

PROMS



august 20-24, 2015, fukuoka, japan

Conference Program

Table of Contents

WORKSHOP SCHEDULE (20 & 21 AUG.)	1
SATURDAY (22 AUG.) SCHEDULE.....	2
SUNDAY (23 AUG.) SCHEDULE	3
MONDAY (24 AUG.) SCHEDULE	4
CAMPUS MAPS.....	5
THURSDAY (20 AUG.) WORKSHOPS.....	8
FRIDAY (21 AUG.) WORKSHOPS	9
SATURDAY (22 AUG.) PLENARIES	10
SUNDAY (23 AUG.) PLENARIES.....	11
MONDAY (24 AUG.) PLENARIES.....	12
SATURDAY (22 AUG.) PARALLEL SESSIONS	13
SUNDAY (23 AUG.) PARALLEL SESSIONS.....	32
MONDAY (24 AUG.) PARALLEL SESSIONS.....	47

20 AUGUST (THU.)



Time	Room 2E405	Room 2W401	Room 2W402	Room 2W403	Room 2W404
8:30 – 9:00	REGISTRATION / COFFEE (Building 2, 4th Floor)				
9:00 – 12:30			WORKSHOP (1-DAY) Rassoul Sadeghi Item banking using the Rasch Measurement Model	WORKSHOP (2-DAY) Trevor Bond A hands-on introduction to Rasch analysis using Winsteps	WORKSHOP (HALF-DAY) Jackson Stenner Causal Rasch Models
12:30 – 1:30	LUNCH (Building 8, 1st Floor)				
1:30 – 4:30			WORKSHOP (1-DAY) Rassoul Sadeghi (Continued)	WORKSHOP (2-DAY) Trevor Bond (Continued)	WORKSHOP (HALF-DAY) James Sick Exploring Many-Facet Rasch Measurement using Facets

21 AUGUST (FRI.)

Time	Room 2E405	Room 2W401	Room 2W402	Room 2W403	Room 2W404
8:30 – 9:00	REGISTRATION / COFFEE (Building 2, 4th Floor)				
9:00 – 12:30			WORKSHOP (1-DAY) Tetsuo Kimura Implementation of small-scale computer-adaptive test (CAT) with open-source software	WORKSHOP (2-DAY) Trevor Bond (Continued)	WORKSHOP (1-DAY) George Engelhard, Jr. & Jue Wang Invariant measurement with raters and rating scales
12:30 – 1:30	LUNCH (Building 8, 1st Floor)				
1:30 – 4:30			WORKSHOP (1-DAY) Tetsuo Kimura (Continued)	WORKSHOP (2-DAY) Trevor Bond (Continued)	WORKSHOP (1-DAY) George Engelhard, Jr. & Jue Wang (Continued)

Time	Room 2E405	Room 2W401	Room 2W402	Room 2W403	Room 2W404
8:30 – 9:00	REGISTRATION / COFFEE (Building 2, 4th Floor)				
9:00 – 10:00	OPENING RECEPTION (Building 2, 4th Floor)				
10:00 – 11:00	PLENARY Trevor Bond The Pacific Rim Objective Measurement Symposia: Where do we go from here?				
11:00 – 11:30	BREAK (Building 2, 4th Floor)				
11:30 – 12:00		Ming-chia Lin & Eric S. Lin Can college-attendance value explain Taiwanese undergraduates' academic performance and expected degree?	Chen-Wei Liu & Wen-Chung Wang An empirical analysis of choice effects in examinee-selected items	Rie Koizumi et al. Developing a scale of interactive oral ability using multi-faceted Rasch analysis and paired oral tasks	Elizar Applying Rasch measurement: Gender and school location differences in mathematics achievement related to higher and lower order thinking
12:00 – 12:30		Norlly Mohd Isa Science process skill assessment: Teachers practice and competency		Mitsuko Tanaka Developing and evaluating a questionnaire instrument for vocabulary learning motivation using Rasch measurement	Che Yee Lye & Michelle Rivera Lacia Validation of the Attitudes Towards the Learning Situation Scale: Confirmatory factor analysis and Rasch analysis
12:30 – 1:00		Daniel Bergh Measuring adolescent perceptions of the school climate: An analysis of the psychometric properties of a scale using Australian adolescent data	A. Adrienne Walker & George Engelhard, Jr. Using person fit and person response functions to examine the validity of person scores	Hsueh-Chu Chen & Kuan-Yu Jin Multifaceted Rasch analysis on perceptual judgments of Chinese-accented speech	Michelle Rivera Lacia Establishing the utility of the Mathematics Teaching Efficacy Beliefs Instrument (MTEBI) for the Philippine context
1:00 – 2:00	LUNCH (Building 8, 1st Floor)				
2:00 – 3:00	PLENARY Rob Cavanagh & William Fisher Measurement: A medium for communication and social action				
3:00 – 3:30	BREAK (Building 2, 4th Floor)				
3:30 – 4:00			Yue Zhao Comparison of relative precision using Rasch-derived and summative scoring approaches	James Sick & Takaaki Kumazawa The effects of multiple-choice item formats on grammar test performance	Mei-Teng Ling & Vincent Pang Gender differential item functioning in Leadership Competency Scale (LCS)
4:00 – 4:30			Kuan-Yu Jin Accounting for global and local dependence among clustered samples using a new Rasch model	Jack Victor Bower Equating three in-house achievement tests of English reading and listening at a Japanese university using Rasch common item equating	Ming-chia Lin et al. Validating research-abstract writing assessment through multi-faceted Rasch-modeling and rater's lenses
4:30 – 5:00		Hui-Fang Chen The scale of reflective process in social work practicum	Haruhiko Mitsunaga & Yuji Nakamura A comparison of equating method based on IRT model for the placement test of EFL course	Brandon Kramer & Stuart Mclean A mixed-methods approach to validating a vocabulary listening test	Nornazira Suhairom et al. Validity and reliability of culinary competency measurement instrument measuring competencies for superior work performance using Rasch measurement model
5:00 – 5:30		Kamal J.I. Badrasawi et al. Oral cancer awareness among secondary school children: Pilot analysis using Rasch measurement model	Paul M. Horness The influence of repetition type on question difficulty	Trevor Allen Holster et al. Measuring extensive reading text difficulty	Pey Tee (Emily) Oon & Kui Foon (Joseph) Chow Psychometric examination of a cross-national assessment on computer literacy using of Rasch framework
7:30 – 9:30	CONFERENCE DINNER Venue: Maruya (まる家・ http://www.f-maruya.jp/)				

23 AUGUST (SUN.)



Time	Room 2E405	Room 2W401	Room 2W402	Room 2W403	Room 2W404
8:30 – 10:00	REGISTRATION / COFFEE (Building 2, 4th Floor)				
10:00 – 11:00	PLENARY George Engelhard, Jr. Invariant measurement in the human sciences				
11:00 – 11:30	BREAK (Building 2, 4th Floor)				
11:30 – 12:00		Jovelyn Gumatay Delosa Validation of the pre-licensure examination for pre-service teachers in professional education using Rasch analysis	Chia-Wen Chen Controlling within-person exposure in computerized adaptive testing for ranking items	Troy L. Cox From standards to rubrics: Comparing full-range to at-level applications of an item-level scoring rubric on an oral proficiency assessment	Claire Campbell & Trevor Bond Constructing the human figure drawing continuum: One scale is good enough
12:00 – 12:30		Pey Shin Ooi Development and validation of tertiary music performance students' motivation scales using Rasch model	Wong Cheow Cher NLMixed procedure to derive the standard errors of true score equating for partial credit tests	Wen-Yen Yang et al. Development and validity of English speaking self-efficacy scale	Yu-Shu Chen & Yuan-Chi Lai The comparison of the unidimensional and multidimensional models: A Rasch model analysis of 3 × 2 achievement goals
12:30 – 1:00		Zuraini Mat Issa Determination of school cooks' knowledge, attitude and practice in preparing healthy school meal using Rasch measurement model	Eric J. Wu & Jin Yan Apply parallel analysis for factor retention with continuous and categorical data psychometrics	Matthew T. Apple A Rasch analysis of the "Four L2 anxieties"	Hoi Yung Leung Student-centered vs teacher-centered assessment in the context of design education
1:00 – 2:00	LUNCH (Building 8, 1st Floor)				
2:00 – 3:00	PLENARY James Sick Rasch and Causality: Incorporating Rasch measures in SEM and multivariate experiments				
3:00 – 3:30	BREAK (Building 2, 4th Floor)				
3:30 – 4:00		Nor Irvoni Mohd Ishar Customer Voice Retaliation (CVR) test: Constructs verification	Yuanyuan Guan & Trevor Bond Examining the impact of genre on the difficulty of listening subskills	Jue Wang & George Engelhard, Jr. A hyperbolic cosine unfolding model for evaluating rater accuracy in writing assessments	Daniel Bergh The psychosomatic problems scale: An analysis of the psychometric properties using Australian adolescent data
4:00 – 4:30		Shereen Noranee Verifying measure of supervisor-rated leader-member exchange (LMX) relationship using Rasch model	Edward Jay Schaefer Identifying rater types among native English-speaking raters of Japanese university students' EFL essays	Nadia Behizadeh & George Engelhard, Jr. Examining the psychometric quality of a modified perceived authenticity in writing scale with Rasch measurement theory	Eric Dionne Optimization of a script concordance test (SCT) in medical education: Rasch vs CTT
4:30 – 5:00	SYMPOSIUM Invited Speakers Progress and impact of PROMS in the Pacific Rim region				
5:00 – 5:30					

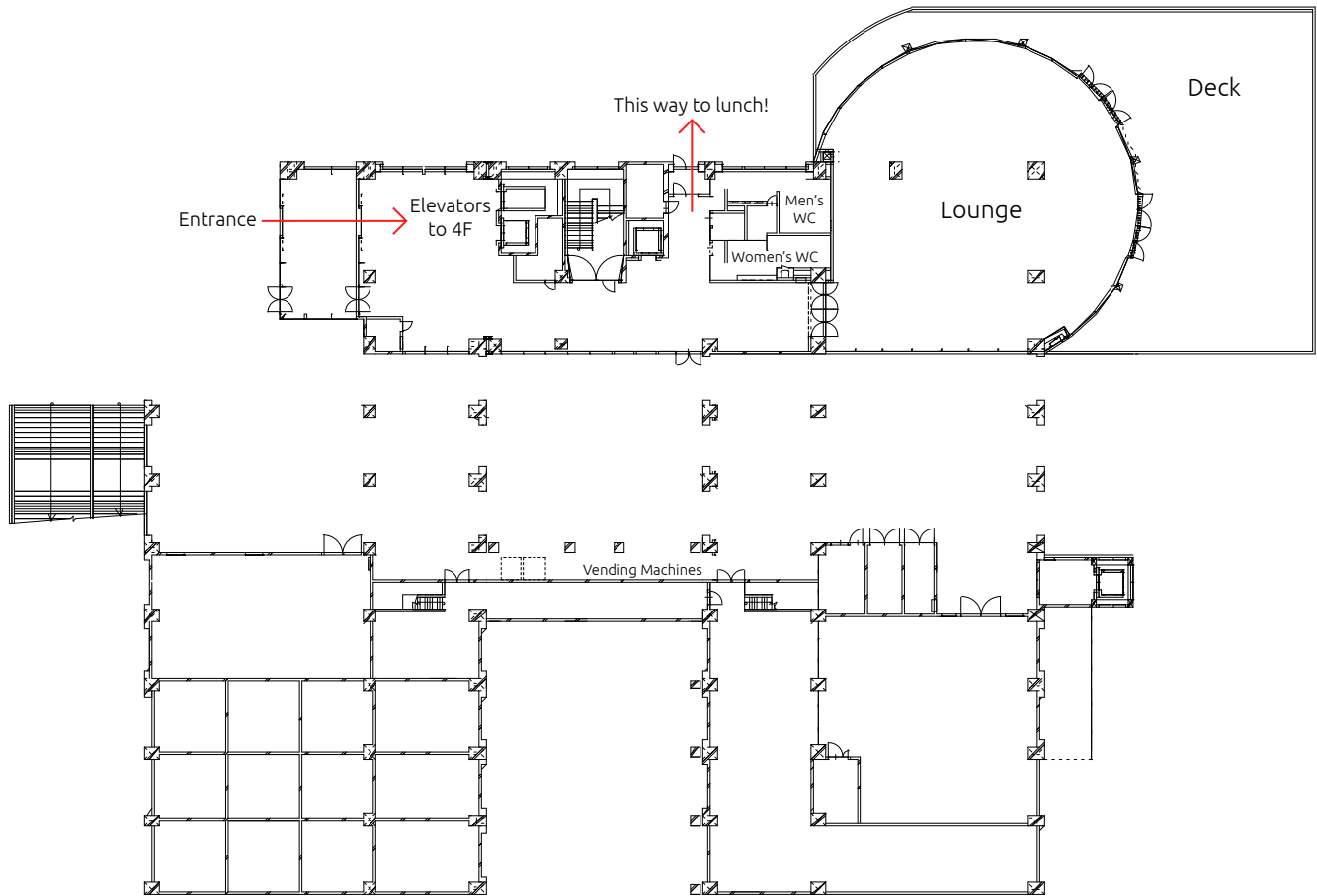
24 AUGUST (MON.)



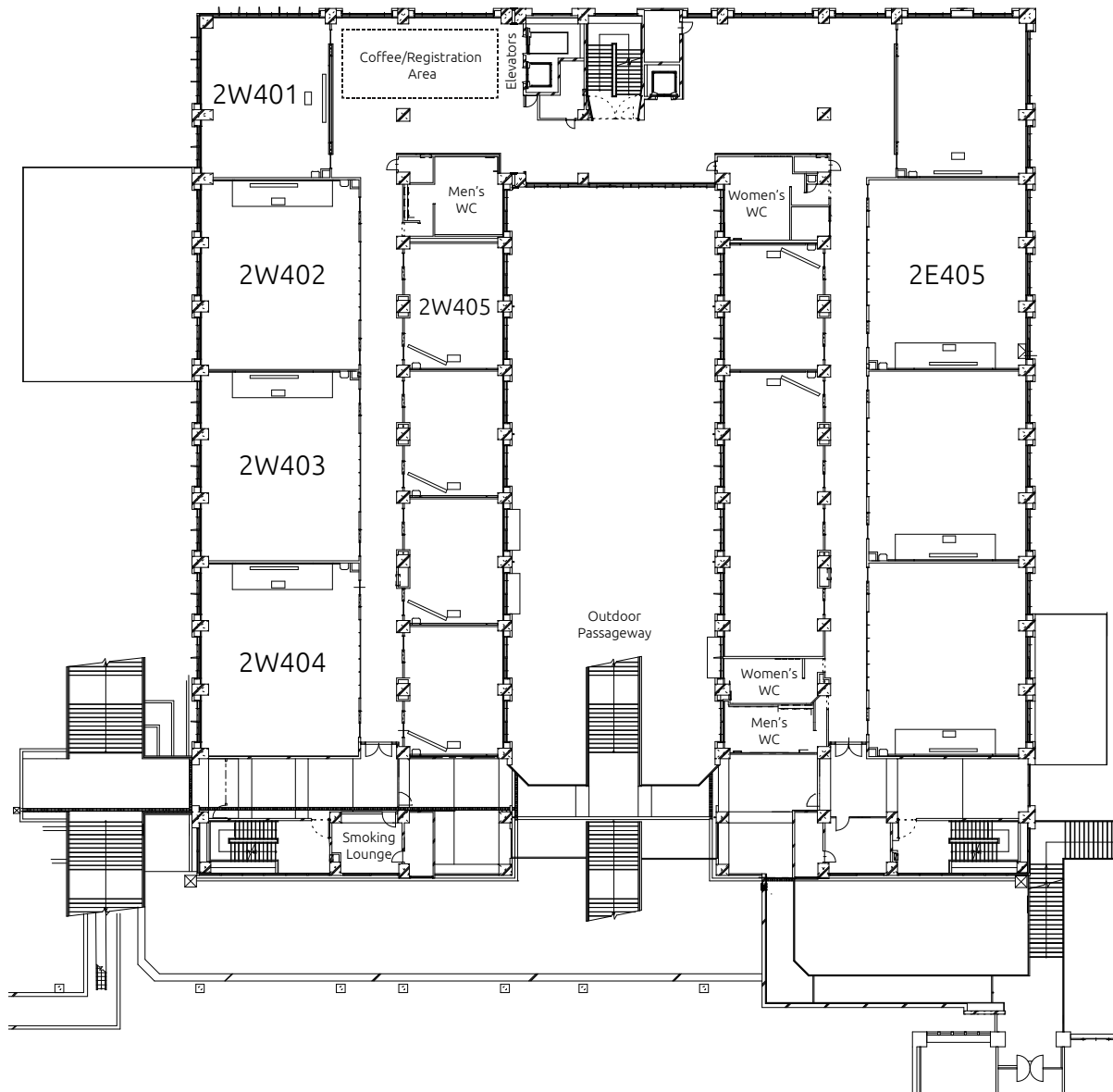
Time	Room 2E405	Room 2W401	Room 2W402	Room 2W403	Room 2W404
8:30 – 10:00	REGISTRATION / COFFEE (Building 2, 4th Floor)				
10:00 – 11:00	PLENARY Tetsuo Kimura Who can be happy with CAT?				
11:00 – 11:30	BREAK (Building 2, 4th Floor)				
11:30 – 12:00		Keita Nakamura Comparability study of different modes of speaking test using the many-facet Rasch model analysis	A.Y.M. Atiquil Islam et al. Factors influencing students' satisfaction in using wireless internet in higher education	Kristy King Takagi Writing assessment in university entrance examinations: The case for indirect assessment	Van Nguyen Effects of field of study background on gender DIF in a university generic skills test
12:00 – 12:30		Graham Robson Use of Rasch to improve a structural model of willingness to communicate (WTC)	Chia-Chi Wang et al. Validation of the Scientific Imagination Test–Verbal	Wei Jie Multidimensional IRT models for L2 computer-based oral english assessment: A comparison between 2PL-PCM and 2PL-RSM	
12:30 – 1:00		Saleh Al-Sinawi & A.Y.M. Atiquil Islam Validation of the employees' service performance scale using Rasch model		Aaron Olaf Batty & Jeffrey Stewart Mitigating the effects of rater severity and examinee familiarity on the Objective Communicative Speaking Test	Vernon Mogol et al. Medical students' approaches to learning: A construct validation from the Rasch perspective
1:00 – 2:00	LUNCH (Building 8, 1st Floor)				
2:00 – 3:00	PLENARY Jackson Stenner Causal Rasch models in language testing: An application rich primer				
3:00 – 3:30	BREAK (Building 2, 4th Floor)				
3:30 – 4:00			J. Lake & Trevor Holster Guessing and the Rasch model	Jinsong Fan Evaluating the validity of the rating scale for an English speaking assessment: An approach combining MFRA and SEM	
4:00 – 4:30			Jeffrey Stewart et al. Examining the vocabulary size test under Rasch and three parameter item response models	Jeffrey Durand Propagation of rater error within a language assessment	
4:30 – 5:00	CLOSING CEREMONY Chair: Robert Cavanagh				

CAMPUS MAPS

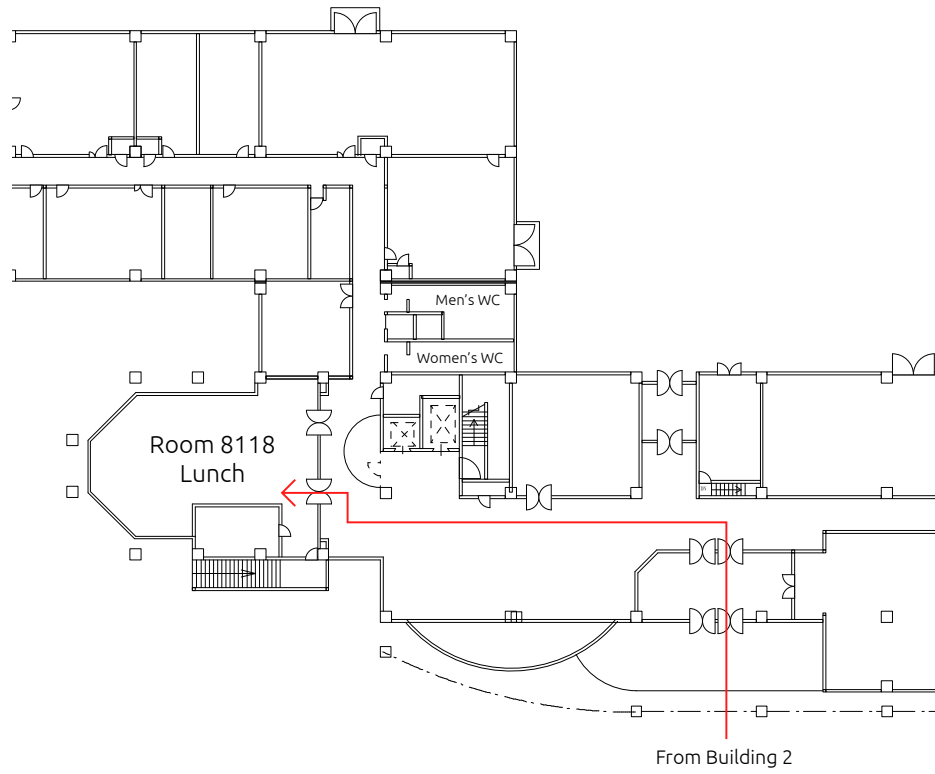
Building 2, 1st Floor (Entrance, Vending Machines, Lounge)



Building 2, 4th Floor (Registration, Presentations, Breaks)



Building 8, 1st Floor (Lunch)



THURSDAY (20 AUG.) WORKSHOPS



Item banking using the Rasch Measurement Model (1-day)

Room: 2W402
Time: 9:00 – 4:30
Instructor: Rassoul Sadeghi

Abstract

Session 1 (9:00 – 10:45) The Rasch Measurement model with an emphasis on its distinctive features

Break (10:45 – 11:00)

Session 2 (11:00 – 12:30) Item banking: applications, advantages and limitations

Lunch (12:30 – 1:30)

Session 3* (1:30 – 3:00) Item banking using the Rasch measurement model 1

Break (3:00 – 3:15)

Session 4* (3:15 – 4:00) Item banking using the Rasch measurement model 2

Session 5 (4:00 – 4:30) Application of item bank in school assessments

* RUMM software will be used to show how an item bank can be created using the Rasch measurement model. See the website for details on how to secure an evaluation copy.



A hands on introduction to Rasch analysis using Winsteps (2-day)

Room: 2W403
Time: 9:00 – 4:30
Instructor: Trevor Bond

Abstract

This workshop provides hands on experience at Rasch analysis using Winsteps software. Prof Bond introduces the rationale for using the Rasch model, and the participants work through a series of guided hands on data analysis exercises. The workshop will focus on the dichotomous Rasch model and the function of fit indices. The application of the Rating scale model to Likert-style data will be introduced. Latest Rasch software will be available for all participants to download. Tutorial worksheets will be available in English / Japanese (some other languages on request).



Causal Rasch Models (Half-day)

Room: 2W404
Time: 9:00 – 12:30
Instructor: Jackson Stenner

Abstract

This workshop introduces the concept of Causal Rasch Models. All measurement shares a three part structure: (1) an attribute such as human temperature or reading ability. (2) a measurement mechanism that transmits variation in the attribute to (3) a measurement outcome such as a count correct on a reading test or a count of cavities turning black on a Nextemp™ thermometer. Causal Rasch Models expose the measurement mechanism and enable direct tests of competing construct theories.

This workshop introduces applications over a wide range of attributes including reading ability, mathematical ability and short term memory. Participants are encouraged to read Causal Rasch Models by Stenner, Fisher, Stone and Burdick (2013) in Frontiers in Psychology. The workshop format will be informal with discussion encouraged.



Exploring Many-Facet Rasch Measurement using Facets (Half-day)

Room: 2W404
Time: 1:30 – 4:30
Instructor: James Sick

Abstract

This half-day workshop will provide a detailed introduction to the many-facet Rasch model by exploring features of the Facets software package. Although a general knowledge of Rasch measurement will be assumed, a brief overview of the many-facet model and its applications will be presented. The session will cover:

1. Formatting data for a Facets analysis,
2. Writing a basic Facets control file,
3. Interpreting standard Facets output, including vertical rulers, measurement tables, and category probability charts

A time-limited edition of Facets will be provided to all workshop participants. Facets is a Windows only program. To use Facets, participants will need to bring a Windows laptop, or a Mac equipped to run Windows. Windows programs can be run on Macintosh computers after installing Windows via Apple Bootcamp, or with emulation software such as Parallels Desktop or VMware Fusion.

FRIDAY (21 AUG.) WORKSHOPS



Implementation of small-scale computer-adaptive test (CAT) with open-source software (1-day)

Room: 2W402
Time: 9:00 – 4:30
Instructor: Tetsuo Kimura

Abstract

This workshop will provides participants with an overview of basic concepts of computer adaptive testing (CAT) and an opportunity to implement a Rasch-based small-scale CAT with open-source software. After briefly discussing key concepts in CAT such as item banking, item selection, target difficulty, maximum information, stopping rules and standard error, two Rasch-based CAT programs will be introduced: UCAT (Linacre, 1987) and Moodle UCAT (Kimura, Ohnishi & Nagaoka, 2012). Participants are encouraged to bring their own notebook PCs so that they can practice building an item bank and creating CATs in the open-source learning management system Moodle. A temporary teacher account on the Moodle UCAT server will be given to each participant.

Session 1: Computer-adaptive testing (CAT)—its origins and concepts

Session 2: UCAT and Moodle UCAT

Session 3: Building an item bank on open-source LMS Moodle

Session 4: Creating CATs on Moodle UCAT



A hands on introduction to Rasch analysis using Winsteps (Day 2)

Room: 2W403
Time: 9:00 – 4:30
Instructor: Trevor Bond

Abstract

(Continuation of Thursday's topics.)



Invariant measurement with raters and rating scales (1-day)

Room: 2W404
Time: 9:00 – 4:30
Instructors: George Engelhard, Jr., Jue Wang

Abstract

The use of rating scales by raters is a popular approach for collecting human judgments in numerous situations. In fact, it is fair to say that raters and rating scales in the social, behavioral and health sciences are ubiquitous. Raters appear in applied settings that range from high-stakes performance assessments in education through personnel evaluations in a variety of occupations to functional assessments in medical research. This workshop utilizes the principles of invariant measurement (Engelhard, 2013) combined with lens models from cognitive psychology to examine judgmental processes that arise in rater-mediated assessments. This workshop focuses on guiding principles that can be used for the creation, evaluation, and maintenance of invariant assessment systems based on human judgments.

The purpose of this workshop is to provide an introduction to the concept of invariant measurement for rater-mediated assessments, such as performance assessments. Rasch models provide an approach for creating item-invariant person measurement and person-invariant item calibration. This workshop extends these ideas to measurement situations that require raters to make judgments regarding performance assessments. This workshop provides an introduction to the Many Facet Model, and its use in the development of psychometrically sound performance assessments. Examples will be based on large-scale writing assessments. Participants are encouraged to bring their own data sets for analysis and discussion in the workshop.

The Facets computer program (Linacre, 2007) is used throughout the workshop to illustrate the principles of invariant measurement with raters and rating scales.

References

- Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York: Routledge.
- Linacre, J. M. (2007). A user's guide to FACETS: Rasch-model computer programs. Retrieved from www.winsteps.com.

Website

www.GeorgeEngelhard.com

SATURDAY (22 AUG.) PLENARIES



The Pacific Rim Objective Measurement Symposia: Where do we go from here?

Room: 2E405
Time: 10:00 – 11:00
Presenter: Trevor Bond

Abstract

The Pacific Rim Objective Measurement Symposia (PROMS) were established in 2005 having grown out of the earlier International Objective Measurement Workshops (IOMW) which are held primarily in the US. The 2004 IOMW workshop was held at the James Cook University campus in Cairns, Australia, and attended by a surprising number of colleagues (many of them, first-timers) from South East Asia. Although PROMS has grown from strength to strength over the subsequent decade, we must turn our endeavours to supporting our Rasch colleagues in the Pacific Rim to consolidate the gains thus far. This address summarizes the progress to date and canvasses a range of strategies for our joint future in PROMS.



Measurement: A medium for communication and social action

Room: 2E405
Time: 2:00 – 3:00
Presenters: Robert Cavanagh, William Fisher

Abstract

In this address we present a conception of measurement aligned with an ‘amodern’ view of science in which the contradictions incumbent within and between positivist and anti-positivist approaches are transcended. An argument is mounted for conceptualising the exercise of measurement as a communicative process that is fundamental to historical and contemporary societies; a process that exploits an invariant language to enable meaningful dialogue across geographical, temporal and socio-cultural boundaries. We explain how amodern measurement theory applying and extending the Rasch Model is crucial for the construction of the measuring instruments that constitute the media connecting individuals and groups. The vital role of metrological networks and traceability for societal development is evident in the way it unleashes human potential and realises living capital.

SUNDAY (23 AUG.) PLENARIES



Invariant measurement in the human sciences

Room: 2E405
Time: 10:00 – 11:00
Presenter: George Engelhard, Jr.

Abstract

The Pacific Rim Objective Measurement Symposia (PROMS) were established in 2005 having grown out of the earlier International Objective Measurement Workshops (IOMW) which are held primarily in the US. The 2004 IOMW workshop was held at the James Cook University campus in Cairns, Australia, and attended by a surprising number of colleagues (many of them, first-timers) from South East Asia. Although PROMS has grown from strength to strength over the subsequent decade, we must turn our endeavours to supporting our Rasch colleagues in the Pacific Rim to consolidate the gains thus far. This address summarizes the progress to date and canvasses a range of strategies for our joint future in PROMS.



Rasch and Causality: Incorporating Rasch measures in SEM and multivariate experiments

Room: 2E405
Time: 2:00 – 3:00
Presenter: James Sick

Abstract

The Pacific Rim Objective Measurement Symposia (PROMS) were established in 2005 having grown out of the earlier International Objective Measurement Workshops (IOMW) which are held primarily in the US. The 2004 IOMW workshop was held at the James Cook University campus in Cairns, Australia, and attended by a surprising number of colleagues (many of them, first-timers) from South East Asia. Although PROMS has grown from strength to strength over the subsequent decade, we must turn our endeavours to supporting our Rasch colleagues in the Pacific Rim to consolidate the gains thus far. This address summarizes the progress to date and canvasses a range of strategies for our joint future in PROMS.



SYMPOSIUM: Progress and impact of PROMS in the Pacific Rim region

Room: 2E405
Time: 4:30 – 5:00
Presenters: Invited Speakers

Abstract

A panel discussion of PROMS' impact on the use of Rasch around the Pacific Rim.

MONDAY (24 AUG.) PLENARIES



Who can be happy with CAT?

Room: 2E405
Time: 10:00 – 11:00
Presenter: Tetsuo Kimura

Abstract

Computer adaptive testing (CAT) is generally considered to be beneficial because of its ability to shorten test length without losses in accuracy. Most CAT algorithms select such items that each test-taker should be able to answer correctly at the 50% chance level, since it maximizes test information and minimize the number of items to be administered. However, many test-takers report feeling discouraged after taking such CATs.

The author collected questionnaire data from test-takers after CATs were administered in order to gain a clearer understanding of how they feel about CAT. According to a questionnaire given to Japanese university students who had taken a CAT English-language placement test, 90% of them found the test “difficult,” with 60% also feeling “discouraged” or “unsatisfied” with the experience. This suggested that the experience of taking a CAT could discourage students, and perhaps even lead to washback effects such as reduced learning self-efficacy and motivation.

Using the same item bank, the author administered two Rasch-based CATs which differed with respect to target difficulty and test length. The first CAT selected 16 items, which the test-takers should have been able to answer correctly at 50% chance level. The second CAT selected 20 items, which they should have been able to answer correctly at 70% chance level. Theoretically, both CATs had the same measurement precision. After each CAT administration, questionnaires were used to discover how they felt about the CATs. The details will be presented at the conference and the trade-off balance between test takers' self-efficacy and measurement efficiency will be discussed so that everyone can be happy with CAT.



Causal Rasch models in language testing: An application rich primer

Room: 2E405
Time: 2:00 – 3:00
Presenter: Jackson Stenner

Abstract

Rasch's unidimensional models for measurement show how to connect object measures (e.g., reader abilities), measurement mechanisms (e.g., machine-generated cloze reading items), and measurement outcomes (e.g., counts correct on reading instruments). Substantive theory shows what interventions or manipulations to the measurement mechanism can be traded off for a change to the measure for an object of measurement to hold the

measurement outcome constant. A Rasch model integrated with a substantive theory dictates the form and substance of permissible interventions. Rasch analysis, absent construct theory and an associated specification equation, is a black box in which understanding may be more illusory than not. Finally, the quantitative hypothesis can be tested by comparing theory-based trade-off relations with observed trade-off relations. Only quantitative variables (as measured) support such trade-offs. Note that to test the quantitative hypothesis requires more than manipulation of the algebraic equivalencies in the Rasch model or descriptively fitting data to the model. A causal Rasch model involves experimental intervention/manipulation on either reader ability or text complexity or a conjoint intervention on both simultaneously to yield a successful prediction of the resultant measurement outcome (count correct). When this type of manipulation is introduced for individual reader text encounters and model predictions are consistent with observations, the quantitative hypothesis is sustained.



CLOSING CEREMONY

Room: 2E405
Time: 4:30 – 5:00
Chair: Robert Cavanagh

SATURDAY (22 AUG.) PARALLEL SESSIONS

Can college-attendance value explain Taiwanese undergraduates' academic performance and expected degree?

Content Area: Education
Room: 2W401
Time: 11:30 – 12:00
Presenters: Ming-chia Lin, Eric S. Lin

Background

In the field of higher education, value of education has been found rather effective in explaining why one emanates motivation to attend and achieve schooling, and to express expectation for further education. Attempting at explaining schooling success, there are two critical values, collective value (e.g., peer influence, career concern) and personal value (e.g., knowledge development, self-exploration).

Aims

The purpose of the study was to construct a college-attendance value scale (CAVS) for undergraduates in Taiwan.

Methods

Data collection involved sophomores ($N = 729$) who voluntarily participated in a schoolwide online survey, including CAVS of the personal value and collective value subscales, expected graduate degree, and Achievement-Goal Questionnaire, alongside cumulative grade point average (CGPA) of each participant.

Sample

Construct validity evidence was substantiated by results of exploratory factor analysis on a calibration sample ($N = 364$) for identification of two subscales, and by results of confirmatory factor analyses (CFA) on a validation sample ($N = 365$) for a good-fit two-factor model.

Results

Moreover in CFA, convergent and discriminant validity evidence supports the distinction of the two subscales. Predictive validity evidence was substantiated from correlations between collective value and personal value scores, achievement-goal scores, cumulative grade point average (CGPA), and expected graduate degree. Results showed achievement-goals as a predictor for CGPA in multiple regression analysis, while personal value as a sole predictor of expected graduate degree in logistic regression analysis.

Conclusions

Findings suggest that CAVS can be used for predictive purposes of Taiwanese undergraduates' academic performance and choices.

Future Directions

Using CAVS, future studies can launch longitudinal studies to further explore the interrelations between CAVS, academic performances, and future study or career pursuit, which may provide fruitful insight into how far college-attendance value contributes to future academic and career development.

An empirical analysis of choice effects in examinee-selected items

Content Area: Technical

Room: 2W402

Time: 11:30 – 12:00

Presenters: Chen-Wei Liu, Wen-Chung Wang

Background

In the examinee-selected (ES) design, respondents are required to respond to a fixed number of items from a set of given items (e.g., respond to 2 of 5 given items). The ES design may enhance students' learning motivation and reduce students' testing anxiety. However, these advantages come at a price: scores obtained from different selection combinations are not comparable. The resulting incomplete data might be missing not at random (MNAR) so that standard IRT models become inappropriate.

Aims

We conducted an experiment with the "Choose one, Answer all" (COAA) design, in which students were instructed to preview the paired items, indicate their preference of the paired items, explain the reasons, and then answered both items. We fit a recently developed class of IRT models to the data to validate the new models.

Methods

To account for choice effects in ES items, we (Liu & Wang, 2014; Wang & Liu, 2015) developed two classes of IRT models. In the first class of models, the choice effects were accounted for by adding a new latent variable to standard IRT model. The correlation between the latent variable and the intended-to-be-measured latent trait quantifies how stronger the choice effect and how serious the violation of the assumption of missing at random are. In the second class of models, those persons showing different selection patterns on ES items were allowed to have different means and variances on the latent variable.

Sample

513 junior students (aged approximately 14) participated in the experiment. The mathematic test consisted of two mandatory and seven pairs of multiple-choice items. The COAA design as adopted.

Results

Because of the COAA design, the item and person parameters could be estimated from the complete data. When the new and traditional IRT models were fit to the incomplete data where unselected items were treated as missing, it was found that the new IRT models had a better model-data fit than traditional IRT models. A significantly positive correlation between the latent variable and the target latent trait was found, indicating a nonignorable choice effect. Additionally, different selection patterns had different distributions on the latent variable for choice effect. The new models yielded parameter estimates that were closer to those obtained from the complete data than traditional models.

Conclusions

The new models were validated by the COAA design. The choice effect was positive and nonignorable. Different selection patterns followed different distributions on the latent variable for choice effects.

Future Directions

It is likely that students sampled from the same school are more homogeneous on the variable of interest (e.g., mathematical proficiency) than those sampled from different schools. In recent years, multilevel IRT models (Fox, 2005; Fox & Glas, 2001; Wang & Qiu, 2013) have been developed to account for such a multilevel data structure. It is of great interest to embed the new IRT models within the multilevel framework.

Developing a scale of interactive oral ability using multi-faceted Rasch analysis and paired oral tasks

Content Area: Language

Room: 2W403

Time: 11:30 – 12:00

Presenters: Rie Koizumi, Yo In'nami, Makoto Fukazawa

Background

Being able to converse in English as a second language is essential for success in a global society, and this skill should be enhanced through teaching and testing. Paired oral tasks that elicit interactions between learners have been found to be a highly valid method to assess interactive oral ability (e.g., Galaczi & French, 2011). However, we can find few investigations of paired oral assessment for Japanese learners of English. This study refines the assessment procedures of our previous study (Koizumi, In'nami, & Fukazawa, 2014) and expands the number of paired oral tasks available that are calibrated on a logit scale in order to improve the task relevancy and representativeness of the intended construct.

Aims

To establish paired oral assessment procedures that can be used in university classrooms, we aim to develop a highly valid scale of interactive oral ability using multi-faceted Rasch analysis. We pose three research questions.

1. Do paired oral tasks with various difficulty levels have high reliability?
2. Do all test takers, tasks, and raters fit the Rasch model?
3. Does the rating scale function properly?

Methods

A multifaceted Rasch measurement program, Facets (Linacre, 2014), was used; the rating scale model was used to examine the test-takers' abilities, task difficulty, rater severity, rating scale functions, and bias patterns.

Sample

A total of 106 Japanese university students from three universities participated in paired oral tasks. They completed all or part of 11 tasks. Their responses were recorded and evaluated by three raters using a holistic scale.

Results

A sample of responses was rated and showed favorable patterns. There was an overall tendency that the paired oral tasks had diverse difficulty. Most test takers, tasks, and raters fit the Rasch model, and the rating scale functioned properly.

Conclusions

The developed paired oral tasks and procedures, and the resulting scale of interactive oral ability, can be supported by validity evidence and used in university classroom oral assessment.

Future Directions

We will further develop and calibrate more tasks on the logit scale so that we can present teachers a pool of paired oral tasks that enable them to select tasks that fit within their teaching context.

Applying Rasch measurement: Gender and school location differences in mathematics achievement related to higher and lower order thinking

Content Area: Education

Room: 2W404

Time: 11:30 – 12:00

Presenter: Elizar

Background

Mathematics assessments need to be designed so as to not disadvantage students on the basis of gender or school location (urban or rural). To date, little research has been done to analyse mathematics assessment in terms of inclusiveness. Differential Item Functioning (DIF) provides a platform for such an analysis, providing an indication of how each item in the assessment behaves across gender and school location.

Aims

The study aims to investigate the DIF of gender and school location in a mathematics assessment related to Higher Order Thinking (HOT) and Lower Order Thinking (LOT). The findings will be used to identify the patterns of students' mathematics achievement related to HOT and LOT across gender and school location.

Methods

Differential Item Functioning (DIF) is used to detect gender and school location differences in mathematics achievement. The testing consisted of four items related to HOT and four items related to LOT. The DIF analysis is performed by ACER ConQuest software.

Sample

The data used in the study are the mathematics test scores of 1135 Year 9 students in 2014 from the Province of Aceh, Indonesia.

Results

Half of the items related to LOT tend to favour males and the other half favour the females. However, in the items related to HOT, 60% favour males. Similar differences also occur for school location, with 60% of the items favouring the urban students.

Conclusions

Gender and school location directions have been drawn for each of the items related to HOT and LOT. The assessment tended to favour males and urban students.

Future Directions

More investigation will be needed to obtain a comprehensive view of reasons for the gender and school location differences in mathematics achievement of students in the Province of Aceh, Indonesia.

Science process skill assessment: Teachers practice and competency

Content Area: Education

Room: 2W401

Time: 12:00 – 12:30

Presenter: Norlly Mohd Isa

Background

Scientific knowledge is gathered and built from the science. Process refers to the way science works when practised by scientists collect and interpret information that is also known as the scientific method. Before using the scientific method scientists must first master the scientific skills. The activities of teaching and learning based on the assessment needs to be improved in order to form meaningful information sharing to enhance the skills of students. To ensure the implementation of practices and assessments carried out by the implementing accurately, teachers must master the concepts related to assessment, evaluation, measurement and testing.

Aims

This study aimed to assess the overall implementation of the practices of Science Process Skills Assessment in the classroom for subjects of science in Malaysia Secondary School. Teachers Science Process Skills Assessment Practice Inventory was developed for this purpose. This inventory consists of three sections on information about teachers' background, training and knowledge on assessment as well as assessment practices implemented by teachers. In addition, this study also explore whether there are significant differences based on gender, teaching experience, and training options in terms of the level of services and Science Process Skills Assessment practices.

Methods

Rasch Measurement Model is used to examine, validate and analyze person and instrument items relating to teachers' assessment practices and competency in Science Process Skills. An assessment practice was measured by the 66 items of five-point Likert scale while Science Process Skills competency was measured by their ability to answer the 28 items correctly or incorrectly.

Sample

The sample for this pilot study made up of 50 science teachers of secondary school from the southern region of Malaysia.

Results

The Rasch analysis showed person reliability index of 0.81 and item reliability index of 0.94. Item fit analysis showed that none of the items needed to be dropped since infit mean square values are between 0.83 and 1.28, and the outfit mean square values are between 0.80 and 1.42.

Conclusions

Majority of the teacher have positive assessment practices even though their Science Proses Skill competency is moderate.

Future Directions

Items in the questionnaires are ordered in a continuum of increasing intensity for the measurement of the teacher competency construct. This shows the validity of the constructs in this instrument.

Developing and evaluating a questionnaire instrument for vocabulary learning motivation using Rasch measurement

Content Area: Language

Room: 2W403

Time: 12:00 – 12:30

Presenter: Mitsuko Tanaka

Background

Self-determination theory (SDT; Deci & Ryan, 2000) is one of the most influential motivational theories in the field of educational psychology. Although numerous researchers have investigated L2 learning motivation using the SDT framework (e.g., Noels, 2001), there is no research examining L2 vocabulary learning motivation from this perspective. As such, there is no SDT questionnaire instrument focusing on L2 vocabulary learning motivation.

Aims

The present study aims to develop and evaluate a SDT questionnaire instrument for vocabulary learning using Rasch measurement.

Methods

First, a Rasch fit analysis was conducted to examine misfitting items and the reliability of each construct measured by the questionnaire. Second, a Rasch PCA of item residuals was conducted to examine the dimensionality of each construct. Third, the original 6-point rating scale was assessed and optimized based on Linacre's (2002) six criteria. Fourth, correlational analysis of the five constructs was performed to examine the simplex-like structure that SDT claims.

Sample

The participants were engineering students from a Japanese technical college ($N = 205$).

Results

The results of the Rasch analyses showed that (1) no items were misfitting, (2) reliability of each construct was adequately high (.59 – .83), (3) an adequate amount of the variance (eigenvalue) was explained by the Rasch model in each construct (41.9 – 66.3% [1.5 – 2.6]), and (4) the original six-category structure was converted into four or five categories for four constructs. The results of the correlation analysis showed that the five constructs have the simplex-like structure that SDT postulates.

Conclusions

The correlation analysis showed that the measurement of the five constructs utilized in this study adequately represents the self-determination theory. The Rasch analysis also showed that each construct is unidimensional and adequately reliable. As such, the developed vocabulary learning SDT questionnaire is valid and reliable.

Future Directions

The instrument was created in order to examine the relationships between motivation, vocabulary knowledge, self-regulation, and self-construal in a Japanese EFL environment. However, it should be adapted and validated in the other L2 learning settings and populations.

Validation of the Attitudes Towards the Learning Situation Scale: Confirmatory factor analysis and Rasch analysis

Content Area: Education
Room: 2W404
Time: 12:00 – 12:30
Presenters: Che Yee Lye, Michelle Rivera Lacia

Background

A reliable and valid instrument is especially crucial for empirical evidence based research. A quality study not only requires a well-designed data collection technique and an appropriate data analysis method, but also requires a psychometrically sound instrument. A similar instrument may have been employed by many researchers in different contexts. It may be reliable and valid in that particular context, but it may not be in other contexts. Therefore, it is essential for researchers to engage in scale development and validation processes.

Aims

The aim of this study was to determine the construct validity of the “Attitudes towards the Learning Situation Scale” by examining its construct at the macro level through Confirmatory Factor Analysis (CFA) and at the micro level through Rasch analysis focusing on person fit and item fit.

Methods

A CFA was conducted to examine the construct validity of the scale. Following the results of the CFA, an initial Rasch analysis based on the rating scale model (RSM) was conducted to examine the person fit. The misfit persons were removed and the second Rasch analysis was carried out, this time examining the item fit. The misfit items were removed and the final structure of the scale was retained for subsequent analysis.

Sample

The study's respondents consisted of 1155 Senior 1 students who were studying the English language subject in the Malaysian Independent Chinese Secondary Schools (MICSS).

Results

The CFA results showed that the best fit model was the 1-factor model with 2 items being removed. This is in contrast to the original model as proposed by Gardner et al. (1997). Analysis of person misfit revealed a total of 22 misfitting respondents. No pattern in the demographics or schools suggested any common characteristics of misfitting respondents. A further analysis of Rasch revealed 2 misfit items. This is consistent with the findings from the CFA.

Conclusions

As evidence by the findings from both the CFA and Rasch analysis, examining the construct validity of a scale in an instrument is of crucial importance. Similar instruments may react differently in different contexts.

Future Directions

Future research should include person fit when examining construct validity so as to develop a psychometrically sound instrument. Future studies examining this scale should be conducted on different samples or contexts to confirm results of this study or of previous studies.

Measuring adolescent perceptions of the school climate: An analysis of the psychometric properties of a scale using Australian adolescent data

Content Area: Education
Room: 2W401
Time: 12:30 – 1:00
Presenter: Daniel Bergh

Background

Adolescents spend a considerable amount of their time in the school environment. Most adolescents are also subjected to compulsory school attendance, implying that they have to deal with the environment on a daily basis. In that sense the school environment is inescapable. There are several different measures on student experiences of the school environment, but School Climate is one of the most prominent. However, there seems to be no agreement upon definition and operationalization of the School Climate concept. Also, it is uncommon to find descriptions of robust psychometric analyses of School Climate measures.

Aims

The purpose of the present study is to examine the psychometric properties of a scale of Adolescent Perceptions of School Climate by means of the Rasch model for ordered response categories.

Methods

A scale consisting of seven polytomous items is analysed by means of the polytomous Rasch model. General fit statistics as well as their graphical representations (ICC) are used to evaluate if the data fit the Rasch model. A particular focus is also directed towards possible Differential Item Functioning (DIF) across sex and grade.

Sample

Using a paper-and-pencil based survey, the data was collected among 758 students enrolled (school year 3 – 7) in schools located in central Perth of Western Australia in 2013.

Results

At a general level of analysis the scale seems to fit the Rasch model fairly well, with good separation of the individuals. Some items showed reversed item thresholds, i.e. the response categories did not work properly and as expected. Also, at a finer level of analysis focusing on DIF, the scale works fairly well, but with exceptions important in order to understand differences between younger and older adolescents.

Conclusions

Although the scale fits the Rasch model fairly well, there is room for improvements. In particular the precision of measurement may be increased by improving the targeting through inclusion of additional items of appropriate severity

Future Directions

As there seems to be a lack of instruments useful for invariant measurement of School Climate across age groups and genders, efforts to develop instruments are required.

Using person fit and person response functions to examine the validity of person scores

Content Area: Technical

Room: 2W402

Time: 12:30 – 1:00

Presenters: A. Adrienne Walker, George Engelhard, Jr.

Background

As the number of computer adaptive tests (CAT) increase around the world, it is important to examine the validity of person scores. Person fit has been used with paper-and-pencil tests, and it offers a promising approach for CAT. The goal is to provide support for inferences regarding how well each person's responses accord with the scores obtained in the CAT system.

Aims

This study explores the validity of individual scores based on a two-stage procedure that includes statistical and graphical aspects. The primary research question focuses on whether or not person fit and person response functions can detect model-data misfit in CAT.

Methods

Person fit is compared across two groups of examinees: those whose responses fit the model and those whose responses do not fit the model. First, three person fit statistics, Unweighted Total (UT), Weighted Total (WT), and Unweighted Between (UB), are used to categorize person response vectors as fitting or misfitting the model. Then, expected and observed person response functions (PRF) are created for selected misfitting and fitting persons. These PRF are visually inspected for discrepancies between model expectations and observed responses.

Sample

A simulation design is used to reproduce an operational test where most persons fit the model, but some persons do not. Five hundred items are generated to represent a unidimensional item bank calibrated with the Rasch model. Dichotomous responses are generated for 5000 examinees drawn from a standard normal distribution.

Results

Preliminary results suggest that visual examinations of PRF aid in the interpretation of UT and WT detected misfit in CAT, but in different ways. The results were inconclusive for UB. The statistical and graphical components provided complementary information about misfit in CAT.

Conclusions

PRF are useful tools for supporting inferences about the validity of individual scores. The patterns of misfit that were found in this study can help practitioners make judgments regarding person fit in CAT, and consequently about appropriate inferences and uses of the test scores.

Future Directions

Future research will include both simulation studies and analyses of operational CAT assessment systems.

Multifaceted Rasch analysis on perceptual judgments of Chinese-accented speech

Content Area: Language

Room: 2W403

Time: 12:30 – 1:00

Presenters: Hsueh-Chu Chen, Kuan-Yu Jin

Background

The term “foreign accent” might be readily characterized as the subjective impressions of a native listener or a learner of a foreign language. The precise nature of foreign accent remains largely unexplored. As listeners (raters) play an important role in assessing foreign accents, rater effects cannot be ignored.

Aims

With the acknowledgement that the observed rating score is exactly a three-way interaction among ratee, rater, and criterion, the multifaceted Rasch model (Linacre, 1989) was employed to analyze the ratings on foreign accent.

Methods

An experiment was conducted, in which 16 recordings were divided into four blocks, and each rater was randomly assigned to two blocks and requested to evaluate the accent of recordings according to five criteria.

Sample

Forty-eight students with different language backgrounds were invited as raters to provide accent ratings.

Results

Among the five criteria for accent ratings, “No foreign accent at all” was the most difficult to attain, while “Very easy to understand” and “This accent is very familiar” were the easiest. The 48 raters exhibited very different degrees of severity ($SD = 0.74$ logit), suggesting rater effects should be carefully considered. By means of the fit statistics, it was found that two raters, one from Hong Kong (outfit = 1.96) and the other from mainland China (outfit = 1.93), could not maintain stable severity throughout the ratings. A mainland China speaker, who exaggeratedly and purposely imitated the accent of English native speakers, received considerably lower ratings from the non-Chinese speakers than from the Chinese speakers (outfit = 2.05).

Conclusions

Fitting the multifaceted Rasch model to rating data in accent studies helps identify aberrant rating behavior and obtain fair measures.

Future Directions

Other than accent, the impression of whether a person can speak a language fluently may be influenced by other factors such like lexical, grammatical and discourse features. Further studies can focus on judgment in terms of these factors and apply the multifaceted Rasch model to the rating data likewise.

Establishing the utility of the Mathematics Teaching Efficacy Beliefs Instrument (MTEBI) for the Philippine context

Content Area: Education
Room: 2W404
Time: 12:30 – 1:00
Presenter: Michelle Rivera Lacia

Background

The concept of self-efficacy used in studies dates back from 1976 when RAND Foundation had constructed two questions about self-efficacy. The results revealed that teachers' sense of self-efficacy had positive strong relations to students' performance, achievement of program goals, and other positive (educational) outcomes (Armor, 1976). These two questions serve as the springboard for several researchers to develop different instruments that measure self-efficacy of teachers. The MTEBI was one of the subject-specific instruments that sprang out from this. In this study, it was administered to different groups of respondents and in different context. Thus, validation of the instrument is warranted. Two of the more advanced techniques in scale validation are the CFA and the Item Response Theory (IRT) using the Rasch Model.

Aims

The purpose of this study was to establish the construct validity of MTEBI developed by Enochs, Smith and Huinker (2000). This further aims to evaluate whether the same model structure fits the Philippines data.

Methods

SPSS 20.0 was used to analyse the internal consistency reliability using Cronbach's coefficient alpha (α). To carry out CFA, LISREL 8.80 by Jöreskog and Sörbom (2006) was employed. Rasch analysis, on the other hand, was performed using Conquest 2.0 (Wu, Adams, Wilson and Haldane, 2007).

Sample

The questionnaire was administered to 326 elementary and secondary mathematics teacher in both private and public schools in Region XII, Philippines.

Results

Using several fit indices as overall model fit to the data, the CFA results show that the structure of both the PMTE and MTOE consists of two correlated constructs. At the item level, the results of the Rasch Analysis suggested that the data best fits the model by removing item 1 from MTOE subscale and item 7 from PMTE subscale.

Conclusions

The result was still consistent with Bandura's (1977) two dimensions of teacher self-efficacy, the self-efficacy expectations and outcome expectancy, only that they branched out into two different but correlated factors. This only show that the Southern Filipino Mathematics teachers have clearly made a line between the positive and negatively worded questions and between the outcomes caused by teachers' effectiveness and effort.

Future Directions

This instrument will be used to gather usable information about the Southern (and other region) Filipino teachers' beliefs of their capacity to teach Mathematics to achieve desirable student's learning outcomes. Examination of the invariance and DIF could likewise be carried out to investigate if the instrument behaves the same way according to gender and grade level.

Comparison of relative precision using Rasch-derived and summative scoring approaches

Content Area: Technical
Room: 2W402
Time: 3:30 – 4:00
Presenter: Yue Zhao

Background

Rasch models have attracted increasing interest in the patient-reported outcome (PRO) measures in recent decades, and carry potential advantages over conventional psychometric models such as classical test theory. There has been however a debate as to the extent to which Rasch-derived scoring could offer greater accuracy and responsiveness than conventional summative scoring in measuring PRO in clinical applications.

Aims

The purpose of this study is hence to compare the discriminatory ability of Rasch-derived scoring and summative scoring in differentiating clinically-diagnosed patient groups in the context of depression measurement.

Methods

A series of relative precisions (RPs), a ratio of pairwise *F*-statistics, were computed based on Rasch-derived scores and summative scores to determine how well each measure discriminated among clinically-defined groups of patients. Statistical significance of RP was quantified by a 95% bootstrap confidence interval.

Sample

A clinical sample of 209 Chinese outpatients seeking treatment for mood and anxiety disorders were invited to complete the Structural Clinical Interview (SCID) and the Hospital Anxiety and Depression Scale (HADS; Zigmond & Snaith, 1983).

Results

Findings from the Rasch analysis demonstrated adequate fit of the HADS items. Findings from the RP index showed that the Rasch-derived scoring was considerably favored over the summative scoring in differentiating various patient groups in the studied context.

Conclusions

In conclusion, Rasch-derived scoring offered more precision in differentiating patient groups than summative scoring in the Chinese clinical sample.

Future Directions

Future directions include the examination of the precision of the Rasch-derived scoring approach over summative scoring in assessing patients' progress over time, and in distinguishing patients with extreme severity from those with mild severity.

The effects of multiple-choice item formats on grammar test performance

Content Area: Language

Room: 2W403

Time: 3:30 – 4:00

Presenters: James Sick, Takaaki Kumazawa

Background

Item formats such as short response, and cloze items are a facet that influences examinees test performance. The most commonly used item format for assessing grammar is multiple-choice (MC), in part because the format is easy to score and can assess test-takers' knowledge of English grammar practically and efficiently.

Aims

In this study, six types of MC item formats were adopted, and 55 multiple-choice grammar items were developed and administered to 608 university students for placement purposes. The research questions were: to what extent do item formats differ in terms of the difficulty and how valid are the items for placement purposes.

Methods

Grammar included 16 features learned up to high school assessed with 6 items formats: fill-in MC, English-Japanese (E-J) translation MC, Japanese-English (J-E) translation MC, error identification MC, correct order MC, and cloze MC. A FACETS analysis was employed with three facets: test-takers, item difficulty, and item format difficulty.

Sample

The test-takers were 608 first-year engineering university students at a private university. Most test-takers were beginner or low-intermediate proficiency levels.

Results

Based on the item analyses, most items functioned well for placement purposes. FACETS (Linacre, 2002) analysis revealed that the six item formats differed in terms of the difficulty in the following difficulty order: cloze MC, error identification MC, correct order MC, J-E translation MC, fill-in MC, and E-J translation MC.

Conclusions

There are a variety of item formats for assessing grammar with MC items. The difficulty of the items formats varied significantly.

Future Directions

The validity of the test score interpretations left remaining questions, as the effects of the grammar features on test-takers' grammar performance were not clearly identified.

Gender differential item functioning in Leadership Competency Scale (LCS)

Content Area: Education

Room: 2W404

Time: 3:30 – 4:00

Presenters: Mei-Teng Ling, Vincent Pang

Background

In Malaysia context, the research of leadership effectiveness or competency are usually on college or University students rather than secondary school student. The gap between male and female leaders success may arise from discrepancies between female gender role and leadership roles. Study by Posner (2012) emphasizes on the instrument to ensure the fairness in leadership measurement.

Aims

The purpose of this study is to explore the differential leadership competency level among Malaysia government secondary school students.

Methods

The instrument used was Leadership Competency Scale known as LCS. The study aims to measure gender bias in LCS by investigating differential item functioning (DIF) across genders. The LCS consists of three constructs; core personality, values and leadership skills. Data gathered were analysed by a Rasch-based item analysis program, Bond&Foxsteps.

Sample

There were 877 male and 1306 female secondary school students involved in this study.

Results

The analysed shows that 30 items (44.1%) from 68 items in LCS show the significance of GDIF in value $t \geq 2.0$ logit. In core personality construct, the items were tended to bias on female respondents, values construct on male respondents and leadership skills construct on female respondents. However, as leadership competency was measured by the three construct, it should be seen as a whole. Overall, LCS shows 14 items bias towards males while 16 items bias towards females. Nonetheless, the contrast between the groups were not serious as the GDIF index showed less than 0.5 logit.

Conclusions

As such, LCS is proposed to be free from GDIF. Therefore, all the items were maintained. In short, it can be used as an indicator to measure leadership competency among students in secondary schools for males and female.

Future Directions

DIF analysis is important to make sure the instrument is valid and less biased to certain group of respondents. Through DIF, items with extreme levels of DIF are identified and omitted. Further studies are needed to understand the different in DIF items according to the respondents' forms, leadership posts and school location.

Accounting for global and local dependence among clustered samples using a new Rasch model

Content Area: Technical
Room: 2W402
Time: 4:00 – 4:30
Presenter: Kuan-Yu Jin

Background

Persons from the same cluster tend to be more homogeneous than those persons from different clusters. Although multilevel models can account for global person dependence (GPD), they are not applicable for local person dependence (LPD). In this study, we developed a new Rasch model to account for both GPD and LPD.

Aims

The new model is called the Rasch model for clustered samples (RMCS), in which a multilevel structure is built on the intended-to-be-measured latent trait to account for GPD, and a random variable across clusters are added to account for LPD for each item.

Methods

A brief simulation was conducted to examine parameter recovery of the RMCS. The computer program WinBUGS was used for parameter estimation. Afterwards, a test of students' civic knowledge was analyzed with the multilevel Rasch model (MRM) and the RMCS.

Sample

The Hong Kong sample of 4,990 students joining the IEA Civic Education Study 1999 was selected for illustration.

Results

As expected, the parameters in the RMCS were recovered fairly well, suggesting the RMCS; ignoring LPD by fitting the MRM yielded a shrunken scale. In the empirical example, it was found that there was a substantial amount of LPD on some items among the students.

Conclusions

The RMCS is feasible and useful for the assessment of LPD among clustered samples.

Future Directions

Further studies can be conducted to develop general models for polytomous responses and to demonstrate their applicability to other data sets.

Equating three in-house achievement tests of English reading and listening at a Japanese university using Rasch common item equating

Content Area: Language
Room: 2W403
Time: 4:00 – 4:30
Presenter: Jack Victor Bower

Background

Three in-house standardized tests of English reading and listening ability, called the Bunkyo English Tests (BETs) are used at Hiroshima Bunkyo Women's University. A major purposes of these achievement tests is to track changes in student English language ability across two years of study in a General English (GE) program. The BET1 is administered to students before they enter the GE program, BET2 is administered at the end of their first year of study, and BET3 is administered at the end of their second year of study. These three tests are intended to be of equivalent difficulty, but it is not currently possible to make inferences about changes in student ability based on the raw test scores. In order to assess changes in student ability across the three tests it is necessary to equate the test forms.

Aims

This study aims to equate the three 2015 BETs.

Methods

Firstly, the reading and listening sections of the BETs are checked for unidimensionality through principal component factor analysis. Secondly, the anchor items are examined for quality. Then shared anchor items are put into a scatter plot. Items lying far from the line of best fit are eliminated until the r-squared value is close to one. Finally, the three tests are equated using common item equating in Winsteps.

Sample

The sample consists test results from 278 finishing first-year students, 302 finishing second-year students, and approximately 300 entering first-year students.

Results

BET2 and 3 anchor items appear sufficient for test equating. There is no statistically significant difference between mean equated item measures on BETS 2&3. Equated BET3 reading section person measures are statistically significantly higher than BET2, but the difference between equated BET2 and 3 listening section mean person measures is not statistically significant. Result from the BET3 administered in April 2015 will also be presented.

Conclusions

Preliminary results show that equating BETs 2&3 is viable. Initial results also lend some quantitative evidence to support a validity argument for the BETs.

Future Directions

Results from equating the 2015 BETs will be used to improve the quality and variety of anchor items, and to build a BET item bank.

Validating research-abstract writing assessment through multi-faceted Rasch-modeling and rater's lenses

Content Area: Education

Room: 2W404

Time: 4:00 – 4:30

Presenters: Ming-chia Lin, Sieh-Hwa Lin, Yuh-Show Cheng

Background

Research-article (RA) writing often poses great difficulties to most second language (L2) graduate students or Taiwanese students. A plethora of research efforts in applied linguistics have been made to tackle these difficulties, such as development in: the genre theory on abstract, introduction-method-results-discussion (IMRD); instructional material for RA-writing; language-support website for RA-abstract writing; and a RA-abstract writing measure. The measure, Research-Abstract Writing Assessment (RAWA), included a timed-task and two rating scales (i.e., a global move of rhetorical purpose, a local pattern of language use). The RAWA appeared severely limited in scope and discipline. The RAWA simply scratched surface of writing competences in key RA-sections (i.e., rhetorical purposes in IMRD) in the RA-abstract task, not examining the competences in various elaborated tasks. The RAWA exclusively targeted the students in applied linguistics as examinees for a discipline-specific prompt, having yet to include students in other disciplines.

Aims

Despite these limitations, it was a fresh attempt at operationalizing the RA-abstract writing into a measure. In validating the RAWA (i.e., a rater-mediated measure), the study aims to investigate how the examinees' responses relate to examinee-ability, rater-severity, scale-difficulty, and score-step by scale via a mixed method that includes qualitative post-rating interview data to supplement quantitative score data.

Methods

The rating scales will be applied to 60 responses of 30 master's and 30 doctoral students. Five raters with varying degrees of expertise will be invited. A total of 600 data-points are planned for the Multi-faceted Rasch Model (MFRM) analysis via FACETS program.

Sample

There will be five raters involved.

Results

The MFRM results will address the study purpose by testing 4 hypotheses. First, the doctoral students outperform the master's. Second, the expert raters demonstrate higher severity than the developing. Third, the global-move

scale entails higher difficulty than the local-pattern. Fourth, there is an interaction effect between score-step difficulty and scale (i.e., Scores 0 to 1 in the local-pattern of lower probability, Scores 4 to 5 in the global-move so). The rater-interview results will reveal major rating strategies and decision-making process. These findings will show how effective the RAWA score is in manifesting the L2 RA-abstract writing competence, supporting the structural aspect of construct validity of RAWA.

Conclusions

The RAWA validation will support the major constructs of the competence that may be decomposed into the global move and local pattern sub-competences.

Future Directions

This validation will have general implications for: (a) theory advancement in L2 RA-writing or L2-writing, (b) measure development in L2 RA-writing, (c) methodology framework for validating rater-mediated measure by considering rater-effect alongside other key factors, and (d) pedagogical insights into how to effectively develop L2 RA-writing by instruction or material compilation.

The scale of reflective process in social work practicum

Content Area: Health

Room: 2W401

Time: 4:30 – 5:00

Presenter: Hui-Fang Chen

Background

The goal of higher education in social work is to help students be ready for the complexities and challenges in real-life practice (Dolan, Canavan, & Pinkerton, 2006; Yip, 2006; Ruch, 2007), and students' reflection ability has been considered an important component in social work education (Oltedal, 2010). Previous literature has noted that the reflective process is critical to social work learning and professional practice, such as the alleviation of negative emotions, development of new perspectives and solutions, and development of professional suitability (Schön, 1993; Yip, 2006). However, there is a lack of quantitative measurement tools that assess students' reflective process in fieldwork practicum.

Aims

The aim of this study was to develop and validate a self-administered tool (termed as the Reflection Process in Practicum Scale, RPPS) to assess students' level of reflection in their practicum.

Methods

The 10-item Reflection Process in Practicum Scale (RPPS) was self-developed. Six experts were invited to review the scale during expert consultation phase, using a 5-point scale (1= Totally disagree to 5 = Totally agree) and to provide insights in terms of item writing and contents. The participants completed the scale during and after practicum. Content validity of the RPPS was examined by six experts. Rasch analyses were conducted to evaluate psychometric properties of the RPPS.

Sample

Students enrolled in a university in Hong Kong and participated in field work practicum in the academic year of 2013-2014 were invited to participate in the present study.

Results

The content validity of the RPPS was confirmed with evidence of high agreeableness among experts. Results suggested that the assumptions of unidimensionality and local item independence were held. All the 10 items showed satisfactory item fit and the reliability was 0.70.

Conclusions

Results showed that the self-developed scale for reflective process was a fairly valid and reliable scale, with one-factor structure. This study may help provide a complete picture about the significance of reflection in social work.

Future Directions

In order to further examine how reflection is related to social work, in future study, reflective practice outcomes can be studied, by examining how they are related to social work, such as the variables as proposed in this study

A comparison of equating method based on IRT model for the placement test of EFL course

Content Area: Technical
Room: 2W402
Time: 4:30 – 5:00
Presenters: Haruhiko Mitsunaga, Yuji Nakamura

Background

The Keio English Placement Test (KEPT) is a placement test which determines the class level of EFL courses for freshmen. KEPT is designed to measure grammar, vocabulary and reading skill using IRT-based scales. We had constructed an item bank before 2012. Since 2013, we have been holding KEPT every April for the placement test (PT) at the beginning of the semester, and the next February for a confirmation test (CT) at the end of the semester. Both PT and CT use items from the item bank which contains item parameters based on the standardized scale.

Aims

This study confirms the best method to equate KEPT 2014 item parameters to the original item bank estimates through the comparison of three equating techniques (mean-mean, mean-sigma and ICC method). Though it appears that there are DIF (differential item functioning) or irrelevant estimators (i.e. threshold parameter exceeds 5.0, etc.) from the parameter estimation result of KEPT 2014, an ideal testing method may provide proper post-equating parameter estimates theoretically.

Methods

Items from a KEPT item bank, whose standardized item parameters were known, was shown the freshmen of 2014 from Keio University and the responses were collected. PT and CT have no common item, but the item bank contains a parameter set which is a standardized scale, so we can compare PT and CT score via item bank common scales. We estimated item parameters assuming a 2PL model using 2014 KEPT data, and estimated post-equating parameters applying three ways of equating methods to equate 2014 KEPT scale to the original KEPT item bank standardized scale for PT and CT separately.

Sample

Freshmen from Keio University, faculty of letters (854 for PT, 794 for CT).

Results

In every measure of PT and CT, ICC method provided closer estimates to the original scale as a post-equating parameter.

Conclusions

ICC method may be the best way to equate 2PL model of KEPT.

Future Directions

IRT assumes local independence in each item. In this study, the effect of local dependence was not taken into account.

A mixed-methods approach to validating a vocabulary listening test

Content Area: Language
Room: 2W403
Time: 4:30 – 5:00
Presenters: Brandon Kramer, Stuart Mclean

Background

An important gap in the field of second language vocabulary assessment concerns the lack of validated tests measuring aural vocabulary knowledge.

Aims

The primary purpose of this study is to introduce and provide preliminary validity evidence for the Listening Vocabulary Levels Test (LVLT), which has been designed as a diagnostic tool to measure knowledge of the first five 1000-word frequency levels and the Academic Word List (AWL).

Methods

Quantitative analyses based on the Rasch model utilized several aspects of Messick's validation framework. Follow-up qualitative interviews assessed the accuracy of the test as a measure of vocabulary knowledge as recommended by Schmitt (1999).

Sample

The test was administered at three Japanese universities ($N = 214$), with 22 students participating in the follow-up interviews.

Results

The findings indicated that (1) the items showed sufficient spread of difficulty, (2) the majority of the items displayed good fit to the Rasch model, (3) items and persons generally performed as predicted by a priori hypotheses, (4) the LVLT correlated with Parts 1 and 2 of the TOEIC listening test at .54, (5) the items displayed a high degree of unidimensionality, (6) the items showed a strong degree of measurement invariance with disattenuated Pearson correlations of .97 and .98 for person measures estimated with different sets of items, and (7) carelessness and guessing exerted only minor influences on test scores. The follow-up interviews and qualitative analyses indicated that the LVLT measures the intended construct of aural vocabulary knowledge, the format is easily understood, and the test has high face validity.

Conclusions

This test fills an important gap in the field of second language vocabulary assessment by providing teachers and researchers with a way to assess aural vocabulary knowledge.

Future Directions

In the future, further validity evidence is necessary to clarify the interpretations of score meaning and its implications on general listening comprehension. Furthermore, bilingual variants in languages other than Japanese should be made with subsequent validation evidence provided before generalizing any results to other populations.

Validity and reliability of culinary competency measurement instrument measuring competencies for superior work performance using Rasch measurement model

Content Area: Education

Room: 2W404

Time: 4:30 – 5:00

Presenters: Nornazira Suhairom, Aede Hatib Mustaamal, Nor Fadila Mohd Amin, Adibah Abdul Latif

Background

The current study emphasize on the development of a comprehensive measurement instrument for workers' competencies. In vocational and technical professions, competency-based assessment throws up some challenges to the professions; however the rewards are potentially very substantial. The creation of a genuinely valid competency-based assessment strategy can yield great benefit, not only to the professions, but to the whole community. Under a competency-based assessment system, assessors make judgments, based on evidence, about whether an individual meets criteria specified in the profession's competency standards. Skilled workers are recognized as quality workers when they have a unity between technical and non-technical competencies.

Aims

The purpose of the current study is to serve as a strong evidence to support the validity of the instrument prior to the actual study. In detail, the specific objectives are to explore the psychometric properties of Culinary Competency instrument and to examine the validity and reliability of the newly developed Culinary Competency instrument.

Methods

The study was conducted using a quantitative survey approach. A survey technique was employed in the data collection utilizing Culinary Competency instrument.

Sample

The Culinary Competency instrument was administered to 114 hotel Chefs, Cooks and Commis who work in the kitchen operations of hotels in Peninsular Malaysia.

Results

The value of item reliability for the Culinary Competency instrument is 0.91 with the item separation index of 3.21. The value for person reliability is 0.98 with person separation index of 8.01. These values indicate that each of the items is highly acceptable as suggested by Bond and Fox (2007).

Conclusions

Analysis of the content validity of the 164 items revealed that several items not demonstrate acceptable goodness-of-fit to the Rasch measurement model, meaning that the respondents' scores on this particular item were inconsistent with their overall response patterns. Tentatively, the Rasch measurement model recommends these items to be deleted or rephrasing. These actions will be executed after considering the study objectives and purpose of measurement. Generally, the Culinary Competency instrument is able to achieve the aims as a good instrument to measure the competencies of the hotels back-of-the-house personnel (Chefs, Cooks and Commis). Analyses of validity and reliability demonstrate that psychometric properties of Star-Chef Competency are good, thus demonstrates the instrument able to produce meaningful measurement.

Future Directions

In an organization, assessment is important as one method to justify a clear standard of employee's ability to perform the best at workplace. Using the right methods and tools, organization are able to ensure they are hiring the right person for the job.

Oral cancer awareness among secondary school children: Pilot analysis using Rasch measurement model

Content Area: Health

Room: 2W401

Time: 5:00 – 5:30

Presenters: Kamal J.I. Badrasawi, N. Abu Kasim, N.L. Abu Kassim, Rosnah Binti Zain

Background

Oral cancer is among the most common cancers worldwide. School children and adolescents are more vulnerable to the main oral cancer risky behaviors: smoking cigarettes, smokeless tobacco use, and alcohol consumption. A school-based intervention is being conducted to examine its effectiveness on the level of school children awareness of oral cancer and its risky behaviors.

Aims

This paper reports the pilot analysis of the survey used to assess the level of school children awareness of oral cancer and its risky behaviors.

Methods

Rasch Partial Credit Model was used to examine the validity and reliability of the survey.

Sample

A convenient sample comprising 63 secondary school children aged 12 – 16 years old completed the survey on oral cancer and its risky behaviors.

Results

The Rasch analysis revealed that after deleting three misfit persons, item reliability was .96 and separation 5.03, and person reliability was .82 and item separation 2.11. This indicated the consistency of the results if the survey given to samples with same characteristics. Unidimensionality was not violated (the items on the first contrast were 2, and Eigenvalue was less than 10%, and the explained variance was 59.3% which is high. All the items were working in the same direction to measure the latent construct as all the item correlation values were positive and ranged between (.53 – .77); and they were productive/ meaningful to the measurement as they had infit statistics values within the recommended range (.7 – 1.3), except item (Know anyone with oral cancer, 1.32) which was closer to the recommended range. Gaps were seen in the middle of the item distribution and at the upper end of the scale.

Conclusions

Overall, the instrument as it stands provides useful measures for the purpose of the study.

Future Directions

The findings will be considered in the real study.

The influence of repetition type on question difficulty

Content Area: Technical
Room: 2W402
Time: 5:00 – 5:30
Presenter: Paul M. Horness

Background

Henning (1991) conducted a study to measure the test-takers' ability to comprehend meaning while varying (a) the effects of memory load through the use of repetitive and non-repetitive aural presentation procedures and (b) passage length. The rationale for using repetition was that it would lessen the memory burden for longer passages and higher-order questions and that listening comprehension would therefore be measured more effectively. One aspect that influences cognitive item difficulty is the interaction between the item type and test-taker.

Aims

The aim of this study was to examine the effect of repetition on question difficulty, lower- or higher-order questions. Similar to Henning's definition, lower-order questions were those that required understanding a specific word within one sentence. Higher-order questions required understanding beyond the word level of sentences; test-takers needed to infer information across one or more sentences.

Methods

250 first-year students at a Japanese university took a 50-item multiple-choice listening comprehension test under one of three conditions: control, massed repetition, and spaced repetition.

Sample

There were 10 listening passages with five multiple-choice questions for each passage. There were 25 lower- and higher-order questions each.

Results

The overall results indicated that higher-order questions were more difficult than lower-order questions. In addition, repetition did not seem to help make the questions significantly easier, but there are indications that repetition on the whole made the items easier.

Conclusions

Repetition had a limited affect on question difficulty relating to higher- and lower-order question types.

Future Directions

Question difficulty was limited two types, so other aspects of difficulty could be examined, specifically TOEFL or TOEIC question types.

Measuring extensive reading text difficulty

Content Area: Language
Room: 2W403
Time: 5:00 – 5:30
Presenters: Trevor Allen Holster, J. Lake, William Pellowe

Background

Extensive reading (ER) aims to increase reading automaticity through processing large quantities of text. This requires matching students to books of appropriate difficulty, implying that both student ability and text difficulty can be measured on a shared scale. The Lexile framework provides such a scale, but was developed for first-language reading, raising questions about its applicability for second-language reading.

Aims

This study aimed to compare Lexile measures of ER texts with other measures of text difficulty in order to find the best predictors of reading difficulty.

Methods

Supported by a grant-in-aid for scientific research from the Japan Society for the Promotion of Science and the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) (Kakenhi grant #25370643), the presenters developed an on-line ER monitoring system, allowing students to rate the difficulty and interest level of

books they read. This allowed many-faceted Rasch analysis to produce logit measures of perceived text difficulty. These were then compared with Lexile measures and other measures of text difficulty.

Sample

Operational data was obtained from 668 Japanese university students, providing logit measures of the difficulty of 1016 texts. Detailed analysis was conducted on 336 popular fiction titles representing a range of publishers and book levels.

Results

Reliability coefficients of .96 for persons and .91 for texts were adequate for the measurement purpose. Lexile measures accounted for 34% of variance in perceived difficulty, compared with 9% for Flesch reading ease scores. However, sentence length alone accounted for 40% of variance, while text length accounted for 62%.

Conclusions

These results support the use of Lexile measures to estimate ER text difficulty, but suggest that the lexical simplification used in ER text production produced texts that are qualitatively different from the authentic English texts used to generate Lexile vocabulary frequency measures. These results suggest that ER texts could be improved through more effective lexical control.

Future Directions

Future analysis will focus on improving the measurement of the lexical difficulty of texts.

Psychometric examination of a cross-national assessment on computer literacy using of Rasch framework

Content Area: Education

Room: 2W404

Time: 5:00 – 5:30

Presenters: Pey Tee (Emily) Oon, Kui Foon (Joseph) Chow

Background

Reliable and valid assessment is essential for international comparative study. The ultimate goal is to position students' ability across countries in a universal scale reliably. One of the aspects that attributed to good quality of assessment lies in its psychometric properties of test use to measure the students' ability. This is undesirable to any cross-national comparative study as this will put the countries' ranking into question.²⁰ The Rasch model has been gaining its popularity in examining psychometric properties of international comparative test items (eg. Glynn, 2012). The Third International Mathematics and Science Study (TIMSS), conducted by IEA, is the first international comparative large scale assessments that fully applied the Rasch model (Wendt, Bos, & Goy, 2011) as the underlying psychometric model. Rasch framework has been gaining popularity in large scale assessment over the classical test theory (CTT) mainly due to its invariance property.

Aims

The goals of this study were twofold. First, overall psychometric properties to the scale were examined. Second, differential item functioning (DIF) will be examined in the items to identify potential sources of DIF.

Methods

Secondary analysis was performed on data garnered from Australia and Korea through the International Computer and Information Literacy Study (ICILS), which is a large-scale international research project under the auspices of the International Association for the Evaluation of Educational Achievement (IEA).

Sample

A total of 8214 students from Australia and Korea participated in the study.

Results

Items showed reasonably good fit to Rasch model but a few items were flagged as DIF items.

Conclusions

A few items may need to be revised for the subsequent study.

SUNDAY (23 AUG.) PARALLEL SESSIONS

Validation of the pre-licensure examination for pre-service teachers in professional education using Rasch analysis

Content Area: Education

Room: 2W401

Time: 11:30 – 12:00

Presenter: Jovelyn Gumatay Delosa

Background

Teacher education is recognized as having an important role in preparing future teachers and in shaping quality education. One of the crucial tasks a teacher education institution has to do is preparing its teachers for the national licensing test and meeting national standards.

Aims

The primary purpose of this study was to provide preliminary validity evidence for a 200-item Pre-Licensure Examination Test for Teachers which is designed to measure the pedagogical knowledge of 152 pre-service teachers on Professional education as preparation for taking the real Licensure Exam.

Methods

Rasch Analysis is used to examine the 200-item test.

Sample

There were 152 pre-service teachers involved in the study.

Results

Out of the 7 subjects, generally, on average the female achieved higher than the males. All the courses got a separation reliability above 0.95.

Conclusions

In conclusion, each of the 7 construct measures what it is supposed to measure and each construct has good psychometric qualities. Three of the constructs (curriculum, educational technology and teaching profession) have all items which have the required fit; social dimension has only 1 item which was under fitting; assessment has 2 "worst" items, social dimension has 1 under fitting item, principle of teaching having 1 under fit item and theories of education with 1 over fitting item. Items of each dimension possess a good discrimination power of separating the student which has a high ability from the less performing students. However putting all the items together to measure the general construct on understanding and application of professional education created a problem thus, there is a need to review the competencies in each construct vis-avis the general goal of professional education as a whole.

Future Directions

Hence, it is recommended that an assessment committee in the School of Education should be created; teachers do pilot testing of the items before these items be included in the final pool of items for Educ 60 and later develop an item bank; review of the LET Primer; review the level of difficulty of the items in terms of order and arrangement in the test; and review of the items which were too difficult that no one answered them and some even skipped them and the easiest items where all the participants were highly able to answer them correctly because the difficulty of these items were far below their ability level.

Controlling within-person exposure in computerized adaptive testing for ranking items

Content Area: Technical
Room: 2W402
Time: 11:30 – 12:00
Presenter: Chia-Wen Chen

Background

Ranking items are so sometimes used, especially in nonability tests. A special case of ranking items is pairwise-comparison items, in which only two statements are ranked. Chen and Wang (2013, 2014) developed the Rasch ipsative model for multidimensional pairwise-comparison items and implemented corresponding computerized adaptive testing (CAT) algorithms. Qiu and Wang (2015) further extended this model to ranking items with more than two statements.

Aims

To increase the feasibility of this model, in this study we aimed to develop corresponding CAT algorithms and investigate their performance. We paid special attention to item selection methods and within-person item exposure control procedures. How to select an item to administer in time is crucial because there are often many dimensions and item sizes are often huge. Our previous attempt was proven successful in pairwise-comparison items (Chen, Wang, & Ro, 2015) and these item selection methods were adapted in this study. It is likely that a person may see a statement many times, which can be very annoying. Thus, control procedures are needed to prevent overexposure of statements to a person.

Methods

In this study, we developed two methods, namely the freeze method and the modified Symptom-Hetter online (MSHO) method. In the freeze method, those statements appearing too many times to a person (e.g., four times) would be frozen. In the MSHO method, an exposure parameter was used to adjust the probability of being administered to a person. We conducted a series of simulations to evaluate the performance of these CAT algorithms.

Sample

The 1500 respondents with six-dimensional latent traits were simulated in CAT with the pool size of 26620 triad-ranking items generated from 66 statements pool.

Results

Results showed that the adapted item selection methods were feasible, the freeze method was effective, and the MSHO method made better usage of the statement than the freeze method, and there was little loss in measurement precision for these algorithms.

Conclusions

MSHO was suggested in the practice since it didn't sacrifice measurement precision a lot and performed better in statement usage than freeze method.

Future Directions

In the future, the real data of CAT need to be collected to examine the relationship between motivation and within person exposure rate.

From standards to rubrics: Comparing full-range to at-level applications of an item-level scoring rubric on an oral proficiency assessment

Content Area: Language
Room: 2W403
Time: 11:30 – 12:00
Presenter: Troy L. Cox

Background

Standards-based proficiency frameworks such as the CEFR are rising in importance in language assessment. The choice of rubric use and its application with performance assessments has important measurement implications.

Aims

This study compares the effect of using a full-range rubric that covers a set of performance standards in their entirety with an at-level or restricted-range rubric that targets the level the specific task is intended to address.

Methods

An expert panel rated a series of prompts to predict their level based on performance standards from which Expert-predicted Item Difficulties (EIDs) were calculated. A speaking test was created with prompts from each EID level and administered to ESL students. The speech samples were rated in an incomplete, spiraled rating schedule. To examine the full-range rubric, the raters used an eight-category scale based on the proficiency scale. To examine the at-level rubric, the ratings were converted to a five-category scale based on the item's intended difficulty level. Those item difficulty parameters were compared to the initial EIDs.

Sample

The analysis of both rubrics was conducted in FACETS with three facets design: examinees ($n = 201$), raters ($n = 10$) and items ($n = 10$). The ratings and scores of the speech samples from both rubrics were analyzed with Rasch measurement to evaluate the functionality of the scales and the separation reliability of the examinees, raters, and items.

Results

Both rating scales performed similarly in terms of functionality and separation reliability with examinee separation reliability being .94 with the full-range and .93 with the at-level. The item difficulty parameters of the full-range rubric did not correlate as expected with the EIDs ($r = -.43$), however the correlation with the at-level rubric was quite strong ($r = .98$).

Conclusions

With a fixed test form that is administered to every student, either method would result in reliable separation of examinees, however the at-level rubric has distinct advantages especially in terms of test equating and development.

Future Directions

Replicating the study with another test and gathering qualitative data from raters using the scales would provide valuable insight.

Constructing the human figure drawing continuum: One scale is good enough

Content Area: Education

Room: 2W404

Time: 11:30 – 12:00

Presenters: Claire Campbell, Trevor Bond

Background

Florence Goodenough's doctoral student, Dale Harris, augmented the original Goodenough Draw-a-Man Test (DAMT) (Goodenough, 1926) to create the Goodenough-Harris Drawing Text (GHDT) (Harris, 1963). The revised GHDT required children to draw an adult female as well as a self-portrait, which is scored against the sex-appropriate DAM or DAW scoring criteria in addition to a drawing of a man.

Aims

The aims of this study were to examine: (1) the psychometric properties of the GHDT from a modern test theory perspective and verify the level of test unidimensionality; (2) the developmental nature of young children's HFD; and (3) the effectiveness of each of the four GHDT sub-tests (DAM, DAW, SPM and SPF) to determine the extent to which each one contributed towards the understanding of the construct.

Methods

All children's drawings were collected, examined and scored in accordance with the GHDT scoring guides (Harris, 1967).

The cross-sectional aspect of the project facilitated the gathering of a broad range of HFD produced by children of different ages and abilities in each phase of data collection. The longitudinal aspect involved three phases of data collection over a 12-month time frame, which was useful for checking the results from the phase one analysis and for investigating the development of children's HFD over time.

Sample

Children ($n = 107$) were recruited from a large Preparatory to Year 12 school in Queensland, Australia (Preparatory, or "Prep," is the name used to describe the first year of full-time schooling prior to Year One in Queensland, Australia). All children were aged within 4 to 10 years, the most appropriate age range for the GHDT (Goodenough, 1926; Harris, 1963), and had informed parental consent to participate in the study. The sample size, whilst comparatively small, was considered sufficient to reflect trends in the data.

Results

Results indicated that the GHDT components were generally psychometrically sound. Consequently, in the interests of parsimony and lessening test-load, a more culturally, socially and educationally relevant prototype Human Figure Drawing Continuum (HFDC) was constructed and examined.

Conclusions

Rasch analysis results revealed that the researcher-developed 45-item HFDC was just as effective as the three component GHDT (217-items in total) and yielded an easier, after and more child-friendly approach to testing.

Future Directions

Future research could involve: replication to investigate whether similar results can be achieved; a larger sample size including children from diverse backgrounds and with diverse needs; and an extended longitudinal aspect that spans longer than 12 months.

Development and validation of tertiary music performance students' motivation scales using Rasch model

Content Area: Education
Room: 2W401
Time: 12:00 – 12:30
Presenter: Pey Shin Ooi

Background

Motivation is an important element in students' learning process. Music performance, in particular, requires students to undertake extensive independent practice in addition to their formal one-to-one tuition. Therefore, students' motivation to persist and engage in their music learning plays a vital role. Although there have been many research studies within the primary and secondary school sectors, research into students' motivation in the field of music performance in the tertiary sector is still lacking.

Aims

To develop and validate an instrument for examining students' motivation within the tertiary education context. This study is also conducted in the hope of laying a foundation for future expansion of the instrument and to promote research in tertiary music education.

Methods

Item analysis based on the Rasch Rating Scale Model was carried out to examine the utility of the motivation scales.

Sample

Tertiary music students who undertake music performance as part of their music courses.

Results

Overall, the statistical results exhibited reasonably good measurement properties. However, there are several items indicating misfit which need attention, either to be revised or removed.

Conclusions

Apart from the misfitting items, the developed motivation scales conform to the measurement requirements of reliability and validity, allowing the measures of these scales to produce meaningful and useful inferences.

Future Directions

It is recommended to revise or replace the misfitting items and conduct similar validation process in any future empirical studies. Differential Item Functioning analysis is also suggested to detect item bias in future study.

NLMixed procedure to derive the standard errors of true score equating for partial credit tests

Content Area: Technical

Room: 2W402

Time: 12:00 – 12:30

Presenter: Wong Cheow Cher

Background

In equating tests, it is informative to know the amount of error in the equated score due to sampling. There are two main approaches to derive standard errors of equating —using re-sampling methods or applying analytical formulas. The former methods (e.g. bootstrap method) involve extensive simulations but are easier to understand, in contrast with the mathematical complexity of the analytical formulas. This complexity is accentuated in equating involving polytomous items, which was recently derived (see Wong, 2015) and validated using the SAS NLMixed procedure.

Aims

This paper aims to make the analytical formulas (see Wong, 2015) to compute true score equating standard errors more accessible to the research community, using a simplified example to illustrate the use of the adapted formulas for the Rasch model.

While this sharing makes use of a particular software and model, the approach is general and could be adopted in other software or models. Scenarios and possible extensions that could make use of this approach, not just for equating, will be discussed in the context of Rasch models.

Methods

A step-by-step guide will be provided, to show how to implement the approach in SAS NLMixed, for tests equated using the Partial Credit Model (PCM). SAS codes, tables and diagrams will be presented, to help researchers relate the procedure to the adapted formulas.

Sample

Simulated tests will be used to demonstrate how to compute standard errors of true score equating, for tests modeled using the Partial Credit Model. SAS NLMixed will be used to calibrate the responses. The intended is to use a simplified example, to make the method more accessible to researchers.

Results

The method has been validated (see Wong, 2015), by comparing empirically derived and analytically derived standard errors. Nonetheless, a replication of the validation process will be conducted for the simplified example used in this paper.

Conclusions

Results shows that the adapted formulas could also apply to the Partial Credit Model.

Future Directions

Scenarios and possible extensions that could make use of this approach, not just for equating, will be discussed in the context of Rasch models.

Reference

Wong, C Cher (2015), Asymptotic Standard Errors for Item Response Theory True Score Equating of Polytomous Items, *Journal of Educational Measurement* Volume 52, Issue 1, pages 106–120, Spring 20

Development and validity of English speaking self-efficacy scale

Content Area: Language

Room: 2W403

Time: 12:00 – 12:30

Presenters: Wen-Yen Yang, Su-Pin Hung, Hung-Yu Huang

Background

Aims

Self-efficacy beliefs determine how people feel, think, motivate themselves and behave. Self-efficacy has been seen as a key factor which may affect English learning motivation (e.g., Gardner, Tremblay, & Masgoret, 1997; Crookes & Schmidt, 1991). Since self-efficacy beliefs are domain specific, the scale of English Speaking Self-efficacy is developed. The present study is aimed to develop a Scale with good reliability and validity to measure English Speaking Self-efficacy of high school students.

Methods

There are 40 items in the English speaking self-efficacy scale. Items with six response categories from 6 “strongly agree” to 1 “strongly disagree.” The Rasch partial credit model was used to assess model-data fit. Moreover, A differential item functioning (DIF) analysis was conducted to assess if any items in the English Speaking Self-efficacy scale favors specific gender group.

Sample

There are three hundred local Taiwanese high school students were recruited under convenience sampling in the present study.

Results

The result shows that most of items fit the PCM fairly well. According Linacre and Wright (1994) suggested that the value of MNSQ around 0.7 – 1.3 can be treated as fit well. Thus, it can be concluded that the English Speaking Self-efficacy scale meet the unidimensional construct. Besides, the scale did not exhibit gender DIF. Finally, mean latent trait of female students was lower than that of the male students on English Speaking Self-efficacy scale.

Conclusions

The results revealed that English Speaking Self-efficacy scale showed appropriate model-data fit and has no differential items.

Future Directions

Suggestions for future research to revise and apply the English Speaking Self-efficacy Scale were proposed.

The comparison of the unidimensional and multidimensional models: A Rasch model analysis of 3 x 2 achievement goals

Content Area: Education

Room: 2W404

Time: 12:00 – 12:30

Presenters: Yu-Shu Chen, Yuan-Chi Lai

Background

While theoretical models of achievement goals thrive, findings of empirical studies of them and their relations to relevant outcomes are inconsistent. Relations between achievement variables and achievement goal differ across studies. While some studies report positive relations of performance goal with performance, others found no relations. Researchers have also found that a mastery goal is positively related to achievement outcomes. However, not all studies have found consistent positive relations. One possibility for these inconsistencies is that they result from limitations in quantifying achievement goals.

Aims

The aims of this study it to conduct both unidimensional and multidimensional rating scale models to examine the newly developed 3 x 2 achievement goal questionnaire (AGQ).

Methods

We used the 18-item achievement goal questionnaire (Elliot, Murayama, & Pekrun, 2011) to assess teacher's achievement goals for teaching task. The questionnaire is rooted in the definition and valence components of competence, and encompasses 6 achievement goals. Teachers completed the AGQ by indicating the extent to which they judge an item was "not at all true of me"=1 to "very true of me"=6. The rating scale models yield estimates of respondents' ability and the difficulties for each item.

Sample

Participants are 694 teachers from junior high schools in Taiwan. Among respondents, there are 68.1% female and 31.9% male. The participants' average age is 37.99-year-old.

Results

The data were analyzed separately for both Unidimensional and Multidimensional models. We compared the fit of two models. The unidimensional has used 23 parameters, and the multidimensional model has used 43 parameters. A formal statistical test of the relative fit of these models was undertaken by comparing the deviance of these two models. The unidimensional model deviance is 4139.18 greater than the deviance for the multidimensional model.

Conclusions

The value is significant which means the unidimensional model is significantly worse than the fit of the Multidimensional model.

Future Directions

Results from the rating scale model analyses revealed a number of issues that need to be addressed in future research. We suggest the AGQ be revised to include more self-approach and self-avoidance items that teachers are less likely to endorse.

Determination of school cooks' knowledge, attitude and practice in preparing healthy school meal using Rasch measurement model

Content Area: Health

Room: 2W401

Time: 12:30 – 1:00

Presenter: Zuraini Mat Issa

Background

School cooks play an important role in ensuring the meals prepared and served to the school children are safe and nutritious.

Aims

The purpose of this study was to examine the level of knowledge, attitude and practice (KAP) of the school cooks at primary schools in Kelantan, Malaysia in preparing healthy school meals. In addition, correlation between the KAP domains was also investigated.

Methods

A validated 47-item self-administered KAP-questionnaire was introduced to 301 school cooks in Kelantan. The dichotomous and polytomous data were analyzed using Rasch measurement model.

Sample

301 school cooks

Results

The results indicate that majority of the school cooks had moderate level of knowledge (50.7%) and a good level of attitude (97%) and practice (88.7%) in preparing healthy meals. Despite having a good attitude and practice, almost one-third (38%) of them had poor knowledge in preparing healthy school meals. The study also detected weak positive correlation between all the three domains investigated (at $p = 0.05$).

Conclusions

As a conclusion, although the cooks were shown to have a very good attitude towards healthy meal preparation, extra effort should be made in educating them related to nutrition and healthy meal preparation education in order to ensure that other foods prepared and sold at the school canteens are also healthy and nutritious.

Future Directions

It is also important to know the dimensions of knowledge items already captured by the school cooks and also to close gaps among the KAP domains.

Apply parallel analysis for factor retention with continuous and categorical data psychometrics

Content Area: Technical

Room: 2W402

Time: 12:30 – 1:00

Presenters: Eric J. Wu, Jin Yan

Background

Parallel Analysis (PA) is a widely studied method for factor retention yet not being routinely adopted. It should be a nice addition to Exploratory Factor Analysis (EFA) and Exploratory Structural Equations Modeling (ESEM). Recent development in commercial and open source software makes it available for practical applications. However, systematic study on how well it recovers the number of factors or latent traits and with what selection criteria has not been thoroughly conducted.

Aims

This proposal attempts to find out how well PA recovers number of factors through simulation with normal and non-normal data. The selection criteria to which these factors are recovered will also be studied.

Methods

A simulation is proposed to create data with CFA and multidimensional IRT (Item Response Theory) type of models with and without cross loadings. The data generation mechanism includes normal data, normal data contaminated with non-normal elements, and binary and graded categorical data. The Parallel Analysis will be used to analyze these data with various selection criteria.

Sample

The data generation will devise different sample sizes, number of factors or dimensions, cross loading vs non-cross loading, normal vs non-normal data. Sample of College English Test-Band 4 in China will also be used to study the dimensionality via simulation and bootstrapping.

Results

Preliminary simulation indicates PA is doing well on recovering number of factors if the factor loading of an indicator is high (i.e. ≥ 0.4) under normal data. The number of factors tends to collapse if the factor loading is small. This behavior will be carefully studied to understand how PA will perform under different model conditions.

Conclusions

The Parallel Analysis seems to be a worthy tool to understand and recover number of factors under specific circumstances.

Future Directions

Combining PA and EFA with various factor extraction and rotation methods could advance this topic further and try to better recover factor structures to form a Confirmatory Factor Analytic (CFA) model.

A Rasch analysis of the “four L2 anxieties”

Content Area: Language

Room: 2W403

Time: 12:30 – 1:00

Presenter: Matthew T. Apple

Background

Foreign language anxiety is considered a situation-specific variable that arises within foreign language contexts, related to but separate from general trait / state anxiety and test anxiety. Traditionally, FLA measurement instruments are “validated” through correlation to existing instruments and split-half reliability analysis using Cronbach’s alpha. However, neither correlational analysis nor Cronbach’s alpha measure construct validity, and items comprising the separate FLA instruments have not been subject to fit analysis.

Aims

The present study aims to investigate (1) the degree to which instruments measuring the “four L2 anxieties” and test anxiety are valid, and (2) whether the four anxieties and testing anxiety are distinct constructs.

Methods

A Likert-type questionnaire instrument with six points was used, with 51 items chosen from five existing FLA instruments. Data were subjected to Rasch Rating Scale Model analysis in WinSteps for Likert category functioning, item fit, and Rasch PCA.

Sample

The questionnaire was distributed online to 315 first and second-year students in 12 intact EFL classes at two undergraduate universities in Kyoto, Japan. Roughly equal numbers of male and female students participated, with an intermediate English proficiency level as measured by the TOEIC ($M = 570$).

Results

Preliminary results showed misfit for the top point of the Likert categories for all of the constructs. Three items additionally misfit their intended constructs. Rasch PCA supported construct validity for the five constructs when measured separately; however, when data from all five constructs were input into the Rasch model simultaneously, speaking and writing anxiety items loaded positively and listening, reading, and testing anxiety items loaded negatively.

Conclusions

While the top performing items of existing FLA instruments overall function well, there is still room for improvement. Whether the “four anxieties” truly measure separate anxieties may still be open for debate; it is perhaps not too surprising that the Rasch PCA indicated the relationship between speaking and writing (i.e., active, interactive skills) and listening, reading, and testing (i.e., the passive skills tested by TOEIC).

Future Directions

Future directions include the winnowing or rewriting of poorly functioning questionnaire items, deeper analysis of the relationship among the “active” and “passive” anxieties, and DIF analysis for gender, major, or English proficiency.

Student-centered vs teacher-centered assessment in the context of design education

Content Area: Education

Room: 2W404

Time: 12:30 – 1:00

Presenter: Hoi Yung Leung

Background

In this empirical study, a class of 25 design students was peer assessed the intended learning outcome in their final assignment in 5 domains. Supporting by the literatures in educational measurement, the arguments from design students and employers that the assessments are subjective and design graduates are lack of creativity. The study discusses the two paradigms of assessment, teacher centered and students centered learning having very diverse interpretation of the best student and ranking in the same class of students.

Aims

The aims of this study are:

1. to explore the difference of a new paradigm shift from teacher centered learning (TCL) to students centered learning (SCL);
2. to discuss the difference of two measurement result from TCL and SCL;
3. to explain in educational measurement approach the reliability and validity of TCL and SCL;

Methods

The methodology of this study in two stages: 1) Applied the Many-facet Rasch model for peer assessment; 2) Compare the ranking of student centered and teacher centered across the same class, teacher and assignment.

$N=25$ design students was assessed in formative assessment in 14 weeks. The final designed product will be assessed in 5 domains during the last lecture. Class teacher and students was evaluated all other design products.

Sample

Researcher is the only class teacher of the 25 design students in a leading design school in Hong Kong.

Results

The pilot study has been found that the reliability of students centered peer assessment is 0.9 in 5-pt scale in 5 assessment criteria. The teacher centered assessment with a single teacher is 0.5. The preliminary finding addressed that students centered approach and assessment has higher reliability in class size assessment.

Conclusions

The finding is supported the paradigm shift of students centered learning.

Future Directions

This study provide new evidence from teachers and educational leaders to review the classroom assessment methods in term of reliability and validity. Our future direction is to investigate how to improve the validity of students as a rater in peer assessment of classroom.

Customer Voice Retaliation (CVR) test: Constructs verification

Content Area: Business

Room: 2W401

Time: 3:30 – 4:00

Presenter: Nor Irvoni Mohd Ishar

Background

Customer complaining behaviour is universal and has received substantial attention over the past decades. Practically, customers expected fair treatment from organization for the effort invested in the relationship. Therefore, perceived unfairness would lead to the impression that they have been betrayed, thus motivate them to respond through complaining as a way of expressing dissatisfaction. In certain cases, they might also resort to performing aggressive complaining to compensate unfairness. For this study, the term customer voice retaliation (CVR) is used to replace customer aggressive complaining. Based on previous literature, a framework was developed to measure CVR which consists of 6 constructs.

Aims

The aim of this study is to verify the newly CVR framework and determine the important constructs for the framework.

Methods

Rasch model was used to examine reliability for both respondents and items. It should give us the list of items that should be included in measuring the CVR constructs.

Sample

Sample used for the study are 50 respondents who have experienced dissatisfaction and have to some extend performed CVR behaviour.

Results

From the analysis, item polarity indicates that all items are measuring in the same direction. Summary statistic indicated that item reliability and item separation is 0.87 and 3.63 respectively, while for person reliability and person separation is 0.81 and 3.57 respectively.

Conclusions

The test conducted indicates that items for measuring CVR needs to be reviewed and instrument construct validity call for further refinement.

Future Directions

Logit measure obtained from the analysis will be used to test mediation effect using smartPLS.

Examining the impact of genre on the difficulty of listening subskills

Content Area: Language

Room: 2W402

Time: 3:30 – 4:00

Presenters: Yuanyuan Guan, Trevor Bond

Background

Genre is a set of communicative events with shared communicative purposes. When audience listens with highly distinctive purposes, it can be assumed that they are likely to apply different cognitive skills to process the spoken input. Although Richards (1983) attempted to develop a hypothetical list of micro-skills in conversational and academic listening, it is questionable whether these generic micro-skills are empirically separable. Moreover, it remains unknown regarding the effect of genre on the difficulty of various listening sub-skills.

Aims

This paper aims to investigate the separability of the difficulty of different listening subskills and their interaction with genres.

Methods

The study is based on the listening component of the DELTA test (Diagnostic English Language Tracking Assessment), which comprises of 207 text-based multiple-choice questions that measure six listening subskills across three spoken genres: daily conversations, radio interviews and academic lectures. The Many-Facet Rasch Model was adopted to calibrate the items and perform interaction analysis with the FACETS software.

Sample

The DELTA test was administered to approximately 2500 students in the 2013 – 14 academic year.

Results

The main analysis revealed that the subskills exhibit good infit mean square values close to 1 and their difficulty measures range from -.67 to .79, with identifying specific information as the easiest subskill and inferring speaker's reasoning as the most difficult. Findings from the interaction analysis suggest that inferring speaker's reasoning scores highest when tested in conversation and lowest in lecture, which has a p -value = .00 of happening by chance. Significant but minor difference was also found in the difficulty of the other subskills relative to different genres.

Conclusions

The results of the study tend to corroborate the assumption that the listening subskills requiring lower-level cognitive processing may pose less challenge than those higher-level subskills which involve interpreting and inferring the implicit meaning based on the message heard. Although slight discrepancies were identified in the subskill difficulties across different genres, the ordering of the subskill measures shows complex and significant pattern.

Future Directions

Implications of these findings with respect to language test design and use in higher education and instruction of listening will be discussed.

A hyperbolic cosine unfolding model for evaluating rater accuracy in writing assessments

Content Area: Language

Room: 2W403

Time: 3:30 – 4:00

Presenters: Jue Wang, George Engelhard, Jr.

Background

Engelhard (1996, 2013) proposed the use of Rasch measurement theory based on the Many Faceted Rasch Model for measuring rater accuracy. Accuracy is defined as the correspondence between observed ratings and benchmark (true) ratings that are assigned by experts. A limitation of previous approaches for quantifying rater accuracy is that they do not differentiate the direction of inaccuracy.

Aims

The current study describes an unfolding model-Hyperbolic Cosine Model (HCM; Andrich, 1997) as a new interpretive framework for examining rater accuracy within the context of rater-mediated assessments.

Methods

A reparameterized form of the Hyperbolic Cosine Model (HCM) with a zone of accuracy parameter is used to examine rater accuracy. Dichotomous accuracy ratings (0 = inaccurate, 1 = accurate) are unfolded into three latent categories by HCM: inaccurate below benchmark ratings, accurate ratings, and inaccurate above benchmark ratings.

Sample

The data analyzed are based on the essays obtained from 8th grade students ($N = 50$) rated by randomly selected raters ($N = 20$) from a large-scale statewide writing assessment. The domain of meaning and style is used in this illustration with dichotomized accuracy ratings (0-inaccurate, 1-accurate).

Results

The HCM provides estimates of the locations of both benchmarks and raters on the underlying unfolded accuracy continuum, and a unit parameter (zone of accuracy) for each rater that can be used to identify the benchmarks having a probability greater than .50 of being rated accurately. It also provides information about the benchmarks that tend to be rated lower than expected relative to the benchmarks (inaccurate below), and higher than expected based on the benchmarks (inaccurate above). Results show different accuracy locations for Raters even though they have same accuracy rates.

Conclusions

The HCM approach for examining rater accuracy provides a useful interpretive framework for evaluating the quality of ratings obtained within the context of rater-mediated assessments.

Future Directions

The idea applying unfolding models to evaluate rater accuracy is new, and we believe that it offers a promising approach for evaluating the quality of rater-mediated assessments. Future research is needed on the deliberate creation of benchmarks that meaningfully represent the underlying continuum.

The psychosomatic problems scale: An analysis of the psychometric properties using Australian adolescent data

Content Area: Education
Room: 2W404
Time: 3:30 – 4:00
Presenter: Daniel Bergh

Background

The PsychoSomatic problems Scale (PSP-Scale) has frequently been used in the Scandinavian countries in order to monitor adolescent psychosomatic health. According to Psychometric analyses based on the the polytomous Rasch model, the PSP-scale shows good measurement properties (see Hagquist, 2008). However, the properties of the PSP-scale have not been examined for non-European samples and for younger adolescents.

Aims

The purpose of the present study is to examine the psychometric properties of the PsychoSomatic Problems Scale by means of the polytomous Rasch model using an Australian sample of younger adolescents (school year 3 – 7).

Methods

The PSP-scale consists of eight polytomous items intended to tap information about student experiences of psychosomatic health complaints. The PSP-scale was analysed by means of the polytomous Rasch model. General fit statistics as well as their graphical representations (ICC) are used to evaluate if the data fit the Rasch model. A particular focus is also directed towards possible Differential Item Functioning (DIF) across school year and sex.

Sample

Using a paper-and-pencil based survey, the data was collected among 758 adolescents enrolled (school year 3 – 7) in schools in central Perth of Western Australia in 2013.

Results

At a general level of analysis, the scale seems to fit the Rasch model fairly well, with good targeting and separation of the individuals. However, some of the items showed reversed item thresholds, indicating that the response categories did not work as expected in the Australian setting. Further there seems to be some tendencies of Differential item functioning by grade.

Conclusions

In comparing the psychosomatic problems among different age groups, in particular younger and older, the analyst needs to be particularly cautious. Also, cultural and language aspects need to be addressed if an instrument is to be used in a different setting than the one it was developed in.

Future Directions

There seems to be a lack of instruments useful for invariant measurement of psychosomatic health among adolescents in different age groups. However, in order to achieve invariant measurement across age groups, efforts to develop instruments are required, in particular if older and younger adolescents are to be compared.

Verifying measure of supervisor-rated leader-member exchange (LMX) relationship using Rasch model

Content Area: Business

Room: 2W401

Time: 4:00 – 4:30

Presenter: Shereen Noranee

Background

An important and unique feature of leader-member exchange (LMX) theory is its emphasis on dyadic relationships. Yet, research on supervisor-subordinate relationships has shown convincingly that leaders do not behave consistently and similarly toward all subordinates. Instead, leaders form different quality relationships with their subordinates. High-quality LMX dyads exhibit a high degree of exchange in superior-subordinate relationships. Subordinates in these dyads are often given more information by the superior and reported greater job latitude. Lower-quality LMX relationships are characterized by a more traditional “supervisor” relationships based on hierarchical differentiation and the formal rules of the employment contract.

Aims

A good and valid instrument helps to determine how clearly leaders behave toward their subordinates. Therefore, the objective of this paper is to verify this instrument, the supervisor-rated LMX of their subordinates.

Methods

The LMX measures were adopted from a 7-item (LMX7) construct of Scandura and Graen (1984) and additional 12 items were adopted from Bernerth, Armenakis, Feild, Giles, and Walker (2007). Several iterations were done by deleting the items identified as misfits. The better instrument was constructed, showing marked improvement across various fit statistics.

Sample

The sampling technique used was by stratified random sampling on 210 supervisors at public universities. Data collected were analyzed using WINSTEPS version 3.72.3.

Results

The results showed a good reliability for both item and person measured at 0.98 (*SE* 0.11) and 0.78 (*SE* 0.54) respectively. The PCA of explained variance improved from 42.7% to 48.3%, determining strong measurement dimension.

Conclusions

Quality control procedures have resulted in an item reduction from 19 to only 9 items, thereby producing a better instrument in measuring the supervisor-rating of LMX of their subordinates.

Future Directions

The results using the revised instrument may yield awareness and understanding among subordinates how their supervisors perceive LMX of them; hence, necessary action can be executed by the employees, supervisors, and the organization, to improve LMX relationship among employees.

Identifying rater types among native English-speaking raters of Japanese university students' EFL essays

Content Area: Language
Room: 2W402
Time: 4:00 – 4:30
Presenter: Edward Jay Schaefer

Background

The bias measurement function of many-facet Rasch measurement (MFRM) has long been used to explore rater bias in EFL speaking and writing tests, and researchers have identified various rater types based on systematic variation in rater-person and rater-category interaction (Kondo-Brown, 2002; Schaefer, 2008). Eckes (2012) criticized Schaefer (2008) for not investigating the reasons for rater bias patterns. He used a combination of MFRM and cluster analysis to explore these reasons, and to identify and label subgroups of raters he called “operational rater types.”

Aims

In a previous study I used MFRM to explore rater bias patterns in NES ratings of Japanese university students' English essays. Following Eckes (2012), the purpose of this study is to reanalyze the data from my previous study using cluster analysis in an attempt to clearly identify rater subgroups among this group of raters.

Methods

Using FACETS, MFRM was performed with 40 student essays and 40 raters rating all the essays. Twenty-four raters showed significant rater-category bias (t -scores of at least ± 2), with 57 significant bias interactions in total. The significant t -scores were input into cluster analysis to determine the existence of different rater types.

Sample

Forty English essays written by Japanese female university students were collected. The raters were 40 Assistant Language Teachers (ALTs) working in the Tokyo area. Each rater rated all 40 essays using a six-category scoring rubric developed by the researcher. The categories were: Content, Organization, Style and Quality of Expression, Language Use, Mechanics, and Fluency.

Results

Cluster analysis was able to identify three rater types, which I have labeled a Rhetorical Features type, a Linguistics Features type, and a Mechanics type.

Conclusions

The use of cluster analysis in combination with the bias analysis function of MFRM seems to be a promising method of studying how the existence of rater bias patterns in EFL writing assessment points to the existence of definite rater types.

Future Directions

Aside from the implications for rater training, another question is whether or not there are pedagogical possibilities for this type of study. If numerous studies identify too many different rater types, then there might be few practical implications. However, if it can be shown that there are a limited number of definite rater types with differential bias patterns toward writing assessment, it may be helpful for EFL students to learn about these types when preparing for high stakes exams such as TOEFL.

Examining the psychometric quality of a modified perceived authenticity in writing scale with Rasch measurement theory

Content Area: Language
Room: 2W403
Time: 4:00 – 4:30
Presenters: Nadia Behizadeh, George Engelhard, Jr.

Background

High-stakes, large-scale testing has proliferated in the United States, and a plethora of studies indicate that instructional practices have suffered (e.g., Darling-Hammond, 2012). In particular, researchers theorize that although US students are writing more, the writing experiences are not highly authentic to students, especially for

urban, minority students who are economically disadvantaged. Student perspectives are needed to explore the authenticity of writing instruction in the United States.

Aims

The purpose of this study is to modify the Perceived Authenticity in Writing (PAW) Scale that was designed to measure perceived authenticity in writing instruction for adolescents (Behizadeh & Engelhard, 2014). The Modified Perceived Authenticity in Writing (MPAW) Scale asks students to evaluate their overall impression of the authenticity of the current writing instruction that they are receiving. The purpose of this study is to examine the psychometric properties the MPAW Scale for use in a larger study of perceived authenticity among minority students in the US.

Methods

Invariant measurement (Engelhard, 2013) is based on Rasch Measurement Theory (Rasch, 1960/1980), and this framework is used to investigate the psychometric quality of the MPAW Scale. The Facets computer program will be used to produce variable maps and model-fit statistics (Linacre, 1989).

Sample

Approximately 100 students at one school site will complete the MPAW Scale during an after-school program this spring. These students mostly identify as Black or African American, 99% participate in free and reduced school lunch programs, and their ages range from 11 to 14 years old.

Conclusions

We may need to alter or drop items based on our analyses.

Future Directions

This study examines the reliability and validity of the MPAW Scale for use with minority students in an urban setting. These analyses will serve as the base for future studies that will examine teacher and student perspectives on writing instruction and assessment.

Optimization of a script concordance test (SCT) in medical education: Rasch vs CTT

Content Area: Education
Room: 2W404
Time: 4:00 – 4:30
Presenter: Eric Dionne

Background

Script concordance tests (SCT) are used in medical education to assess clinical reasoning and decision-making capacity. This instrument is made of a vignette, which presents a clinical case and different questions regarding the case. In each question, there is a hypothesis and a new piece of information about it. After they consider the new information, subjects must position themselves regarding the initial hypothesis. To determine their result, the researcher compares their judgment to the one made by a panel of experts. In the validation process of a SCT, researchers try to optimize the test by removing the items that seem inadequate. In most studies, the classical test theory is utilized to operate this optimization, and practically no research has used the Rasch model to accomplish such a task.

Aims

The objective of this study is to determine whether there are differences between the optimization of items in a SCT when the researcher uses the classical test theory and the Rasch model.

Methods

The metric properties of SCTs have been studied through the classical test theory (CTT) by examining the most common statistics: the item-total correlation, the internal consistency, etc. For this study, we have utilized the partial credit Rasch model (Masters, 1982). Once we verified the assumptions for the use of this model, we examined the person-item map, the item and person fit (infit and outfit) and the ordering of categories. Lastly, we compared our Rasch and our CTT analyses regarding the items that have been discarded.

Sample

Our data comes from a SCT that was created at the University of Liège in Belgium and completed by students in a medical training program of the same institution. The instrument was administered once a year from 2010 to 2014 and was completed by a total of 160 students in medicine.

Results

The preliminary results show that there are important differences in the optimization process of SCTs if the CST or the CTT is used.

Conclusions

The optimization of CSTs seems to depend upon the measurement model that is utilized.

Future Directions

This exploratory study indicates that it would be pertinent to continue our analyses to determine the conditions under which the Rasch model would be more adequate to analyse SCT scores. We also believe that such analyses would help researchers in the decision-making process when it comes to the optimization of this type of instrument.

MONDAY (24 AUG.) PARALLEL SESSIONS

Comparability study of different modes of speaking test using the many-facet Rasch model analysis

Content Area: Language

Room: 2W401

Time: 11:30 – 12:00

Presenter: Keita Nakamura

Background

In the field of language testing, there have been many previous studies conducted on the issue of comparability of tests utilizing different mode of operation. The main focus of those studies was to investigate the comparability of the results from Paper-based test and Computer-based test. The results were not homogenous showing the variability of scores derived from different modes of tests mainly because of PC familiarity. In addition, not so many studies exist regarding the comparability of the results of speaking test derived from different modes.

Aims

In this study, newly developed computer-based speaking test was examined in terms of the comparability of the result against the result of the currently operationalized speaking test. These two modes of tests are planned to be operationalized complementarily such that test takers who cannot take the currently operationalized speaking (PBT) would be able to take CBT speaking.

Methods

All test takers took both PBT and CBT, but in different order. One group went through the experiment starting from the PBT speaking, while the other group started from CBT speaking. Different test prompts were used for the PBT and CBT. In order to take the examiner's rating severity and speaking test prompt difficulty into consideration before calculating test takers' ability measures, multi-facet Rasch model analysis was used in this study. One-way repeated ANOVA was conducted with mode (CBT, PBT) as the independent and test takers' ability measures as the dependent variables. In addition, test takers were asked to respond to a set of questionnaire after the experiment session.

Sample

109 Japanese EFL learners were recruited in this study as test takers. They were randomly assigned to one of the two groups.

Results

The result showed that there was not statistically significant difference between the test takers' ability measures. However, questionnaire result showed that test takers prefer PBT to CBT.

Conclusions

These results, together with other questionnaire survey results, suggested some points of revision to be made to CBT speaking.

Future Directions

The implications are discussed in terms of the actual method to conduct similar studies to validate a testing program.

Factors influencing students' satisfaction in using wireless internet in higher education

Content Area: Education

Room: 2W402

Time: 11:30 – 12:00

Presenters: A.Y.M. Atiquil Islam, Xiuxiu Qian, Hai Leng Chin

Background

In its endeavor to foster the use of technology in higher education, the Jiaying University (JU), Zhejiang Province, P.R.China, provides wireless internet facility to cater to the needs of its students. However, since its execution, there has been no research conducted to assess the successful integration and satisfaction in using of this emerging educational technology within its environment, therefore, creating a gap for investigation.

Aims

In so doing, the aim of this study is to validate the Technology Satisfaction Model (TSM) to examine factors that influence students' satisfaction in using wireless internet in higher education for their learning purpose.

Methods

The instrument's reliability and validity were established through a Rasch model using Winsteps version 3.94. The TSM was validated applying Structural Equation Modeling using AMOS version 18 to test the hypotheses as well as the casual relationships among the constructs.

Sample

A total of 283 students from five colleges (Foreign Studies, Business, Education, and Mathematics and Information Engineering) were collected for this study.

Results

The results of this study revealed that students' satisfaction was directly influenced by perceived usefulness and perceived ease of use of wireless internet. Besides this, students' perceived ease of use and perceived usefulness of wireless internet were directly affected by their computer self-efficacy. On the other hand, students' computer self-efficacy had a significant indirect influence on their satisfaction mediated by perceived ease of use and perceived usefulness, respectively.

Conclusions

The TSM was viable to measure students' satisfaction in using wireless internet in a different culture.

Future Directions

The findings of TSM also suggested that the model could be applied by future researchers, practitioners and academicians to measure new areas in detail such as digital library services, online education, online shopping, mobile learning, social networks, and any other innovative technological services.

Writing assessment in university entrance examinations: The case for indirect assessment

Content Area: Language
Room: 2W403
Time: 11:30 – 12:00
Presenter: Kristy King Takagi

Background

The current project was a follow-up to the writer's 2014 PROMS presentation highlighting the limitations of direct writing assessment, specifically in terms of rubric and rater performance. Because of the limitations of direct writing assessment, the object of the current investigation was to examine the potential value of indirect writing assessment for placement purposes.

Aims

The aim of the current project was to develop a multiple-choice objective test of writing that could serve as a substitute for direct assessment. It was hypothesized that the indirect objective test, which focused on knowledge of sentence form, would not only function well as a test, but also demonstrate validity as a writing test.

Methods

In order to test these hypotheses, the fit, difficulty, and reliability of the test were assessed using the Rasch model. In order to verify whether the indirect objective test demonstrated validity as a writing test, correlations between indirect test scores and essay ratings (analytic and holistic) were examined. A Principal Components Analysis was also conducted to examine whether the indirect test scores loaded together with essay ratings onto the same component.

Sample

The participants were 50 female freshman students at a university in the Kanto area in Japan; 45 were Japanese and 5 were Chinese exchange students. All were approximately 19 years of age and had studied English for seven years.

Results

Results indicated that the indirect writing test performed well as a test. The items moved generally from easier to more difficult, and most items discriminated well between high and low proficiency students. The indirect test scores also demonstrated criterion-related validity, as evidenced by their correlations with both analytic and holistic essay ratings, as well as by PCA loadings.

Conclusions

An indirect test of writing has many obvious advantages for writing placement. It can be an efficient and reliable supplement or substitute for traditional direct tests of writing.

Future Directions

In order to flesh out the construct of writing ability, more study of the relation of indirect and direct measures of writing to each other, and to other measures of language proficiency seems warranted.

Effects of field of study background on gender DIF in a university generic skills test

Content Area: Education
Room: 2W404
Time: 11:30 – 12:00
Presenter: Van Nguyen

Background

uniTEST has been developed by the Australian Council for Educational Research to assist universities with the often difficult and time consuming processes of student selection. The test has been developed to assess generic reasoning and thinking skills that underpin studies at higher education and that are needed for students to be successful at this level. The test consists of three components of 30 multiple-choice items each—Quantitative Reasoning, Critical Reasoning and Verbal-Plausible Reasoning.

Aims

Together with demographics backgrounds, examinees' field of study can be a significant predictor for their performance on generic skill university tests (Hambur, Rowe, & Le, 2004). In this study, gender DIF in uniTEST was examined in relation to field of studies.

Methods

The analysis was conducted at both test score level and at item level. At the item level, candidates were classified into four groups by gender and by two main fields of studies: Psychology and Medicine. Using the Rasch model, item calibrations were taken for each group, and then gender DIF was computed in each the field of studies separately and being compared to each other.

Sample

Data used included for about 1800 Demark university candidates in uniTEST 2014, with 33% males and 67% females.

Results

The results showed that field of studies could affect gender difference in test scores. However, the gender DIF patterns looked consistently. No item showed significantly favouring males in this field of study but significantly favouring females in the other field of study and vice versa.

Conclusions

Results from this study showed that uniTEST gave a fairly discrimination by candidates from different groups of gender and field of study backgrounds. The field of study backgrounds did not yet demonstrate a significant effect on gender DIF.

Future Directions

Multiple regression or Multilevel analysis will be taken for uniTEST scores in relation to such factor as gender, age, and field of study backgrounds when the data sample is large enough.

Use of Rasch to improve a structural model of willingness to communicate (WTC)

Content Area: Language

Room: 2W401

Time: 12:00 – 12:30

Presenter: Graham Robson

Background

Many structural models in second language have incorporated WTC (Willingness to Communicate), but these have tended to focus mainly on the use of more general trait measures to make such models. I believe that the classroom is the most important venue for providing communicative opportunities in the foreign language setting in Japan. Therefore, any subsequent model of WTC should use measures of constructs like perceptions and motivations that are tied to what happens in the class.

Aims

The aims of this study are to show that Rasch and FA can be used to analyze constructs that go into a model of situational WTC.

Methods

I employ FA and Rasch analysis to investigate the reliability and unidimensionality of the factors for the model. In also incorporate Rasch measures into a structural model of those factors.

Sample

461 Japanese university students from a middle-rank university.

Results

Rasch and FA both helped to establish reliability and unidimensionality of the factors. However, the structural model based on raw data did not converge. It was only after using Rasch measures for the structural model that convergence took place.

Conclusions

Rasch and FSA helped to identify which parts of the constructs worked well and which didn't as well as providing a plausible model of situational WTC.

Future Directions

Use of Rasch to investigate other important second language constructs and incorporation of more situational based constructs using situational-based instruments.

Validation of the Scientific Imagination Test–Verbal

Content Area: Education

Room: 2W402

Time: 12:00 – 12:30

Presenters: Chia-Chi Wang, Hsiao-Chi Ho, Ying-Yao Cheng

Background

Imagination has a great influence on people's thinking, language, and life experiences. Today's science education is the best opportunity for emphasizing imagination and innovation. The Scientific Imagination Test-Verbal (SIT-Verbal; Wang, Ho, & Cheng, in press) was designed specifically for 5th and 6th grade elementary school students, is a situation test that measures students' scientific imagination. However, various daily life experiences and cognitive development to the imagination for different ages should be considered.

Aims

In order to verify the SIT-Verbal as a reference for scientific imagination curriculum development and future innovation studies, the present study aimed to provide multiple validity evidence of the SIT-Verbal for different ages and identify the hierarchy of scientific imagination process through Rasch analyses.

Methods

Participants in this study were 767 3th to 6th grade elementary school students from Taiwan. For instrument, the SIT-Verbal was used in this study. The SIT-Verbal was covered four key components of scientific imagination process: brainstorming, association, transformation/elaboration, and conceptualization/organization/formation. For analysis, the multiple validities (Wolfe & Smith, 2007) of the SIT-Verbal were assessed using the Rasch partial credit model (PCM; Master, 1982). Differential item functioning (DIF) analysis was conducted to examine the invariance of item calibrations across genders and grades.

Sample

A total of 767 3th to 6th grade elementary school students (407 males and 357 females) in Taiwan were administered the SIT-Verbal.

Results

The results indicated that the SIT-Verbal showed good model-data fit and supported students' scientific imagination progressed "from brainstorming, association, transformation/elaboration, to conceptualization/organization/formation." In addition, no items exhibited substantial DIF across genders and grades. Furthermore, no significant difference in scientific imagination was evident between genders or grades.

Conclusions

It was concluded that the SIT-Verbal is a reliable and valuable open ended tool for assessing individual's scientific imagination, future research should address how the SIT-Verbal can be applied to the development of training courses to foster students' scientific imagination. Furthermore, the empirical data supported students' scientific imagination progressed "from brainstorming, association, transformation/elaboration, to conceptualization/organization/formation."

Future Directions

Based on the findings, the researcher suggested that develop a series of materials for teaching practices integrated assessment feedback. Suggestions for future research on revision of the SIT-Verbal were proposed.

Multidimensional IRT models for L2 computer-based oral english assessment: A comparison between 2PL-PCM and 2PL-RSM

Content Area: Language
Room: 2W403
Time: 12:00 – 12:30
Presenter: Wei Jie

Background

The key assumption of the UIRT models is that test items measure a unidimensional latent trait. UIRT models are appropriate for items that involve a single underlying ability or combination of abilities that are constant across items (Embretson & Reise, 2000). While in practice, students achievement on oral English test tasks are multidimensional in nature, which can also be treated as consecutive unidimensional estimates. This multidimensionality may happen where each oral test task measures more than one latent ability. Parameters for these abilities can be estimated within a unidimensional or multidimensional IRT model framework.

Aims

This research attempts to make a comparison between 2PL-PCM and 2PL-RSM derived from multidimensional models in a computer-based oral English test in a Chinese university.

Methods

The two IRT models were estimated and compared to verify the dimensional structure of the oral scale. Structural parameters are estimated for each model, and model fit is compared. Further simulation of above models is examined to testify the robustness and consistency.

Sample

The research has taken a sample of 2,000 participants with different majors and levels in computer-based oral English test in a Chinese university.

Results

The multidimensional analysis indicated best fit to the data and provides more reliable estimates of student achievement than the unidimensional approach.

Conclusions

The multidimensional IRT procedure can be successfully used in estimating L2 learners' ability dimensions on oral English proficiency assessment.

Future Directions

The multidimensional model is likely to produce more accurate and efficient parameter estimates, therefore provide richer information to teacher, test developer about the nature of student achievement.

Validation of the employees' service performance scale using Rasch model

Content Area: Business
Room: 2W401
Time: 12:30 – 1:00
Presenters: Saleh Al-Sinawi, A.Y.M. Atiquil Islam

Background

The world is facing rapid change due to the fast development in human life and phenomenal technology advancement. These changes and developments are forcing human beings to adapt with the rapid development in their milieu through rigorous training, staffing, involvement & participation, performance appraisals, compensation & rewards, caring, concern of customers, concern of employees, helping behavior, customer knowledge, human resource planning, management and development. However, existing researches showed that employees' service performance exhibited to be under development. Besides this, critical analyses to date demonstrated that there has been no research on employees' service performance within the Ministry of higher education context in Oman, thus, creating a gap for exploration.

Aims

As such, the aim of this study is to validate the service performance scale for evaluating employees' performance in ministry of education in Oman.

Methods

A set of questionnaire containing validated items from earlier studies was put together and modified to suit the present study. An eleven-point Likert scale that indicated degrees of agreement/disagreement was used to capture the employees' views about the employees' service performance. Data analysis was conducted through Rasch model using Winsteps version 3.49.

Sample

A total of 514 employees were collected from ministry of education applying purposive sampling procedure.

Results

The findings of Rasch model discovered that (i) the items reliability was found to be at 0.99 ($SD = 294.8$), while the persons reliability was 0.96 ($SD = 108.0$); (ii) the items and persons separation were 8.53 and 4.92, respectively; (iii) all the items measured in the same direction (ptmea. corr. > 0.27); (iv) the majority of items revealed good item fit and constructed a continuum of increasing intensity. The findings also revealed that the variance explained by the measures was 73% which confirmed that the items were competent to endorse employees' service performance in ministry of education in Oman.

Conclusions

The Rasch analysis foster support for the internal consistency, unidimensionality, and measurement properties of the service performance scale which is valid.

Future Directions

The results also contributed to body of the knowledge by validating the service performance scale which could be applied by future researchers and academicians in the dissimilar context of education.

Mitigating the effects of rater severity and examinee familiarity on the Objective Communicative Speaking Test

Content Area: Language

Room: 2W403

Time: 12:30 – 1:00

Presenters: Aaron Olaf Batty, Jeffrey Stewart

Background

The use of rating rubrics in speaking tests introduces aspects of subjectivity to the scores, which can manifest in loss of reliability due to differences in severity/leniency, halo effects, and reduced range in scores. To address these issues, the researchers developed the Objective Communicative Speaking Test, a task-based, tablet-computer-mediated, online test of communicative ability. Examinees are presented with information to explain to a rater who is unaware of what has been presented to the examinee. When the rater selects an answer on his tablet, the response is scored and the time to completion is stored, eliminating any subjectivity from the responses. However, two potential obstacles to this test format are that speed to task completion could be influenced by individual raters and examinees' familiarity with the test format.

Aims

The present research investigates the impact of raters and familiarity on task completion speed and proposes changes to the procedure, if necessary, to mitigate them.

Methods

A six-category rating scale was developed from completion times and many-facet Rasch modeling was employed to produce location estimates. To address the first concern, a one-way ANOVA and post-hoc tests on rater severity was employed. A bias/interaction analysis between examinees' test sessions was carried out to address the second concern.

Sample

The experimental sample was 43 second-language speakers of English.

Results

The ANOVA and post-hoc test on raters indicated that differences in severity were statistically non-significant for all but one rater. The bias/interaction analysis indicated a substantial increase in speed between examinees' first and second test sessions, but relatively little speed-up thereafter.

Conclusions

Although one rater was found to be significantly more severe than the rest, further rater training could likely mitigate this effect. To address the issue of examinee familiarity, a revised model incorporating first and subsequent test sessions as a facet improved Rasch person reliability from 0.87 to 0.88.

Future Directions

Work with this test format is ongoing.

Medical students' approaches to learning: A construct validation from the Rasch perspective

Content Area: Education

Room: 2W404

Time: 12:30 – 1:00

Presenters: Vernon Mogol, Yan Chen, Marcus Henning, Andy Wearn, Jennifer Weller, Warwick Bagg

Background

The self-reported Revised Two-Factor Study Process Questionnaire (R-SPQ-2F) was developed using classical test theory to measure students' deep (DA) and surface (SA) approaches to learning.

Aims

To investigate the extent to which DA and SA scales satisfy the Rasch measurement model. To explore if the full scale (FS), after reverse scoring responses to SA items, could be measuring a unidimensional construct where lower scores indicate SA and higher scores DA.

Methods

Data were fitted to the model using WINSTEPS. Scale assessment included investigation of rating scale functioning, item fit, targeting, reliability, dimensionality and differential item functioning (DIF).

Sample

Year 2 to Year 5 University of Auckland medical students ($N = 882$) were invited to participate in a longitudinal study by responding to an online survey. Data were collected from 327 respondents (mean age 22.2 years old; 53% were females).

Results

The rating scale advanced monotonically and the items showed acceptable fit for the three scales.

SA scale showed poor targeting; the mean person location (-1.16) was far from the mean item location of 0.00 logits. The range of person locations (-4.29 to 2.09) was not adequately covered by item thresholds (-3.54 to 4.32) hence, there were not enough thresholds to accurately estimate locations of persons with very low SA.

DA and FS were well targeted; mean person locations were close to zero and the ranges of person locations were covered by the item thresholds.

All three scales had acceptable reliabilities (0.74 – 0.79) but only DA and SA satisfied the unidimensionality requirement. No curriculum (retired vs reinvigorated) and gender DIF was detected.

Conclusions

Analysis support the original idea of DA and SA as separate dimensions and not polar ends of the continuum of the construct learning approach.

Could programme be attracting/retaining students with low SA, hence, poor targeting? If so, care must be exercised when using the SA scores to evaluate interventions that are meant to discourage medical students from using surface approach to learning.

Future Directions

Racked and stacked analysis of the longitudinal data, with item locations in the current study as anchors; and using interval Rasch estimates in future projects instead of ordinal raw scores could be explored.

Guessing and the Rasch model

Content Area: Technical

Room: 2W402

Time: 3:30 – 4:00

Presenters: J. Lake, Trevor Holster

Background

Stewart (2014) questioned Beglar's (2010) use of Rasch analysis of the Vocabulary Size Test (VST) and advocated the use of 3-parameter logistic item response theory (3PLIRT) on the basis that this models a non-zero lower asymptote for items, often called a "guessing" parameter (Engelhard, 2013, p. 96). In support of this, Stewart presented fit statistics derived from Rasch analysis of random numbers displaying good data-model fit. Stewart argued that the Rasch model was unable to identify problematic guessing. Stewart correctly observed that advice given by Beglar (2010) and Nation (2012) on the rescaling of raw scores as criterion measures of vocabulary size is problematic but provided no empirical evidence to support his preference for 3PLIRT.

Aims

This study used empirical data in two studies to investigate whether Rasch analysis could identify problematic guessing and whether correction for lucky guessing affected measurement.

Methods

In study one, real data were taken from a 50 item vocabulary test. First, 250 random guessing response sets were analyzed independently, then 250 real response sets, and finally the combined 500 response sets. This allowed the three conditions to be compared on the basis of fit, reliability of separation, and dimensionality, reflecting the analyses reported by Beglar (2010). In study two, data were taken from vocabulary and morphology tests. In both studies analysis was conducted under the Rasch dichotomous model using the default estimation settings in Winsteps version 3.81.0 (Linacre, 2014).

Sample

In study one, participants were 250 female students enrolled at a public Japanese women's university. In study two, 349 participants were from a public women's university and a public co-educational university.

Results

Rasch person reliability coefficient, Rasch separation index and fit indexes were able to detect guessing in study one. In study two correction for guessing resulted in increasing the difficulty of the test but the corrected correlated highly with uncorrected results. The transformation from uncorrected to corrected was linear across both persons and items.

Conclusions

In study one replicating the Rasch analyses used by Beglar (2010) easily identified patterns of random guessing. In study two guessing correction did not make a significant difference to the Rasch estimates of relative person ability and item difficulty and would not affect the interpretations we make of the results. This study empirically demonstrated that random guessing does not invalidate Rasch analysis of thoughtfully developed SR tests. Rasch fit statistics allowed diagnosis of unexpected response patterns such as random guessing, and correction for guessing resulted in linear rescaling, just as the use of formula scores would do.

Future Directions

From the perspective of measurement, serious misfit is problematic because it indicates distortion of the measurement scale, the Rasch model allows us to identify which responses are misfitting the overall pattern in the data, providing diagnostic analysis such as Engelhard's (2009) investigation of students with disabilities. For classroom level diagnosis, Winsteps provides "Kidmap" results for individual students. One possibility for future studies would be to explore the diagnostic uses of Kidmaps.

Evaluating the validity of the rating scale for an English speaking assessment: An approach combining MFRA and SEM

Content Area: Language
Room: 2W403
Time: 3:30 – 4:00
Presenter: Jinsong Fan

Background

Despite the increasingly extensive application of both Many-Facet Rasch Analysis (MFRA) and Structural Equation Modeling (SEM) in language testing research (e.g., In'nami & Koizumi, 2011; McNamara & Knoch, 2012), few attempts have been made to combine these two analytic approaches in test validation studies. MFRA and SEM can play a complementary role in examining test validity, and the combination of these two approaches can therefore help to build a more convincing validity argument for a language test.

Aims

This presentation reports on a construct validation study of the analytic rating scale for a university-based spoken English test, using both MFRA and SEM.

Methods

MFRA was first employed to investigate raters' performance and the quality of the rating scale. Following MFRA analysis, SEM was utilized to examine the language ability structure that was reflected in the rating scale. Based on a review of relevant literature (e.g., Sawaki, 2007), four Multitrait Multimethod Confirmatory Factor Analysis models were posited, including the Correlated Trait Factor Model, the Orthogonal Trait Factor Model, the Unitary Trait Factor Model, and the Higher-Order Trait Factor Model.

Sample

Research data were collected from 74 students and two certified raters. The analytic rating scale used in this study had four dimensions: pronunciation, content, grammar, and vocabulary, each dimension using four categories (1 – 4) accompanied by detailed description of language ability at each level.

Results

MFRA results demonstrated that there was no difference in raters' severity, and raters used the rating scale reasonably and consistently in their assessment process. In addition, the MFRA results also indicated that the four-category scale functioned well, and could effectively define distinct levels on the latent trait. SEM results indicated that the Higher-Order Trait Factor Model best fit the empirical rating data.

Conclusions

The results yielded by MFRA and SEM lent crucial support to the construct validity of the rating scale.

Future Directions

In addition, through demonstrating how MFRA and SEM could be used in complementary roles in validation research, this study has important methodological implications for examining the validity of analytic rating scales in performance assessment.

Examining the vocabulary size test under Rasch and three parameter item response models

Content Area: Technical
Room: 2W402
Time: 4:00 – 4:30
Presenters: Jeffrey Stewart, Stuart Mclean, Brandon Kramer

Background

The Vocabulary Size Test (VST) was created to provide a reliable estimate of a second language learner's written receptive vocabulary size, measuring from the most frequent fourteen 1000 word-families of the spoken subsection of the British National Corpus (Nation & Beglar, 2007). English Vocabulary size is estimated by multiplying the raw score on the test by 100. While Beglar (2010) and Elgort (2013) recommend that users should limit the amount of the test taken to only slightly above a students' level, Nation (2012), Karami (2012), Nguyen and Nation (2011), and Coxhead (2014) argue that learners should take every level of the test, as they may know some low-frequency words. However, there have been concerns that correct responses on lower-frequency levels

of the multiple-choice test could largely be attributed to guesses rather than vocabulary knowledge (Stewart, 2014).

Aims

To determine blind guessing rates for the multiple-choice items on the VST, examine the proportion of low-level students' scores on the lowest frequency level tested that can be attributed to guessing under the 3PL model, and conduct a model fit comparison to determine if the 3PL model offers a significantly better description of the data when compared to the Rasch model.

Methods

The first eight 1000-word levels of the VST were analyzed with the software program IRTPro 2.1 (Cai, Thissen, & du Toit, 2011) under the Rasch model and the 3PL model.

Sample

3,373 Japanese university students' responses to the first eight levels of the original VST. The most common department Hensachi (Standard Rank Score on the National university exam) of participants in the present study was 49 ($n = 1030$), the majority of the data came from participants with average or above average department Hensachi of 50 or above ($n = 1940$), with an overall mean department Hensachi score of 53.2 for all participants, which is somewhat above the national mean of 50.

Results

Under the 3PL model, which had superior fit to the data, very low-level learners would be expected to receive a score of approximately 2.32 out of 10 due to guessing on the lowest-frequency band of the test examined (the 8k level). As students in departments with a Hensachi near the national mean of 50 had a mean score of 3.58 out of 10 on this word level, it seems that while average students' scores out of 10 on this word level are above chance, the bulk of their scores can likely be attributed to guessing that is unrelated to vocabulary knowledge.

Conclusions

The results support Beglar (2010) and Elgort's (2013) position that students should not sit every level of the test.

Future Directions

Further investigation, perhaps involving mixed-method research in which students are interviewed about their reasoning for choosing answers, will be necessary to provide detailed evidence of which student populations (perhaps defined by Hensachi) should take which bands of the test.

Propagation of rater error within a language assessment

Content Area: Language

Room: 2W403

Time: 4:00 – 4:30

Presenter: Jeffrey Durand

Background

In rated assessments, all the raters must be part of a fully connected network. This allows determination of rater strictness, which affects estimates of student ability. If a rater is unusually strict with a particular student, however, that student's score may not reflect her true ability. There is error in that student's score. The error is not limited to that one student, however. Because estimates of rater strictness depend on comparisons with other raters, the error with the one student is passed along to other raters, and by extension, to other students.

Aims

The purpose of this study is to examine how characteristics of the rating network affect how the error is distributed throughout the network. The goal is to understand how to create rating networks that minimize error around the network.

Methods

This simulation study used an actual rater network from a rated writing examination. Data was generated from set student ability and rater strictness values, with no error induced in the network. Next, one rater in the network was selected and new data was generated with randomness (error) induced in only this rater's scores. The total error in student and rater measures was calculated for the entire network. This process was repeated for each rater in the network. Degree, closeness, betweenness, and eigenvector centrality measures were also calculated for each rater. These values were correlated with the total error associated with each rater.

Sample

Data was simulated.

Results

Results indicate that having a high degree or eigenvector centrality can lead to less error in rater strictness estimates while having a higher closeness centrality can lead to more network error. When considering error in student ability estimates, only degree and eigenvector measures of centrality were important. Much of the variation in network error is unexplained, however.

Conclusions

In creating or shaping rater networks, making raters more connected is important for reducing error. Changing rating partners more frequently is more important than having rating partners work together.

Future Directions

The results are affected by the way that error is induced in each rater. This is an area that needs further investigation.