1. Education for All

   Robert E.Slavin

   1996 ISBN 90 265 1472 7 (hardback)

   ISBN 90 265 1473 5 (paperback)

2. The Road to Improvement. Reflections on School Effectiveness

   Peter Mortimore

   1998 ISBN 90 265 1525 1 (hardback)

   ISBN 90 265 1526 X (paperback)

3. Organizational Learning in Schools

   Edited by Kenneth Leithwood and Karen Seashore Louis

   1998 ISBN 90 265 1539 1 (hardback)

   ISBN 90 265 1540 5 (paperback)

4. Teaching and Learning Thinking Skills

   Edited by J.H.M.Hamers, J.E.H.Van Luit and B.Csapó

   1999 ISBN 90 265 1545 6 (hardback)

5. Managing Schools towards High Performance: Linking School Management Theory to the School Effectiveness Knowledge Base

   Edited by Adrie J.Visscher

   1999 ISBN 90 265 1546 4 (hardback)

6. School Effectiveness: Coming of Age into the Twenty-First Century

   Edited by Pam Sammons

   1999 ISBN 90 265 1549 9 (hardback)

   ISBN 90 265 1550 2 (paperback)

7. Educational Change and Development in the Asia-Pacific Region: Challenges for the Future

   Edited by Tony Townsend and Yin Cheong Cheng

   2000 ISBN 90 265 1558 8 (hardback)

   2000 ISBN 90 265 1627 4 (paperback)

8. Making Sense of Word Problems

   Lieven Verschaffel, Brian Greer and Erik De Corte

   2000 ISBN 90 265 1628 2 (hardback)

9. Profound Improvement: Building Capacity for a Learning Community

   C.Mitchell and L.Sackney

   2000 ISBN 90 265 1634 7 (hardback)

10. School Improvement Through Performance Feedback

    Edited by A.J.Visscher and R.Coe

    2002 ISBN 90 265 1933 8 (hardback)

11. Improving Schools Through Teacher Development. Case Studies of the Aga Khan Foundation Projects in East Africa

    Edited by Stephen Anderson

    2002 ISBN 90 265 1936 2 (hardback)

12. Reshaping the Landscape of School Leadership Development. A Global Perspective

    Edited by Philip Hallinger

    2003 ISBN 90 265 1937 0 (hardback)

13. Educational Evaluation, Assessment, and Monitoring: A Systemic Approach

    Jaap Scheerens, Cees Glas and Sally M.Thomas

    2003 ISBN 90 265 1959 1 (hardback)