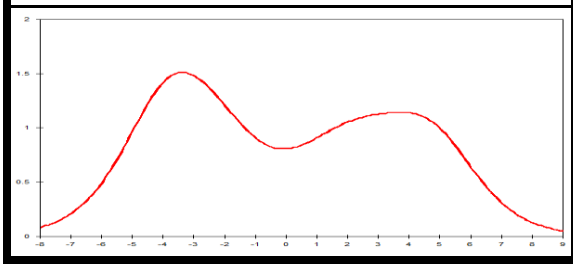
Test Information Function



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 25 No. 1 Summer 2011 ISSN 1051-0796

How Much Do Emotions Alter Our Measurements?

Do situational factors during measurement change measurement estimates (item difficulties, person abilities, standard errors, etc.)? Our research shows that item difficulties are different when one accounts for individual differences in positive affectivity during test administration. We calibrated the items of a Spelling Instrument ignoring, and then including, the influence of positive affectivity. A two-level Hierarchical Generalized Linear Model (HGLM) was used:

Level-1 (Bernoulli) Rasch model for a test of i = 1,*k* dichotomous items:

$$\log (p_{ij} / (1-p_{ij})) = \beta_{0j} + \beta_{1j}X_{1j} + \ldots + \beta_{ij}X_{ij} + \ldots + \beta_{(k-1)j}X_{(k-1)j}$$

where p_{ij} is the probability that person *j* will answer item *i* correctly. β_{oj} is person ability relative to item *k* and is the intercept of the model. β_{1j} is the easiness of the item 1 (relative to item *k*) for person *j* and the coefficient of dummy variable X_1. For p_{ij}, all the dummy variables are 0, except for $X_{ij} = 1$ which flags that this equation models a response to item *i*.

Level-2 model expressing person and item estimates:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10}; \ldots ; \beta_{ij} = \gamma_{i0}; \ldots ; \beta_{(k-1)j} = \gamma_{(k-1)0}$$

γ_{00} is the mean of the person ability distribution relative to item *k*. u_{0j} is the value of the random ability effect specific to person *j*. $\{u_{0j}\}$ are modeled to be normally distributed, $N(0,\tau)$, across the person sample. The item easinesses, $\{\gamma_{i0}\}$ are modeled to be invariant across the sample. When this two-level model is applied to the response by person *j* to item *i*, the probability of a correct response becomes:

$$\log (p_{ij} / (1-p_{ij})) = \gamma_{00} + \gamma_{i0} + u_{0j}$$

In the analysis of our 7-item test of spelling ability, $k = 7$.

In a second "adjusted" analysis, the Level-2 model was modified by adding the term γ_{01}*PositiveAffect$_j$ to β_{0j} in order to account for levels of positive affectivity during the testing situation. PositiveAffect$_j$ is a measure of the positive affect of person *j* assessed just prior to the achievement test.

A comparison of the results of the two analyses is

Figure 1. Test Characteristic Curves for both analyses

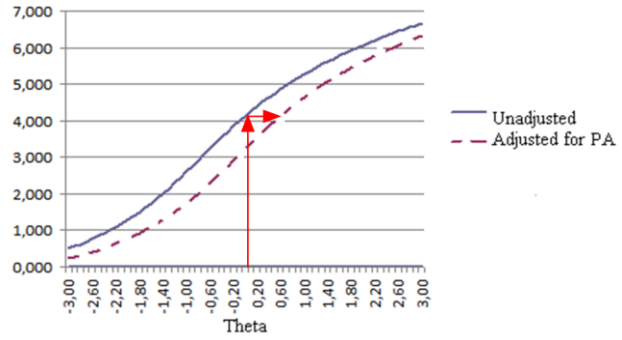


Figure 2. Test Information Functions for both analyses

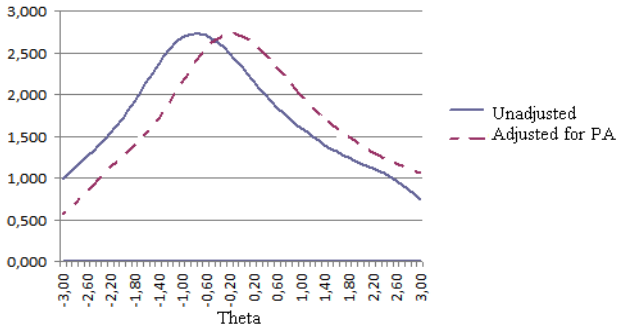


Figure 3. Conditional SEM functions for both analyses

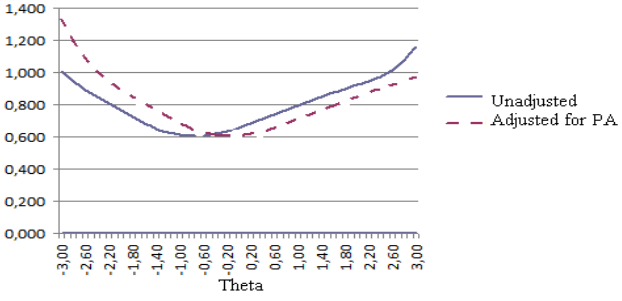


Table of Contents

instructive. The two Test Characteristic Curves (TCCs, Test Response Functions, TRFs, Figure 1) are drawn relative to the apparent difficulty of Spelling item 7. So, someone whose estimated ability is the same as the difficulty of item 7 (theta = 0) has an expected score of 4.2 (out of 7) in the first, unadjusted, analysis, but 3.5 in the second, adjusted, analysis. The effect of positive affect has been to raise the expected score by about 0.7 score-points, equivalent to a theta advance of 0.6 logits, a half-year growth in many educational settings.

The slopes of the TCCs are the Test Information Functions (TIFs). These are off-set by about 0.6 logits (as we would predict). The standard errors of the ability-estimate measurements (SEMs, Figure 3) are the inverse square-roots of the TIF. For most purposes, we would like the SEMs to be approximately uniform, giving equal measurement precision across the ability distribution. Here, this would require flatter TIFs, and more uniform distribution of the difficulties of the 7 items across the target range of abilities. This change would also lessen the impact of the affective bias on measurement precision.

Characteristics such as motivation, emotions, fatigue, and other situational factors can be systematic sources of bias and so can lead to estimates that deviate markedly from the actual abilities of the persons on the intended latent variable. The moral of our story is that care should be taken to watch for, and then adjust for, sources of bias in our measures.

Georgios D. Sideridis
Ioannis Tsaousis
University of Crete

Journal of Applied Measurement
Vol. 12, No. 1 Spring 2011

Using Adjusted GPA and Adjusted Course Difficulty Measures to Evaluate Differential Grading Practices in College. *Dina Bassiri and E. Mathew Schulz, 1-11*

Optimizing the Compatibility between Rating Scales and Measures of Productive Second Language Competence. *Christopher Weaver, 12-24*

Developing a Domain Theory Defining and Exemplifying a Learning Theory of Progressive Attainments, *C. Victor Bunderson, 25-48*

Bringing Human, Social, and Natural Capital to Life: Practical Consequences and Opportunities, *William P. Fisher, Jr., 49-66*

Understanding Rasch Measurement: Distractors with Information in Multiple Choice Items: A Rationale Based on the Rasch Model, *David Andrich and Irene Styles, 67-95*

Richard M. Smith, Editor
JAM web site: www.jampress.org

Rasch-related Coming Events

Aug. 31 - Sept. 2, 2011, Wed.-Fri. IMEKO Conference, Jena, Germany, www.tu-ilmenau.de

Sept. 14-16, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 16 - Oct. 14, 2011, Fri.-Fri. Online course: Rasch Applications in Clinical Assessment, Survey Research, and Educational Measurement (W. Fisher), www.statistics.com

Sept. 19-21, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK,
Sept. 22-23, 2011, Wed.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Oct. 7-8, 2011, Fri.-Sat. In-person workshop: Introduction to Rasch Measurement: Theory and Applications (E. Smith, R. Smith), Minnesota, www.jampress.org

Oct. 21 - Nov. 19, 2011, Fri.-Fri. Online course: Rasch - Core Topics (Linacre) www.statistics.com

Nov. 25, 2011, Fri.. VI Workshop sobre Modelos de Rasch en Administración de Empresas. Nuevas Tendencias. Tenerife, Spain. www.institutos.ull.es

Nov. 30 - Dec 2, 2011, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), UK
Dec. 5-7, 2011, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Jan. 9 - Apr. 27, 2012, Mon.-Fri. Online course: Rating Scale and Questionnaire Design and Analysis (E.V. Smith), education.uic.edu

Jan. 9-15, 2012, Mon.-Wed. In-person workshop: Introductory Rasch course (Andrich, RUMM2030),
Jan. 16-20, 2012, Mon.-Wed. In-person workshop: Advanced Rasch course (Andrich, RUMM2030), Perth, Australia, www.education.uwa.edu.au

Jan. 23-25, 2012, Mon.-Wed. Fifth International Conference on Probabilistic Models for Measurement, Perth, Australia, www.education.uwa.edu.au

March 20, 2012, Tues. 6th UK Rasch User Group Meeting, Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric/News1.htm

Apr. 11-12, 2012, Wed.-Thurs. IOMW International Objective Measurement Workshop, Vancouver BC, Canada, Announcement

Apr. 13-17, 2012, Fri.-Tues. AERA Annual Meeting, Vancouver BC, Canada, www.aera.net

Rasch Analysis of Repeated Measures

Repeated measures are common in rehabilitation studies where patients are scored on assessments at both admission and discharge. There are often intermediate or follow-up data collection periods in addition. The amount of change in patient functional status is an important indicator of rehabilitation quality. In order to determine that it is indeed the patients who have changed and not the item difficulty, constant "anchor" values are needed to fix item difficulties at admission and discharge (or any other time point) within a common frame of reference. Yet creating an anchor file is problematic.

One approach is to create a file of item anchor values by "stacking" the admission and discharge data so that each item corresponds to one column, and each time-point for each person is a row the combined dataset. However this approach may violate the Rasch assumption of local independence in the observations because some characteristics of the patients span time-points . Yet creating item anchor values from either the admission data only, or the discharge data only, and then applying those values to the whole data set may not be reasonable either. Generally, patients are quite disabled at admission to rehabilitation so performance on difficult items of assessment tools are rarely observed or are scored in their lower rating-scale categories. At discharge, patients have often made considerable improvement and most will be scored in the top categories of easier items. At either admission or discharge, some items will be "off-target" compared to patient ability and, for some items (the hardest ones at admission, and the easiest ones at discharge), only one or two categories of the rating scale may be observed.

This suggests a different approach:

1) Create a random sample of patients across the time-points so that each patient is only in the data set once but all time-points are equally represented.

2) Analyze this "random" data set and estimate the item difficulties and Rasch-Andrich thresholds. Save these values in anchor files. They become the definitive set of item difficulties, defining the measurement framework of the latent variable.

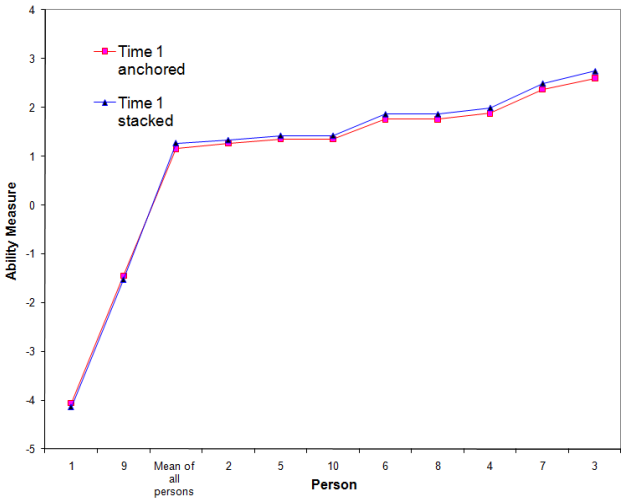
3) Apply the anchor files to the estimation of the person abilities at all time points. This can be done either with each time-point in a separate dataset or with all time-points stacked in one dataset. There will be no interaction between the observations of each person at the different person because they are isolated from each other by the item anchor values.

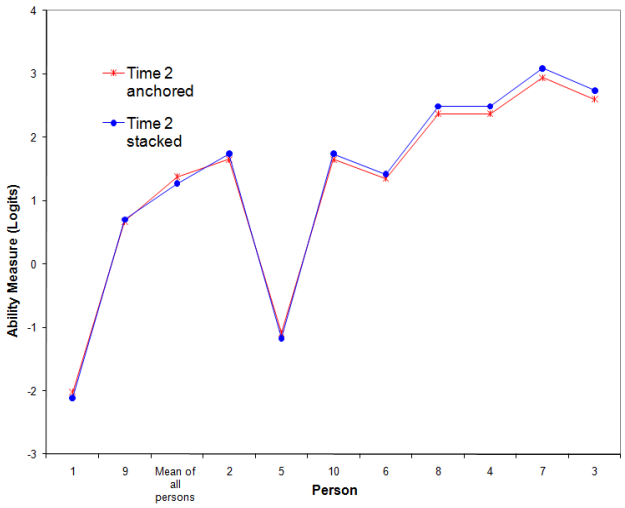
The suggested approach was applied to a dataset of 459 older adults measured on a 13-item self-report survey at 5 time points. Time 1 is before treatment; Time 2 is after treatment. Not all adults were observed at all time-points. All 13 items fit the Rasch model. In accordance with (1), a random sample was selected across all 5 time points so that each person was only in the "random" dataset once but all 5 time points were equally represented. Then (2), this random sample was used to create the anchor files. Finally (3), the anchor files were used in the estimation of 327 adults with both Time 1 and Time 2 records. For comparison, an unanchored "stacked" analysis of all 1527 available records for all adults at all time-points was performed. In this last analysis, the estimates for Time 1 and Time 2 would be influenced by local dependency across time-points, if there is any.

The Figures show the relationship between the "stacked" and "anchored" measures of the first 10 persons with both Time 1 and Time 2 records. We can see that in this dataset the influence of local dependency is small, much less than the S.E.s of the measures which are 0.3 logits or more.

In this dataset, dependencies in the data have little effect on person measures. However, using anchor values from a random sample (selected to be without intra-person dependencies) should satisfy manuscript reviewers that a possible source of time-series dependency has been eliminated.

Trudy Mallinson,
University of Southern California





Testlets and Threshold Disordering

Question: I combined locally-dependent dichotomous items into polytomous "testlet" items. The item-fit statistics look good, but within the polytomous items some of the Andrich thresholds are disordered. Collapsing categories made the item-fit worse. What should I do?

Answer: Ordered thresholds are relevant and central when you have an item format in which the categories are intended to reflect order. However, when you form testlets, you no longer have that situation. You have another structure in which there is no reason for the thresholds to be ordered. In fact, the more local dependence you have accounted for with the testlet form, the more the thresholds will be disordered.

It is good to hear that when you tried to correct the order of the thresholds in **this** situation, that you got worse fit. This is because in your situation, disordered thresholds are not showing anything wrong - they are reflecting the amount of local dependence.

Although I did not call the collection of items testlets, but just subtests, I discussed this in Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. Embretson (Ed.), Test design: Contributions from psychology, education and psychometrics. Academic Press, New York. (Chapter 9, pp. 245-273.)

See also Andrich, D. (2006) Item discrimination and Rasch-Andrich thresholds revisited. Rasch measurement Transactions. 20 (2), 1055 - 1057.

David Andrich, University of Western Australia

IMEKO Symposium: August 31 - September 2, 2011, Jena, Germany

The International Measurement Confederation (IMEKO) Symposium will include presentations of interest to RMT readers. www.tu-ilmenau.de/fakmb/Symposium-Programme.2647.0.html. And in the spirit of provoking more dialogue between the fields of metrology and psychometrics, as President-elect of the Psychometric Society, Mark Wilson has extended an invitation to Luca Mari, Chair of the IMEKO Technical Committee 7 on Measurement Science, to speak at next year's Psychometric Society meeting in Lincoln, Nebraska.

Four approaches in psychometrics (Guttman scaling, classical test theory, Rasch analysis, and construct mapping) and Luca Mari's (2000) sense of a functionally coherent measurement system, Mark Wilson (University of California, Berkeley), *plenary lecture*

Quantity and Quantity Value, Luca Mari (University of Cattaneo, Italy)

Metrological Properties of Classification, Sanowar H. Khan (City University, London, England)

Does Measurement Need its Own System Theory - An Appraisal, Klaus-Dieter Sommer (Physikalisch-Technische Bundesanstalt, Braunschweig, Germany)

Foundational Imperatives for Measurement with Mathematical Models, Nikolaus Bezruczko (Measurement and Evaluation Consulting, Chicago, Illinois)

From Breast-Q to Q-Score: Using Rasch Measurement to Better Capture Breast Surgery Outcomes, Stefan J. Cano (University of Plymouth, England)

How to Model and Test for the Mechanisms that Make Measurement Systems Tick, A. Jackson Stenner (MetaMetrics, Inc., Durham, North Carolina)

The Quantification of Latent Variables in the Social Sciences: Requirements for Scientific Measurement and Shortcomings of Current Procedures, Thomas Salzberger (University of Vienna, Austria)

Measurement, Metrology and the Coordination of Sociotechnical Networks, William P. Fisher, Jr.

Measurement Modeling: Foundations and Probabilistic Approach, Giovanni Battista Rossi (University of Genova, Italy)

The Role of Mathematical Modeling in the Analysis and Design of Measurement Systems, Sanowar H. Khan (City University, London, England)

Application-Oriented Approach to Mathematical Modeling of Measurement Processes, Roman Z. Morawski (Warsaw University of Technology, Poland)

Continuous Quantity and Unit; Their Centrality to Measurement, Gordon A Cooper (University of Western Australia, Crawley, Australia), and William P. Fisher, Jr. (University of California, Berkeley)

Features of the VIM: Application to the Practical Aspects of Measurement, Tetyana Gordiyenko and Oleh Velychko (State Enterprise UkrSREC, Kyiv, Ukraine)

A Technology Roadmap for Intangible Assets Metrology, William P. Fisher, Jr, and A. Jackson Stenner

A Clinical Scale for Measuring Functional Caregiving of Children Assisted with Medical Technologies, Nikolaus Bezruczko, Shu-C. Chen, Connie Hill, Joyce M. Chesniak

Body, Mind, and Spirit are Instrumental to Functional Health: A Case Study, Carl V. Granger (State University of New York, Buffalo), Nikolaus Bezruczko

Reference: Mari, L. (2000). Beyond The representational viewpoint: A new formalization of measurement. Measurement, 27, 71-84.

Using the CUTLO Procedure to Investigate Guessing

Guessing and receiving unearned credit is a possibility with any multiple-choice examination. Rogers (1999) identified three types of guessing: random, cued, and informed. Random guessing refers to blindly choosing a response to an item. Cued guessing refers to making a response based on some sort of stimulus in a test item, such as wording cues, cues associated with item stems, or choices among the distracters. Informed guessing refers to making a response based on some partial knowledge or on misinformation. One would expect an individual who relies solely on random guessing to have the lowest probability of passing an examination; however, cued guessing and informed guessing would likely increase an individual's chance of passing an examination.

Recently, four non-physicians with doctoral degrees in such areas as clinical psychology, educational psychology, evaluation, and curriculum and instruction attempted to pass the American Board of Family Medicine's (ABFM) certification examination in an attempt to determine how savvy test-takers without medical knowledge or training would fare on the 350-item examination (O'Neill, Royal & Puffer, 2011). As expected, the non-physicians failed miserably. In fact, the failures were so dismal that three of the four non-physicians failed to outscore a single physician (from a pool of 10,818 physicians), and the one non-physician who did outscore physicians only managed to outscore four, two of whom were international medical graduates and two US medical graduates who failed to complete the examination by leaving 33 and 79 items unanswered, resulting in incorrect answers. Even then, it can be argued that the reason the highest-performing non-physician outscored any physician at all is because he has a background in clinical psychology, which likely aided his performance on the ABFM examination as 7% of the test items are classified as *psychogenics*.

The minimum passing standard for the 2009 certification examination was a scaled score of 390 on a scale of 200-800. The four non-physicians scored below the reportable range with scores of 20, 80, 90, and 160. To investigate the effects of guessing, four physicians who scored 390 were included in the analysis for comparative purposes.

A Guttman (1944) scale of the 50 most unexpected responses (see figure 1) clearly shows that the four non-physicians managed to correctly guess numerous items that they should have answered incorrectly based on their ability estimates. It should be noted that each "1" represents a correct response when an incorrect response was expected, and each "0" represents and incorrect response when a correct response was expected. Each "." represents an expected response.

To further investigate the effects of guessing, the Winsteps CUTLO procedure was applied. CUTLO allows researchers to exclude responses in cases where it is highly probable that guessing could occur, as indicated by a low probability of success. A CUTLO of 2 was used in this analysis, which excluded any items that were 2 or more logits above a participant's ability estimate. Table 1 compares the non-physicians scaled scores both with and without the CUTLO procedure.

```
MOST UNEXPECTED RESPONSES
Candidate   Scaled Score  |Item: Easier                                  Harder
  MD1          390        |.0..0.............................................
  MD2          390        |....0.............................................1
  MD3          390        |..0...............................................1
  MD4          390        |...0..............................................
  Non-MD1      160        |.....................11.....1.1.1.111...1....1.
  Non-MD2       90        |................1...11111.1..1.......1.111..11...
  Non-MD3       80        |0............1..11.111.11.1.1.1.1.....1...1....1..
  Non-MD4       20        |0....11111111111111.11..1..1.1....1.1.1.1...1.....
                          |-------------------------------------------------
```

Figure 1. Guttman Scalogram of the 50 most unexpected responses.

Two of the non-physicians' scores fluctuated slightly as a result of the CUTLO procedure, while the other two scores remained relatively stable. The unstable scores for Non-MD3 and Non-MD4 provide evidence that these individuals' scores were actually inflated by the influence of guessing, as these two participants received credit for correctly answering items that were beyond their ability using well-targeted items. While it could be argued that all four non-physicians relied heavily on guessing, it is clear that two of the four relied even more heavily on guessing.

Additional evidence to support this claim is found when subtest scoring is investigated. The two non-physicians with backgrounds in psychology (Non-MD1 and Non-MD2) scored considerably higher in the psychogenics area than the two non-physicians with backgrounds in evaluation and curriculum and instruction (Non-MD3 and Non-MD4). This suggests that two of the

Table 1. Comparing Non-Physicians' Performance by Scaled Scores

	Non-MD1	Non-MD2	Non-MD3	Non-MD4
Regular Analysis	169	98	90	29
With CUTLO	167	96	75	14

non-physicians had some content knowledge of psychogenics or that their responses were based in part on informed guessing. Although the analysis using the CUTLO procedure suggests that there was some guessing going on, overall the Rasch analysis proved to be fairly robust.

Critics of the Rasch model often argue the exclusion of the guessing parameter is a limitation of the model. This is simply not true. In cases like this one, unexpected responses are easily identified and persons who are likely to have guessed can be detected quite well. What to do with the guessed responses, on the other hand, is a separate policy issue. In any instance, the fact remains that valid inferences can be made about who was likely to have guessed without any need for additional model parameterization.

Kenneth D. Royal, Thomas R. O'Neill
The American Board of Family Medicine

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Rev*iew, 9, 139-150.

O'Neill, T. R., Royal, K. D., & Puffer, J. P. (2011). Performance on the American Board of Family Medicine Certification Examination: Are Superior Test Taking Skills Alone Sufficient to Pass? *Journal of the American Board of Family Medicine, 24*(2), 175-180.

Rogers, H. J. (1999). Guessing in multiple-choice tests. In G. N. Masters and J. P. Keeves (Eds.). *Advances in measurement in educational research and assessment. (pp. 23-42)* Oxford, UK: Pergamon.

Book: Introduction to Many-Facet Rasch Measurement

Analyzing and Evaluating Rater-Mediated Assessments

by Thomas Eckes, 2011, www.peterlang.com

- Conceptual-psychometric framework for rater-mediated performance assessments
- Foundations of many-facet Rasch measurement Measurement of rater severity/leniency
- Correcting examinee proficiency estimates for rater severity differences
- Examining rater consistency and rating scale effectiveness
- Increasing validity and fairness of performance assessments.

Sample data taken from a writing performance assessment are used to illustrate key concepts, theoretical foundations, and analytic procedures.

Thomas Eckes is Head of the Psychometrics and Research Methodology Department at the TestDaF Institute, University of Bochum. His research interests include rater effects in large-scale assessments, standard setting, and web-based testing.

ORVOMS

Ohio River Valley Objective Measurement Seminar

On May 20, 2011 ORVOMS held its inaugural meeting on the campus of Xavier University in Cincinnati, Ohio. This event which was free of charge was held in order to: (1) provide a regional vehicle for Rasch model users to present their work to an audience who would understand what they are doing, (2) provide a place for people who share this interests to meet and share ideas, and (3) provide a friendly environment for people who have an interest, but not yet a background in the Rasch model to be able to learn more about the model's theoretical foundations and practical applications.

It was attended by approximately 25 people who came from Los Angeles, Iowa City, Toledo, Louisville, Lexington, and Cincinnati. The attendees were from diverse fields including: occupational therapy, criminology, certification testing, institutional research, biostatistics/epidemiology, and psychology (clinical, quantitative, and industrial/organizational).

Tom O'Neill began with a few opening remarks regarding the purpose of the conference, followed by Ed Wolfe giving the keynote presentation, An Introduction to Rasch Measurement. Among the other presentations, were topics such as equating with small sample sizes, construct stability across subpopulations, using very short survey forms, impact of raters' severity on a measure of consciousness, and rating scale category usage on a commitment to health survey.

A special thank you goes to Cindy Kelly, Ph.D. from Xavier's School of Nursing for hosting the conference and helping to make it a success.

The next ORVOMS will be in the spring of 2012 at the University of Kentucky in Lexington. More information on ORVOMS 2012 will follow. If you would like to be included on the ORVOMS email roster, please email Tom O'Neill - *toneill~theabfm.org* - or Brad Schulte - *bschulte~theabfm.org*

jMetrik 2.0 is Here! Free!

jMetrik 2.0 is a significant revision to jMetrik. New features include Ranking Procedures, Test Scaling, Item Response Theory, IRT Test Equating, and Item Maps. The logging and syntax features have been improved and a number of bugs have been fixed.

jMetrik is a free software application for psychometric analysis. It features Rasch Dichotomous, Rating-Scale and Partial-Credit Models. It includes procedures for basic descriptive statistics, graphs, Classical item analysis, Confirmatory Factor Analysis, Non-Parametric IRT, and more. jMetrik is a pure Java application that runs on Windows, Mac OS-X, and Linux platforms.

Patrick Meyer, www.itemanalysis.com