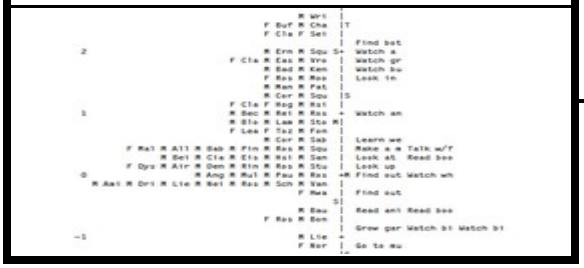
Wright Map



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 25 No. 3 Winter 2011 ISSN 1051-0796

Please Join the Kaggle-Grockit Competition: *"What Do You Know"*

Grockit, https://grockit.com/, is an online social-learning company. Grockit prepares students for the SAT, ACT, GMAT, LSAT and GRE standardized exams, and administer practice tests online. Grockit has made available, through *Kaggle,* a database containing the response strings of 179,106 users (students) responding to 6,045 questions (test items) from practice tests for the ACT (with subtests: English, Math, Reading, Science), SAT (Math, Reading, Writing) and GMAT (Quantitative, Verbal). Each student has responded to between 1 and 2,352 questions. Each question has been answered by between 2 and 14,063 users. The dataset has 4,851,475 reported responses, so that there are 99.6% missing responses. Grockit has also supplied information about each of the questions, such as its subtest, and also about each of the user's responses, such as its start and finish times.

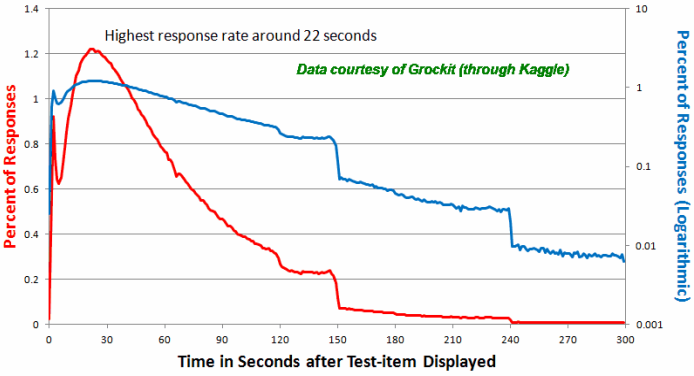


Figure 1. Response times for valid scored responses in the Grockit-Kaggle database .

There are a further 93,100 responses in the database with user and question information, but with blanked out responses. The challenge from Grockit is to predict which of those responses were correct and which were incorrect. For each of those responses, competitors submit the probability that the response was correct. Competitors are allowed to submit two sets of 93,100 response-probabilities each day until the competition ends on 29 February, 2012. Please join the competition at the Kaggle website:

http://www.kaggle.com/c/WhatDoYouKnow.

Kaggle have done their own analysis of this dataset, and have reported the Binomial Capped Deviance, BCD, of its set of predictions as the "LMER Benchmark". They used a random-effects Rasch model implemented with the LMER package of the R-Statistics software. This Benchmark analysis is surprisingly effective, producing a better BCD than a simple run of Winsteps or Facets. As of 12/18/2011, the leading submission appears to predict about 65% successfully. It is not known what technique was used to create that submission.

The Grockit data also reveal aspects of students' behavior. Figure 1 shows how long after an item is displayed on the computer's screen that the student responds. The logarithmic curve is close to linear, but with two steps. This suggests that there may be a time-out mechanism at work. Figure 2 shows the correctness of the responses in Figure 1. Again we can speculate about what mechanisms are at work.

John Michael Linacre

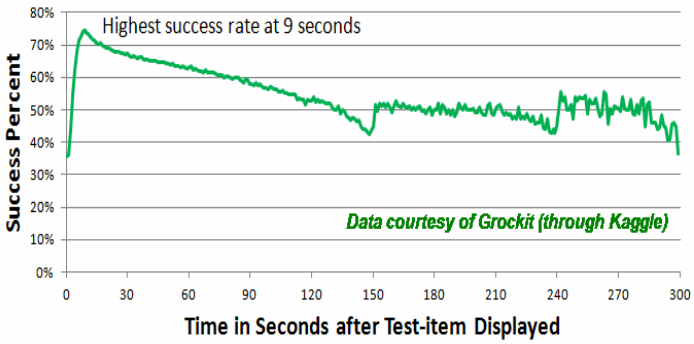


Figure 2. Percentage of correct responses at different response times for valid scored responses in the Grockit-Kaggle database.

Table of Contents

Kaggle-Grockit Competition (Linacre) 1329
Local Item Dependence (Sideridis) 1334
World Standards Day Celebration (Fisher) 1332
Wright Map (Wilson) 1331

Fifth International Conference on Probabilistic Models for Measurement

January 23-25, 2012
Perth, Australia

The University of Western Australia hosts the Fifth International Conference on Probabilistic Models for Measurement in Education, Psychology, Social Science and Health from Monday, 23 January 2012 to Wednesday, 25 January 2012.

The conference is preceded by two weeks of courses on social measurement, in particular Rasch measurement theory and practice, featuring the RUMM2030 software package (January 9-15 and January 16-20. Details at :

www.education.uwa.edu.au/raschconference

This Conference is followed by a Pearson Global Research Conference, "The Role of Technology and Assessment in System-wide Improvement" (January 27-28). Details at:
www.pearson.com.au/marketing/corporate/pearson_global

6th Annual UK Rasch User Group Day March 20, 2012, Leeds, UK

This event will be hosted by the Psychometric Laboratory in the Department of Rehabilitation Medicine at Leeds University. The venue will be Weetwood Hall, on the ring road to the north of Leeds. There are buses from the town to Weetwood Hall, and the 95 bus from the railway station http://www.weetwood.co.uk. Accommodation is also available at Weetwood Hall.

The cost of the day will be £20 and will include lunch (sponsored by the Psychometric Laboratory). It will begin at 10.00 and finish at 16.30. The Abstract call is informal, and is due at the latest by **9th March 2012**, although we would appreciate abstracts earlier than this if possible, in order to help with our planning.

Conference registration forms and Abstract forms are available at:
www.leeds.ac.uk/medicine/rehabmed/psychometric/News1.htm

We welcome all those involved with applying the Rasch model, in whatever discipline. We will consider short talks (15 minutes, or longer by request) in any of the following areas:

- Software development and demonstration
- Software comparisons
- Methodological issues
- Applied Rasch analysis (in health, education, and all other areas)

Alan Tennant BA, PhD.

Rasch-related Coming Events

Dec. 23, 2011, Fri. Submission deadline: IOMW International Objective Measurement Workshop, Apr. 11-12, 2012, Vancouver BC, Canada, Call for Papers

Jan. 6 - Feb. 3, 2011, Fri.-Sat. Online course: Rasch - Core Topics (Winsteps, introductory) online course (E. Smith, Winsteps), www.statistics.com

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), Perth, Australia, www.education.uwa.edu.au

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

March 21-23, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

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

May 23-25, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

May 28-30, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Aug. 6-9, 2012, Mon.-Thur. PROMS2012, Jiaxing University, Zhejiang Province, P.R.China, http://cfs.zjxu.edu.cn/proms/

Sept. 5-7, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 10-12, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Sept. 13-14, 2012, Thurs.-Fri. In-person workshop: Advanced Rasch (A. Tennant, RUMM), Leeds, UK,

Dec. 5-7, 2012, Wed.-Fri. In-person workshop: Introductory Rasch (A. Tennant, RUMM), Leeds, UK,

Dec. 10-12, 2012, Mon.-Wed. In-person workshop: Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Some Notes on the Term: "Wright Map"

I had heard the term "item map" being sometimes used to describe the representation of items and persons on the same continuum. I am not sure the origin of the term, nor of the idea. But I know that, for many years, Ben Wright had championed this approach to interpreting the results of measurement analyses, and also that he had made significant contributions to that approach, including enhancements and adaptations such as Kidmaps, fit maps, maps for polytomous items, etc. It seemed to me that, in fact, Ben had made his most significant contributions to measurement in the area of conceptualizing measures, and interpreting the results of measurement analyses, and that his central tool in doing so were these (many forms of) items maps. In addition, I knew of no one else who had made an equivalent contribution, especially not in terms of item mapping. Hence, I coined the term "Wright Map" in honor of Ben Wright and his very deep contributions to measurement.

This was about the end of 1999 and the beginning of 2000. After that, I used the term in my class (EDUC 274A at UC Berkeley) to get used to the sound of it, and, as I was drafting the text of my book *Constructing Measures* (Wilson, 2004), it became embedded in the text. The first time I used the term in a formal presentation was at a conference in Banff, Canada:

> Wilson, M., & Draney, K. (2000, May). Standard Mapping: A technique for setting standards and maintaining them over time. Paper in an invited symposium: "Models and analyses for combining and calibrating items of different types over time" at the International Conference on Measurement and Multivariate Analysis, Banff, Canada.

As far as I know, the first time the term appeared in print was when that conference paper was published:

> Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), Measurement and multivariate analysis (Proceedings of the International Conference on Measurement and Multivariate Analysis, Banff, Canada, May 12-14, 2000), pp 325-332. Tokyo: Springer-Verlag.

The following presentation is also interesting, as that is the first time Ben Wright was in the audience (he was very moved):

> Wilson, M. (2001, October). On choosing a model for measuring. Invited paper at the International Conference on Objective Measurement 3, Chicago, Il.

This conference paper was later published as:

> Wilson, M. (2003). On choosing a model for measuring. Methods of Psychological Research-Online, 8(3), 1-22. Download: www.dgps.de/fachgruppen/methoden/mpr-online/
>
> (Reprinted in: Smith, E.V., and Smith, R.M. (Eds.) (2004). Introduction to Rasch Measurement. Maple Grove, MN: JAM Press.)

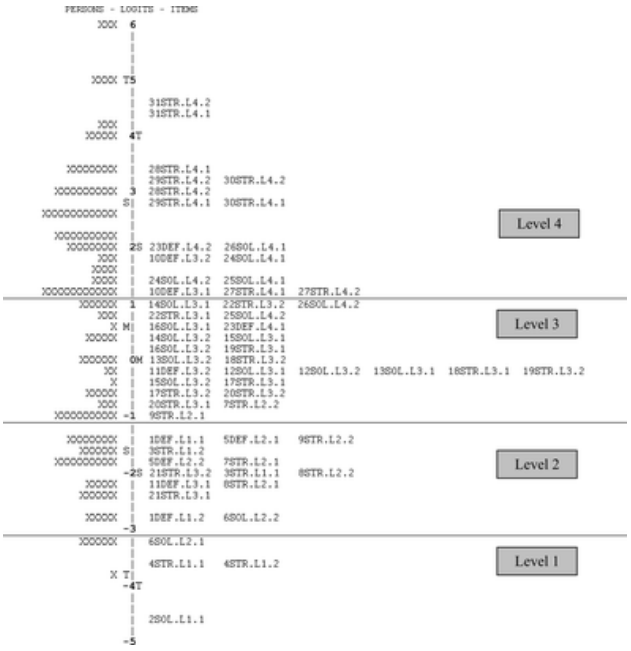
Some might be surprised that Ben didn't invent the term himself, as he was usually far from modest in most matters. But I believe he was indeed quite modest in formal matters, and was delighted to hear his name being used in this way.

Mark Wilson
University of California, Berkeley

Wright Maps

... the most helpful Rasch-based research for mathematics educators [and others!] marries the rigorous measurement demands of the model with the crucial qualitative distinctions demanded by researchers in our field. This marriage is most obvious in the way in which Rasch measurement output is often displayed as variable maps, supported - rather than replaced - by tables. The item-person map, often called a Wright map in honour of Rasch measurement's most tireless advocate, Ben Wright, displays both persons (in terms of their ability) and items (in terms of their difficulty) spaced along a common (usually) vertical axis marked with a scale.

Rosemary Callingham and Trevor Bond (2006) Research in Mathematics Education and Rasch Measurement. Editorial in Mathematics Education Research Journal, 18, 2, 1-10



Wright Map from: *Rittle-Johnson B., Matthew P.G., Taylor , R.S. McEldoon K.L. (2011) Assessing knowledge of mathematical equivalence: A construct-modeling approach. Journal of Educational Psychology, 103, 1, 85-104.*

World Standards Day Celebration:
Rasch Measurement as a Basis for a New Standards Framework

The 2011 U.S. celebration of World Standards Day took place on October 13 at the Fairmont Hotel in Washington, D.C., with the theme of "Advancing Safety and Sustainability Standards Worldwide." The evening began with a reception in a hall of exhibits from the celebrations sponsors, which included the National Institute for Standards and Technology (NIST), the Society for Standards Professionals (SES), the American National Standards Institute (ANSI), Microsoft, IEEE, Underwriters Laboratories, the Consumer Electronics Association, ASME, ASTM International, Qualcomm, Techstreet, and many others. Several speakers took the podium after dinner to welcome the 400 or so attendees and to present the World Standards Day Paper Competition Awards and the Ronald H. Brown Standards Leadership Award.

Dr. Patrick Gallagher, Under Secretary of Commerce for Standards and Technology, and Director of NIST, was the first speaker after dinner. He directed his remarks at the value of a decentralized, voluntary, and demand-driven system of standards in promoting innovation and economic prosperity. Gallagher emphasized that "standards provide the common language that keeps domestic and international trade flowing," concluding that "it is difficult to overestimate their critical value to both the U.S. and global economy."

James Shannon, President of the National Fire Protection Association (NFPA), accepted the R. H. Brown Standards Leadership Award in recognition for his work initiating or improving the National Electrical Code, the Life Safety Code, and the Fire Safe Cigarette and Residential Sprinkler Campaigns.

Ellen Emard, President of SES, introduced the paper competition award winners. As of this writing the titles and authors of the first and second place awards are not yet available on the SES web site (http://www.ses-standards.org/displaycommon.cfm?an=1&subarticlenbr=56). Third place was awarded to William P. Fisher, Jr., for his paper, "What the World Needs Now: A Bold Plan for New Standards." Where the other winning papers took up traditional engineering issues concerning the role of standards in advancing safety and sustainability issues, Fisher's paper spoke to the potential scientific and economic benefits that could be realized by standard metrics and common product definitions for outcomes in education, health care, social services, and environmental resource management. All three of the award-winning papers will appear in a forthcoming issue of Standards Engineering, the journal of SES.

Fisher was coincidentally seated at the dinner alongside Gordon Gillerman, winner of third place in the 2004 paper competition (Gillerman, 2004) and currently Chief of the Standards Services Division at NIST. Gillerman has a broad range of experience in coordinating standards across multiple domains, including environmental protection, homeland security, safety, and health care. Having recently been involved in a workshop focused on measuring, evaluating, and improving the usability of electronic health records (Gillerman, 2011). Gillerman was quite interested in the potential Rasch measurement techniques hold for reducing data volume with no loss of information, and so for streamlining computer interfaces.

Robert Massof of Johns Hopkins University accompanied Fisher to the dinner, and was seated at a nearby table. Also at Massof's table were several representatives of the National Institute of Building Sciences, some of whom Massof had recently met at a workshop on adaptations for persons with low vision disabilities. Massof's work equating the main instruments used for assessing visual function in low vision rehabilitation could lead to a standard metric useful in improving the safety and convenience of buildings.

As is stated in educational materials distributed at the World Standards Day celebration by ANSI, standards are a constant behind-the-scenes presence in nearly all areas of everyday life. Everything from air, water, and food to buildings, clothing, automobiles, roads, and electricity are produced in conformity with voluntary consensus standards of various kinds. In the U.S. alone, more than 100,000 standards specify product and system features and interconnections, making it possible for appliances to tap the electrical grid with the same results no matter where they are plugged in, and for products of all kinds to be purchased with confidence. Life is safer and more convenient, and science and industry are more innovative and profitable, because of standards.

The point of Fisher's third-place paper is that life could be even safer and more convenient, and science and industry could be yet more innovative and profitable, if standards and conformity assessment procedures for outcomes in education, health care, social services, and environmental resource management were developed and implemented. Rasch measurement demonstrates the consistent reproducibility of meaningful measures across samples and different collections of construct-relevant items. Within any specific area of interest, then, Rasch measures have the potential of serving as the kind of mediating instruments or objects recognized as essential to the process of linking science with the economy (Fisher & Stenner, 2011b; Hussenot & Missonier, 2010; Miller & O'Leary, 2007). Recent white papers published by NIST and NSF document the challenges and benefits likely to

Report of meeting by ANSI:
"**U.S. Standards Community Celebrates World Standards Day 2011**"
http://webstore.ansi.org/NewsDetail.aspx?NewsGuid=590a225c-d779-4f81-804e-4d05ef239c37

be encountered and produced by initiatives moving in this direction (Fisher, 2009; Fisher & Stenner, 2011a).

A diverse array of Rasch measurement presentations were made at the recent International Measurement Confederation (IMEKO) meeting of metrology engineers in Jena, Germany (see RMT 25 (1), p. 1318). With that start at a new dialogue between the natural and social sciences, and with the award in the World Standards Day paper competition, the U.S. and international standards development communities have shown their interest in exploring possibilities for a new array of standard units of measurement, standardized outcome product definitions, standard conformity assessment procedures, and outcome product quality standards. The increasing acceptance and recognition of the viability of such standards is a logical consequence of observations like these:

"Where this law [relating reading ability and text difficulty to comprehension rate] can be applied it provides a principle of measurement on a ratio scale of both stimulus parameters and object parameters, the conceptual status of which is comparable to that of measuring mass and force. Thus...the reading accuracy of a child...can be measured with the same kind of objectivity as we may tell its weight" (Rasch, 1960, p. 115).

"Today there is no methodological reason why social science cannot become as stable, as reproducible, and hence as useful as physics" (Wright, 1997, p. 44).

"...when the key features of a statistical model relevant to the analysis of social science data are the same as those of the laws of physics, then those features are difficult to ignore" (Andrich, 1988, p. 22).

Rasch's work has been wrongly assimilated in social science research practice as just another example of the "standard model" of statistical analysis. Rasch measurement rightly ought instead to be treated as a general articulation of the three-variable structure of natural law useful in framing the context of scientific practice. That is, Rasch's models ought to be employed primarily in calibrating instruments quantitatively interpretable at the point of use in a mathematical language shared by a community of research and practice. To be shared in this way as a universally uniform coin of the realm, that language must be embodied in a consensus standard defining universally uniform units of comparison.

Rasch measurement offers the potential of shifting the focus of quantitative psychosocial research away from data analysis to integrated qualitative and quantitative methods enabling the definition of standard units and the calibration of instruments measuring in that unit. An intangible assets metric system will, in turn, support the emergence of new product- and performance-based standards, management system standards, and personnel certification standards. Reiterating once again Rasch's (1960, p. xx) insight, we can acknowledge with him that "this is a huge challenge, but once the problem has been formulated it does seem possible to meet it."

William P. Fisher, Jr.

Andrich, D. (1988). Rasch models for measurement. (Vols. series no. 07-068). Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills, California: Sage Publications.

Fisher, W. P.. Jr. (2009). Metrological infrastructure for human, social, and natural capital (NIST Critical National Need Idea White Paper Series). Washington, DC: National Institute for Standards and Technology. Retrieved 25 October 2011 from http://www.nist.gov/tip/wp/pswp/upload/202_metrological_infrastructure_for_human_social_natural.pdf.

Fisher, W. P., Jr., & Stenner, A. J. (2011a, January). Metrology for the social, behavioral, and economic sciences (Social, Behavioral, and Economic Sciences White Paper Series). Washington, DC: National Science Foundation. Retrieved 25 October 2011 from http://www.nsf.gov/sbe/sbe_2020/submission_detail.cfm?upld_id=36.

Fisher, W. P., Jr., & Stenner, A. J. (2011b). A technology roadmap for intangible assets metrology. In Fundamentals of measurement science. International Measurement Confederation (IMEKO), Jena, Germany, August 31 to September 2.

Gillerman G. (2004) Making the Confidence Connection: Conformity Assessment System Design. http://www.ses-standards.org/associations/3698/files/WSD%202004%20-%203%20-%20Gillerman.pdf

Gillerman (G. 2011) Collaboration and Consensus through Standards – The National Technology Transfer and Advancement Act. http://www.nist.gov/healthcare/usability/upload/EHR-Usability-Workshop-2011-6-03-2011_final.pdf

Hussenot, A., & Missonier, S. (2010). A deeper understanding of evolution of the role of the object in organizational process. The concept of 'mediation object.' Journal of Organizational Change Management, 23(3), 269-286.

Miller, P., & O'Leary, T. (2007, October/November). Mediating instruments and making markets: Capital budgeting, science and the economy. Accounting, Organizations, and Society, 32(7-8), 701-34.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests (Reprint, with Foreword and Afterword by B. D. Wright, Chicago: University of Chicago Press, 1980). Copenhagen, Denmark: Danmarks Paedogogiske Institut.

Wright, B. D. (1997, Winter). A history of social science measurement. Educational Measurement: Issues and Practice, 16(4), 33-45, 52 [http://www.rasch.org/memo62.htm]

The Effects of Local Item Dependence on Estimates of Ability in the Rasch Model

One of the most crucial assumptions of psychometric theory is that the relationship between items is accountable to a specific latent trait. However, a major issue in psychometrics is what happens when items continue to relate with each other, after accounting for their contribution to the latent trait. In the context of the Rasch model, this relationship is termed Local Item Dependence (LID), and represents a prerequisite and assumption of the model. Violation of this assumption states that there is still some covariation between items, although the relationship of each item to the latent trait has been accounted for. The issue of local item dependency relates strongly to the issue of unidimensionality (as that covariation could be easily explained by the presence of a second factor), but could also relate to other sources of measurement error (e.g., situational such as fatigue or rater effects). Furthermore, violation of this assumption has major implications regarding the validity of estimates (e.g., on discrimination) of the Rasch model (Tuerlinckx & de Boeck, 2001; Yen, 1993). The assumption has been described in mathematical form by Tuerlinckx and de Boeck (2001):

$$\Pr(X_1 = x_1,\ldots..,X_k = x_k \mid \theta) = \prod_{i=1}^{k} \Pr(X_i = x_i \mid \theta)$$

Journal of Applied Measurement
Vol. 12, No. 2, 2011

A Comparison between Robust *z* and 0.3-Logit Difference Procedures in Assessing Stability of Linking Items for the Rasch Model, *Huynh Huynh and Anita Rawls, 96-105*

Assessment of English Language Development: A Validity Study of a District Initiative, *Juan D. Sanchez, 106-123*

Equating of Multi-Facet Tests Across Administrations, *Mary Lunz and Surintorn Suanthong, 124-134*

Examining Student Rating of Teacher Effectiveness using FACETS, *Naraihan Mat Daud and Noor Lide Abu Kassim, 135-143*

Exploring Differential Item Functioning (DIF) with the Rasch Model: A Comparison of Gender Differences on Eighth Grade Science Items in the United States and Spain, *Tasha Calvert Babiar, 144-164*

Understanding Rasch Measurement: A Mapmark Method of Standard Setting as Implemented for the National Assessment Governing Board, *E. Matthew Schulz and Howard C. Mitzel, 165-193*

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

This implies that the association between items (adjacent or not) should be zero. In that case, the true latent score of a person should equal the observed estimate. The purpose of the present paper was to evaluate the effects of local item dependence on the ability parameters of a spatial test involving a series of Chinese tangrams (i.e., puzzle). Participants were 94 university students with a major in psychology who completed the Chinese tangrams in response for extra credit. The specific hypothesis posited was that item difficulties of the Puzzle would be overestimated in the presence of local item dependence. The present illustration involves the manipulation of local item dependency on two puzzles only, for simplicity. Prerequisite analyses involve evaluation of the presence of local dependency. Within the framework of Hierarchical General Linear Modeling (HGLM), and as recommended by Johnson and Raudenbush (2006), this evaluation involved examination of the within-person variance σ^2 under the Bernoulli model (with the expectation being that $\sigma^2 = 1$). Applications of both restricted and full maximum likelihood procedures indicated that σ^2 was significantly lower compared to the expectation of the Bernoulli model (actual value of σ^2 was equal to 0,52, after rounding for both solutions), suggesting the presence of local item dependencies.

To evaluate the effects of local item dependence, a Rasch model was initially estimated using the Bernoulli function in Hierarchical Generalized Linear Modeling (HGLM). As Kamata (2002) demonstrated the following two-level HGLM model shown below is equivalent to the Rasch model:

Level-1 (Bernoulli) Model:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}X_{1j} + \beta_{2j}X_{2j} + \ldots\ldots\beta_{(k-1)j}X_{(k-1)j}$$

Level-2 model expressing person estimates:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ . \\ . \\ \beta_{(\kappa-1)j} = \gamma_{(\kappa-1)0} \end{cases}$$

The term p_{ij} reflects the probability that person *j* will answer item *i* correctly. The term X_{ij} describes the *ith* dummy variable for participant *j*. Last, the term β_{oj} reflects the intercept of the model (as in dummy variable regression) and β_{1j} through β_{kj} the coefficients of puzzle items X_1 through X_k. The random term u_{0j} is the error around the intercept, which is expected to be normally distributed (with a mean of zero and variance equal to τ). When the above two-level model is applied to the data of person *j* for a specific item *i*, the probability of that person responding correctly to item *i* is expressed as:

$$p_{ij} = \frac{1}{1 + \exp\{-[u_{0j} - (-\gamma_{t0} - \gamma_{00})]\}}$$

The following two level HGLM model was tested in order to estimate item abilities of the Rasch model:

Level 1

Prob (Response$_{ij}$ = 1 | $\beta_j = \varphi_{ij}$
Log [φ_{ij} / 1 - φ_{ij}) = η_{ij}
$\eta_{ij} = \beta_{0j} + \beta_{1j}$ (Puzzle1) + β_{2j} (Puzzle2) + β_{3j} (Puzzle3) + β_{4j} (Puzzle4) + r_{ij}

Level 2

$\beta 0j = \gamma 00 + u0j$
$\beta 1j = \gamma 10$
$\beta 2j = \gamma 20$
$\beta 3j = \gamma 30$
$\beta 4j = \gamma 40$

As shown above, only four puzzles are included in the model with the 5th one being represented by the intercept. The above model was compared to the model below in order to account for the presence of local dependence between puzzles 4 and 5. However, first there is a description of the interaction model used to account for the interaction between the two items. The model has been referred to as the *constant interaction model* (Tuerlinckx & de Boeck, 2001) because the interaction is presumably equal in magnitude across all participants. It is expressed in the following function:

$$\Pr(X_1 = x_1, X_2 = x_2 \mid \theta) = \frac{\exp(x_1(\theta - \beta_1) + x_2(\theta - \beta_2) + x_1 x_2(-\beta_{12})}{1 + \sum_{j=1}^{2} \exp((\theta - \beta) + \exp(\theta - \beta_1) + (\theta - \beta_2) - \beta_{12})}$$

The above model applies to two binary items denoted by the numbers 1 and 2. The responses to the items are seen as a realization of a bivariate random variable (X_1, X_2) and for a particular realization (x_1,x_2) the model is as shown above. The term β_{12} expresses the interaction between puzzles 4 and 5.

The estimated HGLM model employed to account for the above dependency was the following:

Level 1

Prob (Response$_{ij}$ = 1 | $\beta_j = \varphi_{ij}$
Log [φ_{ij} / 1 - φ_{ij}) = η_{ij}
$\eta_{ij} = \beta_{0j} + \beta_{1j}$ (Puzzle1) + β_{2j} (Puzzle2) + β_{3j} (Puzzle3) + β_{4j} (Puzzle4) + β_{5j} (Puzzle5) + r_{ij}

Level 2

$\beta 0j = \gamma 00 + u0j$
$\beta 1j = \gamma 10$
$\beta 2j = \gamma 20$
$\beta 3j = \gamma 30$
$\beta 4j = \gamma 40$
$\beta 5j = \gamma 50$

The difference between the two HGLM models is on their intercepts. In the Rasch model the intercept expressed the last item (as in dummy regression) whereas in the conditional independence assumption model the intercept represented the interaction (local dependency) between puzzle 4 and 5.

Figure 1 shows the effects of local item dependence on the difficulty levels of puzzle 4 and 5. It is obvious from item 5, the after controlling for local item dependency (i.e., its relation with item 4), the estimated difficulty of the item went down. This finding agrees with the theses of Douglas, Kim, Habing, and Gao (1998) who stated that the difficulty of the item is affected by the interacting item (5 in our case) and not the first item of the interaction (item 4). This finding also agrees with the suggestions of Thissen, Steinberg, and Mooney (1989) who stated that when local item dependencies are positive, and are not accounted for, theta values are greatly overestimated (Yen, 1993 reported the same finding, attributed it to the underestimation of the standard errors of measurement). Tuerlinckx and de Boeck (2001), put it more intuitively: "If two items interact positively, they provide less information than two independent items." (p. 186). That is, if they are treated as being independent their information regarding the latent trait is greatly overestimated. This effect is shown on the puzzle's total response functions when accounting for or ignoring local dependence (Figure 2). The curves on Figure 2 show that at higher levels of ability (i.e., last two puzzles) the two forms become more and more different. Similar information is provided by the Test Information Functions (TIFs) of the two forms, with differences being observed at higher levels of ability (theta).

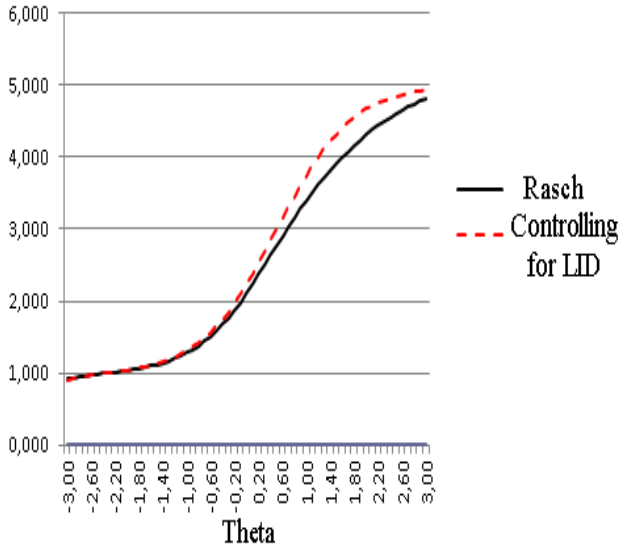


Figure 2. Average (Test) response functions for the Puzzle using (a) the Rasch model (straight line) and (b) the Rasch model which has been controlled for the presence of LID at items 4 and 5 (dashed line).

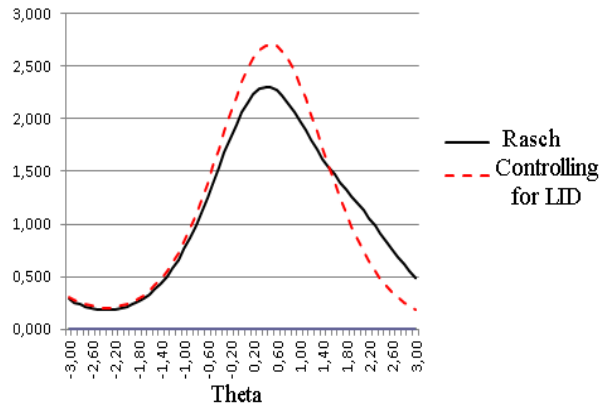


Figure 3. Test information functions for the Puzzle using (a) the Rasch model (straight line) and (b) the Rasch model which has been controlled for the presence of LID at items 4 and 5 (dashed line)

Summary

The purpose of the present paper was to evaluate the effects of local item dependence on the ability parameters of a series of puzzle. Results indicated that the effects of local independence are substantial and likely inflate the ability estimates of items at a given scale. Thus, the presence of LID seriously distorts the *qualities* of the items. Ideally, researchers should examine and control for the presence of LID. In HGLM, one can allow for under-dispersion in order to correct for local item dependency.

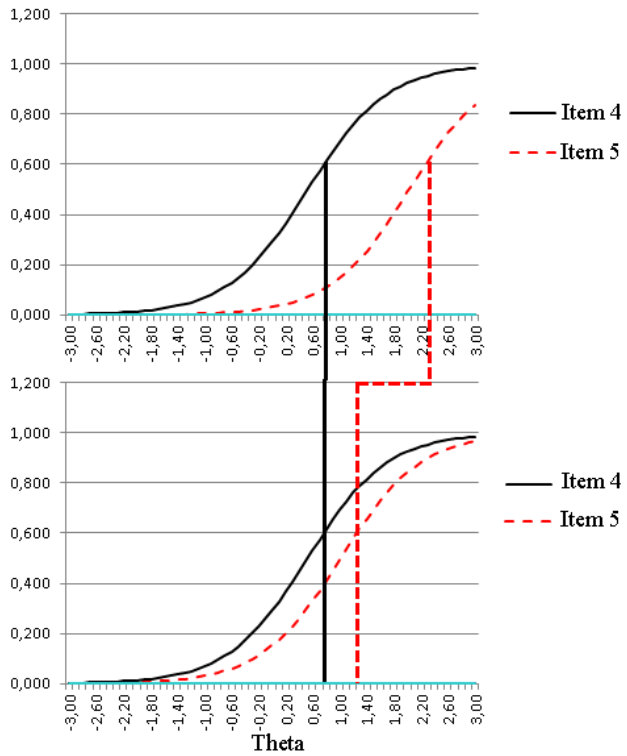


Figure 1. Item response functions for items 4 and 5 or the Rasch model (upper panel) and the model controlling for conditional independence (lower panel).

In the Rasch model one can estimate the likelihood ratio test, but as Tuerlinckx and de Boeck (2001) reported the test has little power and can reveal either large numbers of interacting items or extreme interactions (thus, it leaves several cases of LID undetected). It is concluded that LID represents a serious psychometric nuisance and should be evaluated at all times. It is suggested that hybrid Rasch models should be implemented to account for its deleterious effects on the quality of the items.

Georgios D. Sideridis
University of Crete

Douglas, J., Kim, H., Habing, B., Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics, 23,* 129-151.

Johnson, C., & Raudenbush, S. (2006). A repeated measures multilevel Rasch model with application to self-reported criminal behavior. In C. Bergeman, & S. Boker, (Eds.), *Methodological issues in aging research.* Mahwah, NJ: Lawrence.

Kamata, A. (2002, April). *Procedure to perform item response analysis by hierarchical generalized linear model.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26,* 247-260.

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6,* 181-195.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187-213.

Rasch Measurement SIG

2012 Nominations for Officers are Closed

Thank you to all who have submitted nominations for SIG Chair and SIG Secretary/Treasurer. We have forwarded these to AERA in accordance with their regulations for conducting SIG elections. We anticipate that a ballot will be conducted early in 2012.

At the AERA 2012 SIG Business Meeting, the incoming SIG Officers will appoint the SIG Program Chair(s) and Editor *of Rasch Measurement Transactions.*

Michael J. Young, SIG Chair
Kenneth Royal, SIG Secretary/Treasurer

Comparing Item Calibration Estimates using Winsteps and RUMM2010

Stable estimates of item calibration are important and the results should be consistent using different software programs for research to progress satisfactorily. Items tend to greater consistency although produced by person responses (Fisher, 2010). The database of 3,121 persons and 26 items for the KCT-R (Stone, 2002) produced a test separation value of 38.22 corresponding to a test reliability of 0.99. The person separation value was 2.54 corresponding to a person reliability of 0.87. A sample of 260 persons ages 5 – 60 produced a test-retest coefficient of 0.96.

Two programs for producing estimations – Winsteps (Linacre, 2002) and RUMM (Andrich, D., et al., 2000) – were utilized. A well-defined variable and wide item/ability logit range made the KCT-R data useful for this comparison. Newer releases provide additional enhancements, but the basic estimation algorithms remain constant.

Figure 1 is a plot of item calibration estimations using Winsteps and RUMM2010. This figure accents the differences. The Winsteps estimates were lower than the estimates using RUMM2010 for items 1 – 12, and slightly higher for items 13 – 26. A difference of about 0.8 logits was observed for the two easiest items of the test. These differences decreased to item 12. They reversed thereafter with the difference continuing to about 0.3 for the most difficult items.

Figure 2 identifies the stability in the estimations. The lowest four calibration estimates show the largest deviations that diminish towards 0.0 and then increase with the highest two calibration estimates showing the greatest difference. In spite of the noted differences between the two software programs, the item calibration estimates are highly correlated. The wide-range of the calibration estimates contributes to a high r^2 of 0.99, but with Winsteps estimates slightly more extreme than RUMM2010 estimates. Stability exists between the two estimations processes with the differences well within any meaningful difference. Nevertheless, a slight difference exists in the estimation process.

Futoshi Yumoto
American Institutes for Research
Collaborative for Research on Outcomes and -Metrics

and

Mark Stone, Aurora University, Aurora, IL

Andrich, D., Lyne, A., Sheridan, B., & Luo, G. (2000). RUMM 2010: Rasch Unidimensional Measurement Models. Australia: Rumm Laboratory Try Ltd.

Fisher, W. P., Jr. (2010, 30 September). Distinguishing between consistency and error in reliability coefficients: Improving the estimation and interpretation of information on measurement precision. Social Science Research Network. Retrieved from http://ssrn.com/abstract=1685556.

Linacre, J. M. (2003). Winsteps. Beaverton Oregon: Winsteps.com.

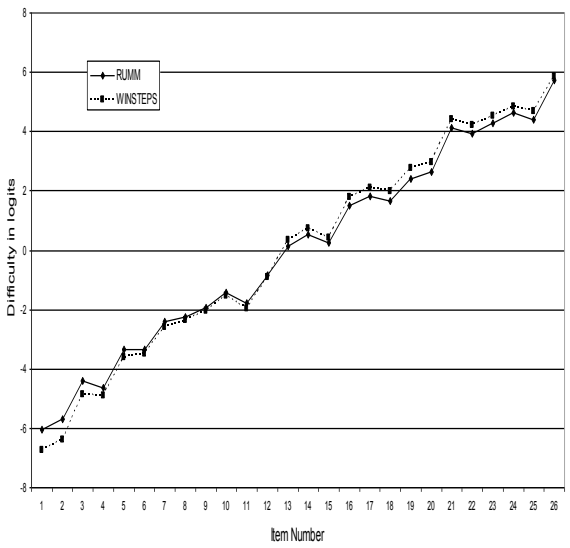


Figure 1. A plot of item calibration estimates using Winsteps and RUMM2010. Points on the solid line are RUMM estimates. Points on the dashed line are Winsteps estimates. The y-axis is "Difficulty in Logits"

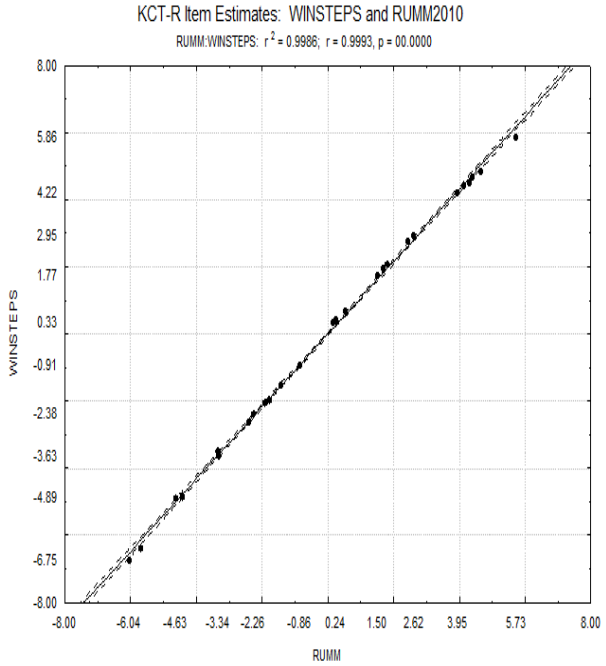


Figure 2. A cross-plot of the KCT-R items calibration estimates using Winsteps (y-axis) and RUMM2010 (x-axis) .

Explaining Discrepancies between the Sum of Subtest Scale-Scores and Total Scale-Scores

When high-stakes examinations are administered, examinees are keenly interested in the accuracy of their test scores, especially when their scores are close to, but below, the pass/fail cutpoint. When exam blueprints are available that outline various content percentages, test-takers will often attempt to reverse engineer their score reports in such a way that they can evaluate the extent to which the sum of their subtest scores is congruent with the overall test score. In some instances, the weighted sum of the subtest scores will be higher than the total score. This discrepancy may give test-takers a seemingly legitimate reason to question the accuracy of the overall score, thus prompting a phone call to the Examiners.

Fortunately, much of the problem can be easily explained as simply due to the range of the reported scale. For instance, if scores are reported on a scale that ranges from 200 to 800, extreme scores (<200 or >800) will be reported as either 200 or 800. Typically, this is the reason for the discrepancy. Examinees, however, will likely be unaware that the actual scale extends beyond the reported range. Therefore, when asked for assistance in interpreting scores, it is helpful to immediately ask if the examinee had any extreme scores on any of the subtests. If the answer is "yes" (and usually it is), the aforementioned explanations should suffice for the inquiring test-taker. For instance, if the pass-fail score is 500, and three equally weighted sub-scores are 700, 600, 200, then is looks like the total score should be 500 (a pass), but, in fact, it may be reported as 490 (a fail), because the reported 200 corresponds to an actual 170..

Issues of reported-score granularity, such as reporting by rounded increments of 5 or 10 may also contribute to a discrepancy. So that reported sub-scale scores of 600, 500, 400 (apparently averaging 500) may actually be 597, 497, 397 producing a rounded total score of 495.

On the other hand, if extreme scores and rounding are not the culprits, then it is helpful to have Wright (1994) and sometimes Bowles (1999) readily available to assist with explaining the technical aspects of the phenomenon in excruciating detail. Usually, the caller does not really care about the mechanics of subtest scoring, but merely wishes to argue his/her case for passing. Having the detailed and technical answer ready for discussion will usually dissuade callers from continuing down that path and hopefully will allow them to refocus on questions that are more salient to their future success on the examination.

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

Bowles, R. (1999). Combining and dropping subtest measures. Rasch Measurement Transactions, 13(1), p. 686. http://www.rasch.org/rmt/rmt131f.htm

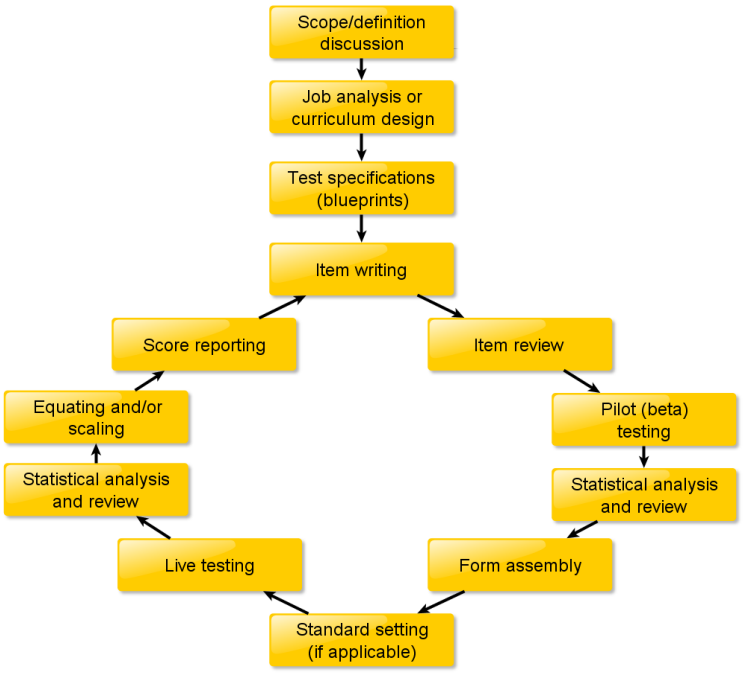
Wright, B. (1994). Combining part-test vs whole-test measures. Rasch Measurement Transactions, 8(3), p. 376. http://www.rasch.org/rmt/rmt83f.htm

Most Published Research Findings Are False!

"Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias."

Ioannidis JPA (2005) Why most published research findings are false. PLoS Med 2: e124. doi: http://dx.doi.org/10.1371/journal.pmed.0020124

Best practices in test development follow a circular process



Reproduced from
http://www.fasttestweb.com/test-development-cycle.php
with permission of Nathan A. Thompson, Ph.D., Vice President, Assessment Systems Corporation.

Rasch Measurement Transactions
www.rasch.org/rmt
Editor: John Michael Linacre

SIG Chair: Michael Young
Secretary: Kenneth Royal
Program Chairs: Daeryong Seo & Stephen Jirka
SIG website: www.raschsig.org