Google's PageRank Algorithm and the Rasch Measurement Model

"PageRank" is a widely-acclaimed algorithm used for determining the ranking or ordering of web pages by web search engines (Langville & Meyer, 2006; Chung, 2008). It was originally developed at Stanford University by Larry Page and Sergey Brin, who later started Google, Inc (Page & Brin, 1998). The algorithm has become a highly celebrated mathematical tool for analyzing networks of all kinds, including biological and social networks (Chung, 2008). The algorithm was the topic of a featured talk at the 2008 Joint Mathematics Meetings (joint meeting of the Mathematical Association of America and The American Mathematical Society) in San Diego.

The PageRank algorithm has much in common with one particular algorithm for estimating the item parameters of the Rasch model – the eigenvector (EVM) algorithm described by Garner and Engelhard (2002, etc.). Both methods depend on pairwise comparisons. Both methods also use the eigenvector of a matrix derived from pairwise comparisons.

An Illustrative Data Set

	Persons									
Items	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	1	1	1	1	1	1
2	0	1	0	1	1	1	1	0	1	1
3	1	0	0	1	1	0	1	1	1	0
4	1	0	0	1	0	1	1	0	1	0

We can represent this data matrix graphically as a directed multi-graph, shown here, with nodes representing items. An arrow (a directed edge) from node j to node i indicates that someone got item i correct and item j wrong. So, the four arrows between items 1 and 2 indicate that 3 people succeeded on item 1, but failed on item 2, and one person succeeded on item 2 but failed on item 1.

The Eigenvector Method (EVM)

The EVM begins with creation of a paired comparison matrix similar to that now implemented in RUMM2020. A paired comparison matrix is constructed. Each entry, $b_{ij,}$ representing the number of people who got item i right and item j wrong divided by the number of people who got item i wrong and item j right. This is an estimate of
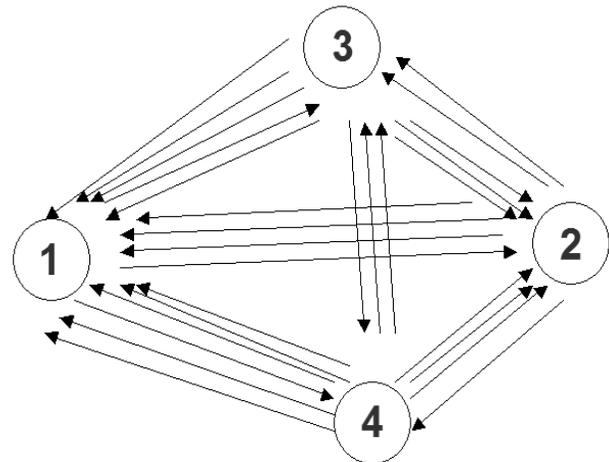
the difference in difficulties between items i and j on an odds-scale. b_{ii} are set to 1.

We can determine the Rasch item difficulties from this matrix by computing the eigenvector associated with the maximum eigenvalue, exponentiating the eigenvector, and subtracting out their mean, so that their sum is zero. The Rasch item difficulties are: -1.018, –.103, .268, .853.

The PageRank Algorithm

The PageRank algorithm is designed to assign weights to web pages so that the pages may be ranked in order of popularity. The web network we'll represent is the one depicted in Figure 1. The web pages A_1, A_2, A_3, and A_4 are represented by the four circles and an arrow would represent the fact that one web page links to the other web page. For example, web page A_2 has four links to web page A_1.

Each web site has a "PageRank" value associated with it. Let a_1, a_2, a_3, a_4 be those values. The PageRank value is conceptualized as the sum:

$$a_i = \sum_j \frac{M_{ij}}{N_j} a_j$$

N_j is the total number of links out of A_j to any other web pages (6 for A_2) and M_{ij} is the number of links out of A_j to A_i (3 from A_2 to A_1) Thus each page in the system contributes a portion of their value to the page they reference. Then, for our illustrative dataset,

$$a_1 = \frac{3}{6} a_2 + \frac{4}{8} a_3 + \frac{5}{10} a_4$$

$$a_2 = \frac{1}{3} a_1 + \frac{3}{8} a_3 + \frac{3}{10} a_4$$

$$a_3 = \frac{1}{3} a_1 + \frac{2}{6} a_2 + \frac{2}{10} a_4$$

$$a_4 = \frac{1}{3} a_1 + \frac{1}{6} a_2 + \frac{1}{8} a_3$$

A convenient way to solve these simultaneous equations is by matrix algebra and eigenvalues, adding a constraint to make the equations identifiable and the unknowns non-zero. A solution is $a_1 = -.65$, $a_2 = -.49$, $a_3 = -.45$, and $a_4 = -.36$. Thus, the easiest-to-link web page is A_1, followed by A_2, then A_3, then A_4.

Observations and Possibilities

There is much excitement over PageRank. Chung points out that PageRank is a "well defined operator on any given graph" (Chung, 2008), and describes the relationship between PageRank and random walks, spectral graph theory, spectral geometry, combinatorics, probability, and linear algebra. Langville and Meyer (2006) state that "models exploiting the Web's hyperlink structure are called link analysis models. The impact that these link analysis models have had is truly awesome."

PageRank webpage weights
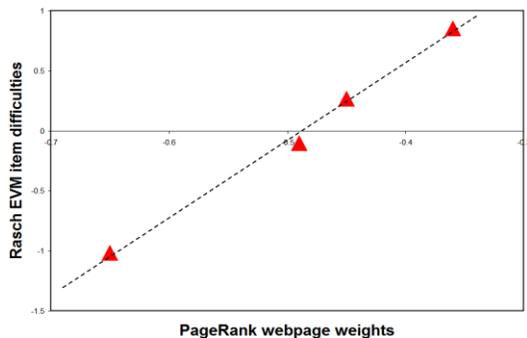
The plot shows the relationship between the Rasch item difficulties and the PageRank weights for our illustrative dataset. This suggests that the relationship is ogival, but close to linear for practical purposes. Notice that the coefficients of the PageRank equations are in the range 0-1, representing empirical probabilities. Rasch theory suggests that expressing those coefficients as log-odds, log(coefficient/(1-coefficent)), would be an immediate improvement to the PageRank equations. Here is a fruitful area of research for a graduate student looking for a dissertation topic.

Tools being developed for PageRank analysis and validation may well prove productive for Rasch measurement. These include techniques for accommodating missing and extreme comparisons, and also for identifying "cut vertices", web pages whose omission causes disconnected subsets of web pages in the analysis. These would correspond to elements (items, persons, raters) on which the linkage in judging plans, adaptive-testing or equating analysis depends.

Mary Garner, Kennesaw State University

Chung, Fan. (2008). The Mathematics of PageRank. Presentation at the Joint Mathematics Meetings. San Diego, CA. http://www.math.ucsd.edu/~fan/

Garner, M., & Engelhard, G. (2002). An Eigenvector Method For Estimating Item Parameters Of The Dichotomous And Polytomous Rasch models. Journal of Applied Measurement, 3(2), 107-128.

Langville, A.N., & Meyer, C.D. (2006). Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press.

Page, L., & Brin, S. (1998). The Anatomy of a Large Scale Hypertextual Search Engine. Proceedings of the Seventh International Web Conference. http://infolab.stanford.edu/pub/papers/google.pdf

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94. www.rasch.org/memo18.htm

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

SIG Officer Nominations

Dear Rasch SIG members,

I am writing to follow up on a previous email concerning the 2010 AERA General Election. As I mentioned in that email, Timothy Muckle and I will complete our terms as Rasch SIG Secretary/Treasurer and Chair, respectively, at the 2010 Annual Meeting. Because all SIG elections are now incorporated into the AERA General Election, and that process requires me to complete and submit a form for each position (Chair and Secretary/Treasurer) by November 16, 2009, the nominations process is now under way. We have received only one accepted nomination to date, so I am writing to solicit additional nominations.

If you know someone who you think would be interested in providing service to the SIG in the way of leadership or if you are interested in doing so yourself, **please send, via email, to** *ed.wolfe -\at/- pearson.com* **nominations for the offices of Chair and Secretary/Treasurer prior to November 1st, 2009.** Please include the individual's name, contact information, and the position for which that person is being nominated.

I will contact those who are nominated to confirm that the are willing to serve and to request a candidate statement prior to the November 16th deadline for submitting nominations to AERA.

The relevant sections of the SIG By-Laws are shown below, and they contain the following points:

There are two elected positions: Chair and Secretary/Treasurer.

* Elections take place via email balloting of the Rasch SIG members 3 months prior to the annual meeting.

* All SIG members are eligible to serve as officers.

* The term of each office is 2 years, commencing and expiring at the Annual AERA Meeting.

* No person shall serve more than 2 consecutive terms in a single office.

* This call for nominations is to be distributed electronically and published in the newsletter.

* The Chair shall be responsible for the general administration of the Rasch SIG and act as liaison between the SIG and AERA, shall preside at all meetings of the Executive Committee and at the annual business meeting, and shall appoint ad hoc committees as needed.

* The Secretary/Treasurer shall be responsible for the safe keeping of all financial documents and any official correspondence and meeting minutes of the Rasch SIG, will be responsible for maintaining the Rasch SIG website or appointing an appropriate representative as needed.

Edward W. Wolfe, Ph.D, Chair, Rasch SIG, AERA

Article VI: Officers

Section 1 -- General. The elected officers of the Rasch SIG ((1) Chair; (2) Secretary/Treasurer) shall be elected by a majority of SIG members voting every second year. Elections are to be conducted via e-mail balloting of Rasch SIG members three months prior to the appropriate AERA annual meeting. In the unlikely event of the failure to hold an e-mail ballot, officers will be elected by a majority of SIG members voting in an election held at the annual business meeting of the Rasch SIG. The elected officers and two other appointed officers ((3) Program Chair; and (4) Newsletter Editor) shall comprise the Executive Committee of the Rasch SIG and shall conduct all business of the SIG in the interim between the annual business meetings.

Section 2 -- Eligibility. All members in good standing of both AERA and the Rasch SIG shall be eligible for election as officers.

Section 3 -- Terms. The term of each office shall be for 2 years, expiring at the end of each year's annual AERA meeting. No person may serve in any single office for more than 2 consecutive terms. The newly elected chair shall in consultation with the secretary/treasurer appoint a (3) Program Chair; and (4) Newsletter Editor to be named at the AERA annual meeting.

Section 4 -- Offices. The following offices shall compose the executive committee of the Rasch SIG: (1) Chair; (2) Secretary/Treasurer; (3) Program Chair; and (4) Newsletter Editor. Persons to assist these officers or to carry out other work of the SIG may be appointed by the Chair.

Section 5 -- Election Procedures. Every second year prior to the appropriate annual AERA meeting the SIG Chair shall cause a call for nominations to be distributed electronically and published in the SIG Newsletter. When more than one valid nomination is received for either of the elected positions of SIG Chair or Secretary/Treasurer; the Chair shall be responsible for ensuring that an e-mail ballot of SIG members is conducted three months prior to the annual meeting. In the unlikely event of the failure to hold an e-mail ballot, officers will be elected by a majority of SIG members voting in an election held at the annual business meeting of the Rasch SIG. The Chair shall announce the outcomes of the elections to the Executive Committee and all candidates at least one week prior to beginning of the AERA annual meeting. Election shall require a majority of votes cast; a tied vote shall be broken by a show of hands conducted at the AERA annual meeting by the Chair. If the tie persists the Chair will exercise a deliberative vote.

Article VII: Duties of Officers

Section 1 -- Chair. The Chair (elected) shall be responsible for the general administration of the Rasch SIG and act as liaison between the SIG and AERA. The Chair shall preside at all meetings of the Executive Committee and at the annual business meeting. The Chair shall appoint ad hoc committees as needed.

Section 2 -- Secretary/Treasurer. The secretary/treasurer (elected) shall be responsible for the safe keeping of all financial documents and any official correspondence and meeting minutes of the Rasch SIG. The Secretary/Treasurer will also be responsible for maintaining the Rasch SIG website or appointing an appropriate representative as needed.

The Concept of a Measurement Mechanism

And in technology, as well as in basic science, to explain a fact is to exhibit the mechanism(s) that makes the system in question tick (Bunge, 2004, p. 182).

In 1557, the Englishman Robert Recorde remarked that no two things could be more alike (i.e., more equivalent), than parallel lines and thus was born the equal sign, as in $3 + 4 = 7$. Equation (1) is the familiar Rasch model for dichotomous data, which sets a measurement outcome (raw score) equal to a sum of modeled probabilities. The measurement outcome is the dependent variable and the measure (e.g., person parameter, b) and instrument (e.g., item parameters d_i's) are independent variables. The measurement outcome (e.g., count correct on a reading test) is observed, whereas the measure and instrument parameters are not observed but can be estimated from the response data. When a mechanismic interpretation[1] is imposed on the equation, the right-hand side (r.h.s.) variables are presumed to characterize the process that generates the measurement outcome on the left-hand side (l.h.s.). An illustration of how such a mechanism can be exploited is given in Stone (2002). The item map for the Knox cube test analysis had a 1 logit gap. The specification equation was used to build an item that theory asserted would fill in the gap. Subsequent data analysis confirmed the theoretical prediction of the Rasch relationship:

$$\text{Raw score} = \sum_i \frac{e^{(b-d_i)}}{1+e^{(b-d_i)}} \qquad (1)$$

Typically, the item calibrations (d_i's) are assumed to be known, and the measure parameter is iterated until the equality is realized (i.e., the sum of the modeled probabilities equals the measurement outcome). How is this equality to be interpreted? Are we only interested in the algebra or is something more happening?

Freedman (1997) proposed three uses for a regression equation like the one above:

1.1) To describe or summarize a body of data,

2.2) To predict the l.h.s. from the r.h.s,

1.3) To predict the l.h.s. *after manipulation* or *intervention* on one or more r.h.s. variables (measure parameter and/or instrument parameters).

Description and summarization possess a reducing property in that they abstract away incidentals to focus on what matters in a given context. In a rectangular persons-by-items data matrix (with no missing data), there are n_p x n_i observations. Equations like those above summarize the data using only $n_p + n_i − 1$ independent parameters. Description and summarization are local in focus. The relevant concept is the extant data matrix with no attempt to answer questions that might arise in the application realm[2] about "what if things were different." Note that if interest centers only on the description and summary of a

specific data set, additional parameters can be added, as necessary, to account for the data.

Prediction typically implies the use of the extant data to project into an as yet unobserved context/future in the application realm. For example, items from the extant data are used to compute a measure for a new person, or person parameters are used to predict how these persons will perform on a new set of items. Predictions like these rest on a set of claims of invariance. New items and new persons are assumed to behave as persons and items behaved in the extant data set. Rasch fit statistics (for persons and items) are available to test for certain violations of these assumptions of invariance (Smith, 2000).

Rasch models are probabilistic models that are fundamentally associational and thus cannot and do not, alone, support a causal interpretation of equation (1) (Woodward, 2003). Note that equation (1) can support a predictive interpretation if the equality is taken to satisfy a simple if-then condition. A causal interpretation of equation (1) requires successful predictions under manipulation of the measure parameter, the instrument parameters, or ideally, under conjoint manipulation of the two parameters. Conjoint manipulation up and down the scale directly tests for the trade-off property that holds only when the axioms of additive conjoint measurement are satisfied (Kyngdon, 2008).

To explain how an instrument works is to detail how it generates the count it produces (measurement outcome) and what characteristics of the measurement procedure affect that count. This kind of explanation is neither just statistical nor synonymous with prediction. Instead, the explanation entails *prediction under intervention*: if I wiggle this part of the mechanism, the measurement outcome will be different *by this amount.* As noted by Hedström (2005), "Theories based on fictitious assumptions, even if they predict well, give incorrect answers to the question of why we observe what we observe" (p. 108). Rasch models, absent a substantive theory capable of producing theory-based instrument calibrations, may predict how an instrument will perform with another subject sample (invariance) but can offer only speculation in answer to the question, "How does this instrument work?" Rasch models without theory are not predictive under intervention and, thus, are not causal models.

Measurement mechanism is the name given to just those manipulable features of the instrument that cause invariant measurement outcomes for objects of measurement that possess identical measures. A measurement mechanism explains by opening the black box and showing the cogs and wheels of the instrument's internal machinery. A measurement mechanism provides a continuous and contiguous chain of causal links between the encounter of the object of measurement and

instrument and the resulting measurement outcome (Elster, 1989). We say that the measurement outcome (e.g., raw score) is explained by explicating the mechanism by which the measurement outcome is brought about. In this view, to respond with a recitation of the Rasch equation for converting counts into measures, to reference a person by item map, to describe the directions given to the test-taker, to describe an item-writing protocol, or simply to repeat the construct label more slowly and loudly (e.g., extroversion), provides a nonanswer to the question, "How does this instrument work?"

Although the sociologist Peter Hedstr*ö*m (2005) was concerned with the improvement of macro theory, several of his reasons for favoring mechanistic explanations apply to measurement science in general:

2.1) Detailed specifications of mechanisms result in more intelligible explanations.

2.2) A focus on mechanisms rather than, for example, item types, reduces theoretical fragmentation by encouraging consideration of the possibility that many seemingly distinct instruments (e.g., reading tests) with different item types and construct labels may in fact share a common measurement mechanism.

2.3) The requirement for mechanistic explanations helps to eliminate spurious causal accounts of how instruments work.

Measurement mechanisms as theoretical claims make point predictions under intervention: when we change (via manipulation or intervention) either the object measure (e.g., reader experiences growth over a year) or measurement mechanism (e.g., increase text measure by 200L). The *mechanismic*[1] narrative and associated equations enable a point prediction on the consequent change in the measurement outcome (i.e., count correct). Notice how this process is crucially different from the prediction of the change in the measurement outcome based on the selection of another, previously calibrated instrument with known instrument calibrations. *Selection* is not *intervention* in the sense used here. Our sampling from banks of previously calibrated items is likely to be completely atheoretical, relying, as it does, on empirically calibrated items/instruments. In contrast, if we modify the measurement mechanism rather than select previously calibrated measurement mechanisms, we must have intimate knowledge of how the instrument works. Atheoretical psychometrics is characterized by the aphorism "test the predictions, never the postulates" (Jasso, 1988, p. 4), whereas theory-referenced measurement, with its emphasis on measurement mechanisms, says *test the postulates, never the predictions*. Those who fail to appreciate this distinction will confuse invariant predictors with genuine causes of measurement outcomes.

A Rasch model combined with a substantive theory embodied in a specification equation provides a more or less complete explanation of how a measurement instrument works (Stenner, Smith, & Burdick, 1983). A Rasch model in the absence of a specified measurement mechanism is merely a probability model; a probability model absent a theory may be useful for (1.1) and (1.2), whereas a Rasch model in which instrument calibrations come from a substantive theory that specifies how the instrument works is a causal model; that is, it enables prediction after intervention (1.3):

"Causal models (assuming they are valid) are much more informative than probability models: A joint distribution tells us how probable events are and how probabilities would change with subsequent observations, but a causal model also tells us how these probabilities would change as a result of external interventions. . . . Such changes cannot be deduced from a joint distribution, even if fully specified." (Pearl, 2000, p. 22)

A mechanismic narrative provides a satisfying answer to the question of how an instrument works. Below are two such narratives for a thermometer designed to take human temperature (3.1) and a reading test (3.2).

3.1) "The Nextemp thermometer is a thin, flexible, paddle-shaped plastic strip containing multiple cavities. In the Fahrenheit version, the 45 cavities are arranged in a double matrix at the functioning end of the unit. The columns are spaced 0.2^oF intervals covering the range of 96.0^oF to 104.8^oF. . . . Each cavity contains a chemical composition comprised of three cholesteric liquid crystal compounds and a varying concentration of a soluble additive. These chemical compositions have discrete and repeatable change-of-state temperatures consistent with an empirically established formula to produce a series of change-of-state temperatures consistent with the indicated temperature points on the device. The chemicals are fully encapsulated by a clear polymeric film, which allows observation of the physical change but prevents any user contact with the chemicals. When the thermometer is placed in an environment within its measure range, such as 98.6^oF (37.0^oC), the chemicals in all of the cavities up to and including 98.6^oF (37.0^oC) change from a liquid crystal to an isotropic clear liquid state. This change of state is accompanied by an optical change that is easily viewed by a user. The green component of white light is reflected from the liquid crystal state but is transmitted through the isotropic liquid state and absorbed by the black background. As a result, those cavities containing compositions with threshold temperatures up to and including 98.6^oF (37.0^oC) appear black, whereas those with transition temperatures of 98.6^o (37.0^oC) and higher continue to appear green" (Medical Indicators, 2006, pp. 1-2).

3.2) "The MRW technology for measuring reading ability employs computer generated four-option multiple choice cloze items "built on-the-fly" for any continuous prose text. Counts correct on these items are converted into Lexile measures via an applicable Rasch model. Individual cloze items are one-off and disposable. An

item is used only once. The cloze and foil selection protocol ensures that the correct answer (cloze) and incorrect answers (foils) match the vocabulary demands of the target text. The Lexile measure of the target text and the expected spread of the cloze items are given by a proprietary text theory and associated equations. A difference between two reader measures can be traded off for a difference in Lexile text measures to hold count correct (measurement outcome) constant. Assuming a uniform application of the item generation protocol the only active ingredient in the measurement mechanism is the choice of text with the requisite semantic (vocabulary) and syntactic demands."

In the first example, if we uniformly increase or decrease the amount of additive in each cavity, we change the correspondence table that links the number of cavities that turn black to a degree Fahrenheit. Similarly, if we increase or decrease the text demand (Lexile) of the passages used to build reading tests, we predictably alter the correspondence table that links count correct to Lexile reader measure. In the former case, a temperature theory that works in cooperation with a Guttman model produces temperature measures. In the latter case, a reading theory that works in cooperation with a Rasch model produces reader measures. In both cases, the measurement mechanism is well understood, and we exploit this understanding to answer a vast array of "W" questions (see Woodward, 2003): If things had been different (with the instrument or object of measurement), what then would have happened to what we observe (i.e., the measurement outcome)?

To explain a measurement outcome, "One must provide information about the conditions under which [the measurement outcome] would change or be different. It follows that the generalizations that figure in explanations [of measurement outcomes] must be change-relating. . . . Both explainers [e.g., person parameters and item parameters] and what is explained [measurement outcomes] must be capable of change, and such changes must be connected in the *right way.*" (Woodward, 2003, p. 234)

The Rasch model tells us the *right way* that object measures, instrument calibrations, and measurement outcomes are to be connected. Substantive theory tells us what interventions/changes can be made to the instrument to offset a change to the measure for an object of measurement to hold constant the measurement outcome. Thus, a Rasch model in cooperation with a substantive theory dictates the form and substance of permissible conjoint interventions. A Rasch analysis, absent a construct theory and associated specification equation, is a black box and "as with any black-box computational procedures, the illusion of understanding is all too easy to generate". (Humphreys, 2004, p. 132).

A. Jackson Stenner, Mark H. Stone, Donald S. Burdick

Footnotes

1. The term *mechanismic* was coined by Bunge (2004) to emphasize the nonmechanical features of some mechanisms.

2. In applied mathematics, we typically distinguish between the mathematical realm and the application realm.

References

Bunge, M. (2004). How does it work? The search for explaining mechanisms. *Philosophy of Social Science 34*, 182-210.

Elster, J. (1989). *Nuts and bolts for the social sciences.* Cambridge, MA: Cambridge University Press.

Freedman, D. (1997). From association to causation via regression. In V. McKim & S. Turner (Eds.), *Causality in crisis? Statistical methods and the search for causal knowledge in the social sciences* (pp. 113-161). Notre Dame, IN: University of Notre Dame Press.

Hedström, P. (2005). *Dissecting the social: On the principles of analytical sociology.* Cambridge, UK: Cambridge University Press.

Humphreys, P. (2004). *Extending ourselves: Computation science, empiricism, and scientific method.* New York: Oxford University Press.

Jasso, G. (1988). Principles of theoretical analysis. *Sociological Theory, 6, 1-20.*

Kyngdon, A. (2008). Conjoint measurement, error and the Rasch model: A Reply to Michell, and Borsboom and Zand Scholten. *Theory and Psychology, 18(1),125-131*

Kyngdon, A. (2008). The Rasch model from the perspective of the representational theory of measurement. *Theory and Psychology, 18(1), 89-109.*

Medical Indicators. (2006). [Technical paper]. www.medicalindicators .com/pdf/nt-fc-tech-bulletin.pdf

Pearl, J. (2000). *Causality: Models, reasoning, and inference.* Cambridge, MA: Cambridge University Press.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1(2), 199-218.*

Stenner, A. J., Smith, M., & Burdick, D. (1983). Toward a theory of construct definition. Journal of *Educational Measurement, 20 (4), 305-316.*

Stone, M. H. (2002). *Knox's cube test: Revised.* Wood Dale, IL: Stoelting.

Woodward, J. (2003). *Making things happen.* New York: Oxford University Press.

The Impact of Rasch Item Difficulty on Confirmatory Factor Analysis

Table 1: CFA of Sections 1 and 4 of the IELTS Listening Module

Model	χ^2	df	χ^2/df	NNFI	CFI	GFI	RMSEA	RMSEA 90% confidence interval
Section 1 (10 items)	29.68*	35	0.85	1.06	1.00	0.95	0.001	0.001 to 0.051
Section 1 (8 items)	10.54	20	0.53	1.12	1.00	0.98	0.001	0.001 to 0.001
Section 4 (10 items)	72.90**	35	2.08	0.94	0.95	0.88	0.100	0.067 to 0.130
Section 4 (9 items)	54.54*	34	1.60	0.96	0.97	0.91	0.076	0.036 to 0.110
Constraint tenable	Non-significant	-	< 2	.95	.90	.90	< 0.06	Narrow interval

Note. $n = 148$. **$p < 0.001$. *$p < 0.01$.
df = degree of freedom. NNFI = Non-Normed Fit Index. CFI = Comparative Fit Index.
GFI = Goodness of Fit Index. RMSEA = Root Mean Square Error of Approximation.

It has been argued that item difficulty can affect the fit of a confirmatory factor analysis (CFA) model (McLeod, Swygert, & Thissen, 2001; Sawaki, Sticker, & Andreas, 2009). We explored the effect of items with outlying difficulty measures on the CFA model of the listening module of International English Language Testing System (IELTS). The test has four sections comprising 40 items altogether (10 items in each section). Each section measures a different listening skill making the test a conceptually four-dimensional assessment instrument.

We observed two items with outlying low Rasch difficulty measures, but poor fit to the Rasch model, in section 1 (measure of item 8 = -1.71, infit MNSQ = 1.43; measure of item 9 = -1.59, infit MNSQ = 1.36) and an item with an outlying high Rasch difficulty measure, and good fit to the Rasch model, in section 4 (measure of item 38 = 3.01, infit MNSQ = 0.99). There was a large gap between these items and the rest of the items in each section on the Wright map.

Initially, we proposed separate CFA models for sections 1 and 4 to investigate the causes of variations in the measurements (items). In each model was a latent trait measured by 10 items. The 10-item CFA model for section 1 had a significant chi-square index (indicating rejection of the null hypothesis of one factor) although other fit indexes fell within the acceptable range (Table 1). The two outlying items did not load significantly on the latent trait at 5%. In a *post hoc* modification stage, we removed items 8 and 9 from the analysis and calculated the fit of the modified 8-item CFA model. We observed a noticeable improvement in the fit of the items to the one-factor model. We expected this because of the bad fit of the items to the Rasch model.

Likewise, we calculated the CFA model for section 4 with 10 items, which also did not exhibit acceptable fit indexes. Item 38 which had a high difficulty measure outlying from the rest of the items was deleted and a noticeably better fit to the one-factor model was obtained.

This was somewhat surprising, because the deleted item exhibited good fit to the unidimensional Rasch model.

This analysis is supportive of the results from previous studies which show item difficulty can affect the fit of the CFA models. Items with outlying difficulty measures can compromise the fit of CFA models. So, it may be useful that we delete items with outlying Rasch difficulty measures prior to conducting any CFA or in the *post hoc* modification stages.

S. Vahid Aryadoust
Nanyang Technological University, Singapore

References
McLeod, L.D., Swygert, K.A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 189-216). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sawaki, Y., Sticker, L.J., & Andreas, H.O. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing 26*(1), 5-30.

Test Theory Reference Materials Online
"The Reference Supplement to the Manual for relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)" is an online resource on test theory and standard setting, published by the Council of Europe at http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp which includes these sections:

B: Standard Setting by Felianka Kaftandjieva
C: Classical Test Theory by Norman Verhelst
D: Qualitative Analysis Methods by Jayanti Banerjee
E: Generalizability Theory by Norman Verhelst
F: Factor Analysis by Norman Verhelst
G: Item Response Theory (mostly Rasch) by Norman Verhelst
H: Many-Facet Rasch Measurement by Thomas Eckes
Thomas Eckes

Rasch-related Coming Events

Oct. 23 - Nov. 21, 2009, Fri.-Fri. Many-Facet Rasch Measurement online course (M. Linacre, Facets), www.statistics.com/ourcourses/facets

Nov. 4-6, 2009, Wed.-Fri. Applying The Rasch Model: Practical workshop in applying and interpreting Rasch analyses (T. Bond), Sydney, Australia. www.winsteps.com/sydney.htm

Nov. 13, 2009, Fri. IV Workshop de Modelos de Rasch en Administración de Empresas, Canary Islands, www.iude.ull.es

Nov. 24, 2009, Tues. Rasch Refresher workshop (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Nov. 25-27, 2009, Wed.-Fri. Introduction to Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Nov. 30-Dec 2, 2009, Mon.-Wed. Intermediate Rasch (A. Tennant, RUMM), Leeds, UK, www.leeds.ac.uk/medicine/rehabmed/psychometric

Dec. 15, 2009, Mon. Deadline for Abstracts for June 13-16, 2010, Mon.-Wed. International Conference: Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Copenhagen, Denmark, www.rasch2010.cbs.dk

Jan. 8 - Feb. 5, 2010, Fri.-Fri. Rasch - Core Topics online course (M. Linacre, Winsteps), www.statistics.com/ourcourses/rasch1

March 5 - April 2, 2010, Fri.-Fri. Rasch - Further Topics online course (M. Linacre, Winsteps), www.statistics.com/ourcourses/raschfurther

Apr. 28-29, 2010, Wed.-Thur. IOMW 2010 International Objective Measurement Workshop, Boulder, CO, USA, www.iomw2010.net

April 30 - May 4, 2010, Fri.-Tues. AERA Annual Meeting, Denver, CO, USA, www.aera.net

June 13-16, 2010, Mon.-Wed. International Conference: Probabilistic Models for Measurement in Education, Psychology, Social Science and Health, Copenhagen, Denmark, www.rasch2010.cbs.dk

Online Programs/Courses in Research Methodology

The *University of Illinois at Chicago* is offering two online programs in research methodology.

The first program is an eight course MEd in Measurement, Evaluation, Statistics, and Assessment (MESA). See http://education.uic.edu/mesaonline-med/ for details.

The second is an Educational Research Methodology (ERM) Certificate, which consists of a minimum of any three courses offered in the MESA online curriculum. Visit http://education.uic.edu/erm/ for more information.

Those interested in taking a course without entering a program can enroll as an Extramural Student.

The current courses are Essentials of Quantitative Inquiry in Education, Advanced Analysis of Variance and Multiple Regression, Multivariate Analysis of Educational Data, Educational Measurement, Rating Scale and Questionnaire Design and Analysis, Educational Program Evaluation, Assessment for Measurement Professionals, and Research Design in Education. We also anticipate adding courses in Hierarchical Linear Modeling and Classroom Assessment.

For course descriptions, see
education.uic.edu//mesaonline-ed/coursedescriptions.cfm

For when courses are offered, starting Spring 2010, see
education.uic.edu//mesaonline-med/onlineschedule.cfm

Students enrolled in these online courses come from a variety of backgrounds:

1) those already holding a PhD or EdD in a non-MESA field wanting to increase their research skills,

2) those who eventually wish to pursue doctoral studies in a MESA area,

3) current PhD and EdD students in a non-MESA field wanting to increase their research skills and career opportunities,

4) those who want to acquire the knowledge and technical skills needed for entry-level positions in academic institutions, state and federal agencies, school districts, and the testing and evaluation industry (e.g., licensure and certification boards, private and not-for-profit testing organizations), and

5) students enrolled in other graduate programs needing coursework they can transfer into their current degree program.

Everett Smith, Ph.D.
Associate Professor, Educational Psychology
Director, Measurement, Evaluation, Statistics, and Assessment Lab, http://education.uic.edu/mesalab/

Unidimensional Models in a Multidimensional World

Question: "Unidimensionality is one of the assumptions underlying most Rasch models. But everything we encounter is multidimensional. Why aren't all Rasch models multidimensional?"

Reply: The world is multidimensional and confusing. A fundamental activity of physical science is to decompose the world around us into unidimensional variables (weight, height, temperature, pressure, ...). Using these unidimensional variables, physicists can think clearly and make strong inferences. The history of the thermometer is an illustrative example of this process. Early thermometers (around 1600 A.D.) combined temperature with atmospheric pressure. They were "multidimensional". It was a major advance when scientists discovered how to separate those two dimensions in order to make both temperature and atmospheric pressure into unidimensional variables.

In Rasch measurement, we are attempting to perform the same process of splitting a multidimensional world into unidimensional variables, but now with social science. Asserting and then building unidimensional variables has been very useful in physical science. We expect it will also be in social science.

What if there are two dimensions?

Question: When we know there are two dimensions in the data, what is the next step - two separate analyses? Then, how can we make it sense out of the two analysis when we only want to report one number?

Answer: Under these circumstances, we need to consider:
1. How big is the difference between the dimensions?
2. How many people, and which people, does it impact?
3. Is it important enough to merit reporting two numbers?

This may require a separate analysis of each dimension. For instance, in an elementary-arithmetic test, we will probably find there is an "addition" dimension and a "subtraction" dimension. Unless the test is intended to identify learning difficulties, it is unlikely we will want to report two numbers. But the dimensionality may have useful information for instruction. In one situation, relatively bad performance on "subtraction" was discovered to be related to poverty. Children in poverty did not like the thought of something being "taken away" (subtracted). This suggests that teaching "subtraction" to impoverished children should avoid using emotive words or personal implications.

It is unusual in a carefully-constructed test that two dimensions are different enough inferentially to merit reporting two numbers. But secondary dimensions may indicate that care should be taken in test-construction in order to balance items between dimensions. For instance, aim for 50% addition items and 50% subtraction items, not 80% addition items and 20% subtraction items.
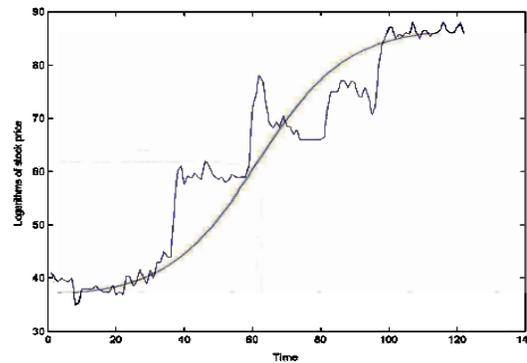
Model and empirical logistic ogives for stock prices. Figures 7 and 17, "Price trajectory for Charter Plc from 22nd April 2003 to 17th October 2003", in Silas N. Onyango (2007) *On the pattern recognition of Verhulst-logistic Itô Processes in Market Price Data.*

NIST Call for White Papers

The US *National Institute for Standards and Technology* has posted a new *Call for White Papers:*
   www.nist.gov/tip/call_for_white_papers_sept09.pdf
as part of its mission "to support, promote, and accelerate innovation in the United States through high-risk, high-reward research in areas of critical national need."

The White Papers are NIST's mechanism for collaborating with practitioners in the development of new areas of research into fundamental measurement and metrological systems. NIST is seeking out areas of measurement research that are not currently a priority and that have the potential for bringing about fundamental transformations in particular scientific areas. The *Call for White Papers* is not a funding opportunity itself, but a chance to influence the focus of future funding, such as expanding existing measurement methodologies into publicly recognized reference standards.

As was evident in its celebration of *World Metrology Day,* May 20, 2009, NIST is well aware of the human, economic, and scientific value of technical standards. Metrological standards for human, social, and natural capital have become an area of critical national need that could be highly rewarding. This is especially so when considered relative to the rewards that could accrue from order-of-magnitude improvements in the meaningfulness, utility, and efficiency of measurement based on ordinal observations.

Deadlines over the next year for White Papers are November 9, February 15, May 10, and July 12, though submissions will be accepted any time between November 9, 2009 and September 30, 2010.

A PDF of a White Paper that builds a case for Rasch-based metrological standards and that was submitted to NIST in its previous round is available at
 www.livingcapitalmetrics.com/images/FisherNISTWhitePaper2.pdf

William P. Fisher, Jr., Ph.D.
LivingCapitalMetrics.com

How Many Rating-Scale Categories?

"First, scales with two or three response alternatives are generally inadequate in that they are incapable of transmitting very much information and they tend to frustrate and stifle respondents.

"Second, the marginal returns from using more than nine response alternatives are minimal and efforts for improving the measurement instrument should be directed toward more productive areas.

"Third, an odd rather than an even number of response alternatives is preferable under circumstances in which the respondent can legitimately adopt a neutral position. Overuse of the neutral category by respondents can generally be avoided by providing them with an adequate number of reasonable response alternatives. *[Ben Wright argued that a neutral category allowed respondents to escape from making difficult or uncomfortable decisions.]*

"Fourth, even a few response alternatives may be too many for the respondent if comprehensible instructions and labeling of response alternatives are not included to enable the respondent to conceptualize and respond in spatial terms."

Cox E.P. III (1980) The Optimal Number of Response Alternatives for a Scale: A Review. Journal of Marketing Research, 17, 4, 407-422

DIF Sample Size for Polytomous Items

Scott, Fayers, Aaronson, et al. (2009) *A simulation study provided sample size guidance for differential item functioning (DIF) studies using short scales.* Journal of Clinical Epidemiology 62, 288-295, make the following recommendations (with many provisos):

Uniform DIF in polytomous items:
"Based on our results, as a general rule of thumb, we would suggest imposing a minimum of 200 respondents per group to ensure adequate performance. If the scale contains just two items, we would suggest a minimum of 300 respondents."

Non-uniform DIF in polytomous items:
500 respondents per group were not enough to detect non-uniform DIF reliably. Further, "it is difficult to know what amount of non-uniform DIF ... represents practically important non-uniform DIF as no published guidelines on this topic were identified."

Foundations of Measurement

suppes-corpus.stanford.edu/measurement.html
links to 18 downloadable video lectures on Measurement Theory. They were given in 1981 by Patrick Suppes, R. Duncan Luce, and Amos Tversky. Two of Duncan Luce's lectures are titled "Conjoint Measurement", reminding us of Luce, R. D. and J. W. Tukey. (1964). "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement," Journal of Mathematical Psychology, 1, 1-27.
Michael Lamport Commons

What is "Scaling" ?

"Scaling" is an ambiguous word in English, even in its psychometric usage.
"Scale" from *"scala"* (a ladder) is a means of "positioning objects in an ascending sequence (up a ladder)" - so it signifies "ordering".
"Scale" from *"skal"* (a bowl) is part of the pan-balance, "weight scales", "scales of justice" - so it signifies "quantification".
A "Guttman Scalogram" is a Guttman ordering, not a Guttman quantification. But "Rasch scaling" is a Rasch quantification, which includes a Rasch ordering, but only secondarily.

Probabilistic Models for Measurement in Education, Psychology, Social Science and Health
International Conference
13 - 16 June 2010
Copenhagen, Denmark

50 years since the publication in 1960 of Georg Rasch's "Probabilistic Models for Some Intelligence and Attainment Tests"

Call for Abstracts
Deadline is Tuesday, December 15, 2009

Submit your abstract at www.rasch2010.cbs.dk

Keynote speakers
• **David Andrich**, The University of Western Australia
• **Denny Borsboom**, University of Amsterdam, The Netherlands
• **Klaas Sijtsma**, Tilburg University, The Netherlands

Conference Tracks
• Psychometrics & statistics, education, social sciences (psychology, sociology, business) and health.

Conference topics
• The history of Rasch and IRT models
• Local response dependence and multi-dimensionality
• Objectivity and invariance
• Analysis of DIF
• Development of measures for cross-cultural comparisons
• Validity of measurement
• Reliability of objective measurement
• Test equating and linking
• Computer adaptive testing
• Item banking

The Organizing Committee
• John Brodersen, Associate Professor, University of Copenhagen, Denmark
• Svend Kreiner, Professor, University of Copenhagen, Denmark
• Tine Nielsen, Senior Advisor on Education, Copenhagen Business School, Denmark

Measurability does not demonstrate existence

"Valid measures are often taken, albeit implicitly, as proof that the assumed variable really does exist. Suppose one could attain evidence of the unidimensionality and linearity of the *QoL* scores from a questionnaire: again, this would still not be evidence that the measurable variable named *QoL* is **QoL**. Naming a variable is a matter of perspective: it relates to the meaning the variable is assigned, rather than to its intrinsic properties."

Tesio, L. (2009) Quality of life measurement: one size fits all. Journal of Medicine and the Person (2009) 7:5–9

Rasch Measurement And Sociological Theory

"Have you ever pondered the ambiguity of "and" in titles? Here I mean, "Rasch Measurement, a Challenge to Sociological Theory." The challenge is to take seriously a measurement model that is attractive in the light of commonly observed patterns in data and also for its fundamental logical and statistical properties. Taking it seriously will mean exploring carefully the conceptual consequences of the assumptions that all responses are probabilistic and that it is possible to separate the measurement of personal traits (such as attitudes) and the measurement of social objects (such as questionnaire items or social. entities or social values))."

Otis Dudley Duncan (1982) Rasch Measurement And Sociological Theory. Lecture at Yale University.
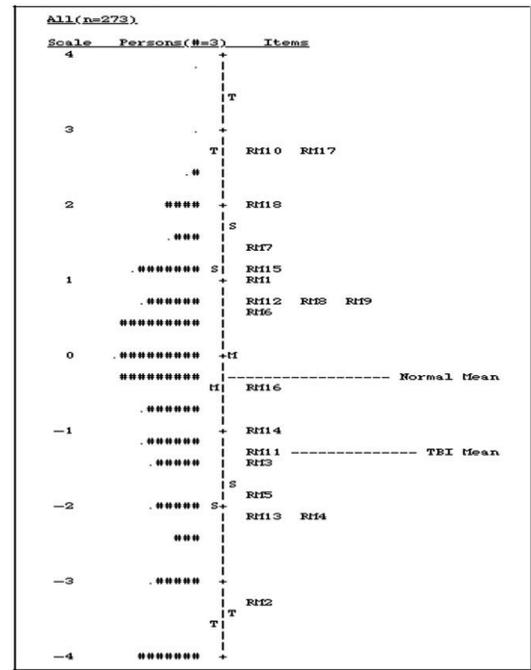
Wright map from Prieto, Gerardo, Delgado, Ana R., Perea, Maria V. and Ladera, Valentina (2009) Scoring Neuropsychological Tests: Using the Rasch Model: An Illustrative Example With the Rey-Osterrieth Complex Figure, The Clinical Neuropsychologist.

Web KIDMAP

Figure 1 in Tsair-Wei Chien, Weng-Chung Wang, Sho-Be Lin, Ching-Yih Lin, How-Ran Guo and Shih-Bin Su (2009) KIDMAP, a web based system for gathering patients' feedback on their doctors. *BMC Medical Research Methodology 2009, 9:38*
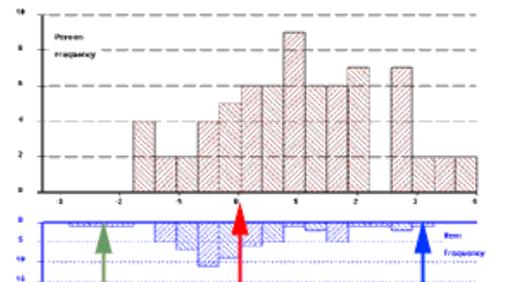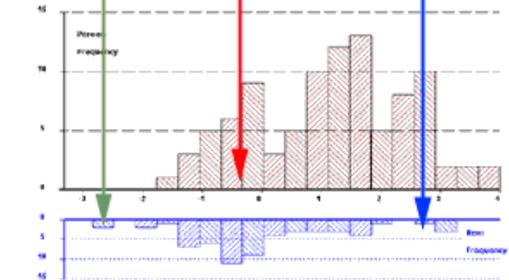
Two tests equated by common person/item linking. Figure in Yu, Chong Ho & Sharon E. Osborn Popp (2005). *Test Equating by Common Items and Common Subjects: Concepts and Applications.* Practical Assessment Research & Evaluation, 10(4). Available online:

http://pareonline.net/getvn.asp?v=10&n=4

Theoretical Complexity vs. Rasch Item Difficulty in Reading Tests

The concept of unidimensionality of reading comprehension (Weir & Porter, 1994) has led scholars to believe that there might be a one-to-one correspondence between item difficulty and the level of cognition the item measures (Alderson, 1990). It is commonplace among reading specialists to divide reading ability into different layers of cognition such that hypothetically labeled lower layers are assumed to be followed by higher ones (Alderson, 1990). The hierarchy assumption is so appealing that tests developers usually calibrate items solely in terms of item difficulty, while ignoring issues related to their level of cognition. Yet, it is often the case that more difficult items represent lower order abilities (at least as predicted by theory) than do easier ones (Weir and Porter, 1994). Paradoxically, harder items seem to contribute *less* to reading ability than do easier ones (Meyer, 1975, cited in McNamara, 1996).

Weir and Porter (1994) suggest that the main reason for limiting the reproducibility assumption to item difficulty in test constructions is 'practical expediency rather than ... a principled view of unidimensionality' (p. 9). Because empirical item hierarchies sometimes contradict theoretical notions of reading comprehension (McNamara, 1996; Weir & Porter, 1994; Alderson, 1990), we approach the issue from a qualitative as well as a quantitative perspective:

1. Does there exist a one-to-one correspondence between item difficulty and the nature of the latent ability the item measures?

2. To what extent do variations in item difficulty reflect qualitative rather than quantitative item differences?

To address these questions we used the SBRT - Forms a and b - which are (mostly) multiple-choice item language tests. The SBRT was developed at the Iran University of Science and Technology (IUST) (Daftarifard, 2000) using over 200 intermediate students for each form. As is shown in Table 1, the SBRTa contains 39 questions that address twenty-four abilities that are frequently referred to in the literature. Items' hypothetical cognitive complexity is indicated by the ordinal number in the last column of this table. The classification of some items is uncertain (e.g., answering factual questions might either be classified as perception or speed reading).

Reading ability as a hierarchy

The results in Table 1 and Figure 1 reveal a clear lack of correspondence between item complexity and the hypothetical level of cognition. Some supposedly cognitively demanding abilities turned out to be less difficult than less cognitively demanding abilities, and some item types are out of order. This is summarized by the finding that the Spearman rank correlation between items' Rasch locations and their hypothetical complexity is just 0.22. Moreover, the average locations for items in complexity groups 1 or 2, 2, 3, 3 or 4, 4, 4 or 5, and 5, are -2.8, -0.2, 0.4, 0.4, 0.8, 1.2, and -0.1 logits, respectively.

The existence of one-to-one relation between empirical (i.e., Rasch) and hypothetical complexity follow is contradicted in many ways. For instance, DFH2 (distinguishing between fact and hypothesis) is harder than IN2 (inferencing), while RF2 (understanding the rhetorical function of the text) is easier than LT1 (understanding the literal meaning). Similarly, the presumably more complex skill of understanding the factual question (here FQ1) is much easier than mere text scanning (both SCB and SCE). Also, skimming (SK1) turns out to be more difficult than SK2 (Rasch measure -0.23) although both belong to speed reading category. Certain items which hypothetically measure higher ability like interpretation ability turn out to be much easier than lower level items like speed reading (here SK1 with the Rasch measure of 0.57). These include items AU1 with Rasch measure of -1.65, CT1 with the Rasch measure of -1.24, MI1 with the Rasch measure of -0.52, TP2 with the Rasch measure of -0.42, RF1 with the Rasch measure of -0.40, and IN2 with the Rasch measure of -0.40.

Another problem found in the data pattern concerns items with the same operational definition but with quite different item difficulties. These items include understanding the audience of the text, i.e., AU1 and AU2 with two different consecutive Rasch measures -1.65 and 0.39, CD1 and CD2 with two consecutive Rasch measures of 0.84 and -2.32, ED1 and DE2 with two different Rasch measures of -0.23 and 0.60 respectively, PA1 and PA2 with two consecutive different Rasch measures of -0.03 and 0.98, and TP1 and TP2 with the Rasch measures of 1.07 and -0.42 respectively. Among these, however, there are only a few items that operationally belong to one category and turn up with almost similar measure like SI1
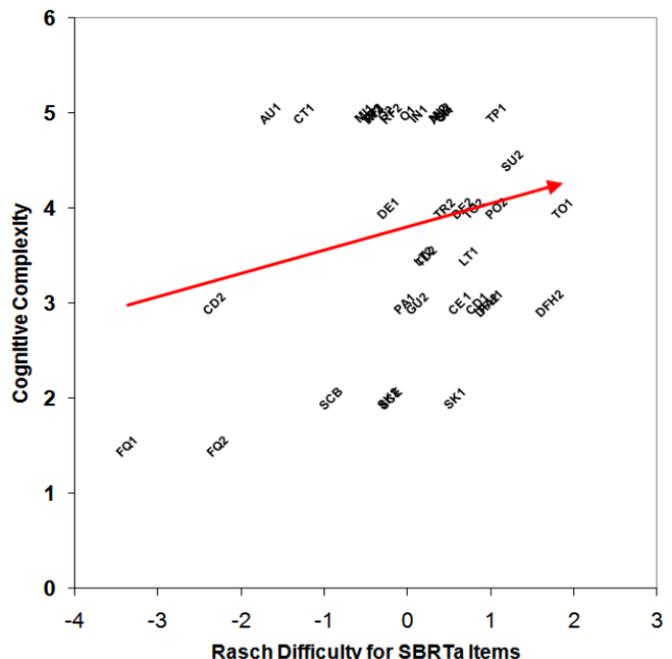
Figure 1. Theoretical Complexity vs. Rasch Difficulty.

and SI2 with Rasch measure of 0.45 and 0.43.

We point out that the findings summarized above cannot be attributed to the particular set of items being used. Firstly, the SBRTa items fit the Rasch model adequately (only one item's outfit exceeds 1.3), thus establishing this test-form's measurement validity. Second, in support of unidimensionality, factor analysis of items' Rasch residuals indicates that just three items (SI1, AU2 and SK2) loaded higher than 0.5 on the most prominent residual factor. Third, the SBRTa shows acceptable classical reliability (coefficient alpha = 0.82). Fourth, students' SBRTa measures are highly correlated with their measures on the widely used IELTS (exemplar, 1994, the academic version of module C , $r = 0.71$, $p < .001$). Finally, the lack of correlation between items' hypothetical and empirical difficulties is replicated for the second test form, the SBRTb. Similar to the value observed for the SBRTa, the rank correlation for the SBRTb is just 0.23.

The present findings thus indicate that while reading is unidimensional and hierarchical, this hierarchy disagrees with theoretical predictions in the literature (for an overview, see e.g., Alderson, 1990). Given this lack of correspondence, we propose that notions of items complexity require careful distinctions between the qualitative and quantitative aspects of reading theory. For instance, it may be necessary to distinguish between the complexity of a concept and the complexity of the question designed to assess this concept. Rasch scaling is likely to remain the tool of choice in this research, but it seems likely that multi-facetted approaches will be needed to accommodate both types of complexity simultaneously.

Parisa Daftarifard
Rense Lange, Integrated Knowledge Systems

References

Alderson, J. C. (1990). Testing reading comprehension skills (part one). *Reading in a Foreign Language, 6(2),* 425-438.

Daftarifard, P. (2002). *Scalability and divisibility of the reading comprehension ability*. Unpublished master's thesis. Tehran, Iran: IUST.

McNamara, T. (1996). Measuring Second Language Performance. New York: Addison Wesley Longman.

Weir C. J., & Porter D. (1994). The Multi-Divisible or Unitary Nature of Reading: The language tester between Scylla and Charybdis. *Reading in a Foreign Language, 10(2),* 1-19.

Editor's Note: These findings contrast with the remarkable success of the Lexile system at predicting the Rasch item difficulty of reading-comprehension items. See Burdick B., Stenner A.J. (1996) Theoretical prediction of test items. *Rasch Measurement Transactions,* 1996, 10:1, p. 475.
www.rasch.org/rmt/rmt101b.htm

Table 1
Items' Rasch Difficulty and Hypothetical difficulty (SBRTa)

	Skills to be measured	Code	Rasch Difficulty	Cognitive Complexity **
1.	Scanning and information search	SCB	-0.92	2
		SCE	-0.20	2
2.	Skimming	SK1	0.57	2
		SK2	-0.23	2
3.	Guessing	GU2	0.12	3
4.	Understanding the factual questions	FQ1	-3.37	1 or 2
		FQ2	-2.28	1 or 2
5.	Interpreting cohesive devices	CD1	0.84	3
		CD2	-2.32	3
6.	Paraphrasing	PA1	-0.03	3
		PA2	0.98	3
7.	Distinguishing between the facts and hypothesis	DFH1	0.98	3
		DFH2	1.71	3
8.	Distinguishing between cause and effect	CE 1	0.63	3
9.	Deduction	DE1	-0.23	4
		DE2	0.66	4
10.	Paragraph organization	PO2	1.07	4
11.	Transcoding information	TR2	0.45	4
12.	Text organization	TO1	1.87	4
		TO2	0.80	4
13.	Understanding the source of the text	SI1	0.45	5
		SI2	0.43	5
14.	Understanding the function of the text	RF1	-0.40	5
		RF2	-0.19	5
15.	Understanding the audience of the text	AU1	-1.65	5
		AU2	0.39	5
16.	Understanding the opinion of the author	O1	0.00	5
		O2	-0.26	5
17.	Choosing the best title for the text	CT1	-1.24	5
18.	Inference	IN1	0.14	5
		IN2	-0.40	5
19.	Choosing Title for paragraph	TP1	1.07	5
		TP2	-0.42	5
20.	Choosing the main idea of the text	MI1	-0.52	5
		MI2	0.37	5
21.	Understanding the propositional meaning (syntactical meaning or literal meaning)	LT1	0.74	3 or 4
		LT2	0.20	3 or 4
22.	Text diagrams	TD2	0.23	3 or 4
23.	Summarizing ability	SU2	1.26	4 or 5

** Numbers in the last column stand for the following in increasing complexity: (1) Perception, (2) Speed Reading, (3) Word-based reading, (4) Analyzing, (5) Interpretation.

A Rasch-Informed Standard Setting Procedure

Items are written for the different levels of ability that exist in the institution. Alternatively, judges assign already written items to the existing levels of ability. In other words, they decide what ability level a person should be to answer each item. If there are say, five levels of proficiency from A to E, A being the highest and E the lowest, then the items are rated on a five point scale from 1-5. Five corresponding to the lowest level, E and 1 to the highest level, A. For instance, if the judges agree that "border line" Level B students can answer an item correctly the item is rated 2 and if they envisage that for answering the item a test-taker should be at least a border line Level A student then the item is rated 1. The average judge ratings of an item is considered as its final difficulty estimate. All the items are rated in this way.

Afterwards, the items are administered to a group of test-takers and then Rasch analyzed and the location calibrations for the items are estimated. **The success and preciseness of the standard setting procedure heavily depends on the accordance between the judge-envisaged item difficulties and empirical student-based item difficulties. Any standard setting procedures in which this accordance is not achieved is futile.**

If the judges have done their job properly then there must be a correspondence between the empirical item estimates and judge-based item difficulties. Figure 1 shows the item estimates hierarchy on an item-person map. The Level A items are clustered at the top of the map and the other levels' items are ordered accordingly. However, there are some items which are misplaced. It is obvious that judge-intended levels of the items never correspond exactly with the Rasch measures. For instance, as can be seen in Figure 1, Rasch has reported some A items down in the B region (or below) and some B items up in the A region (or above).

Standard-setting always requires a compromise between the judges' item hierarchy and the empirical (Rasch) item hierarchy which corresponds to actual examinee performance. Standard-setting also requires negotiation about the location of the criterion levels. There will be several reasonable positions for the criterion level, from least-demanding to most demanding.

We might choose the transition points to be the lines at which the minimum number of items are misclassified between two adjacent levels. For example, the transition point between Level A and Level B is the point where the items predominantly become Level B items (as is done in Figure 1). That is, the difficulty level of item 18A or 97B which is 1.53 logits.

Or we might choose the line corresponding to 60% chances of success on the items that fall in the transition points determined by the procedure above. For example, the items at the transition points between Level A and Level B have a difficulty estimate of 1.53 logits. This is an item of borderline difficulty. In other words, an ability estimate of 1.53 logits can be the minimum cut-off score for Level A. This is the ability level required to have 50% chances of getting this item right. To be on a safe side, one can also define:

"cut-off score" = 60% chances of success on an item of borderline difficulty

Therefore, the cut-off score for Level A will be:

$$P_{ni}(X_i=1 \mid \theta_n, \delta_i) = \exp(\theta_n - \delta_i) / (1 + \exp(\theta_n - \delta_i))$$

$$0.60 = \exp(\theta_n - 1.53) / (1 + \exp(\theta_n - 1.53))$$

$$\log_e(1.5) = \theta_n - 1.53$$

$$\theta_n = 1.93$$

The cut-off scores for the other levels can be determined in a similar way. The items at the point of transition between Level B and Level C are 15C, 41B, 4B, 70B, 81B, 82B with difficulty estimates of 0.68 logits. Therefore the cut-off score for Level B can either be 0.68 logits, if we consider the 50% chances of success on the items at the transition point, or 1.08 logits if we consider 60% chances of success at the transition point.

Purya Baghaei

```
 5           +
             |
             |   46A
             |
             |   92B     99A
 4           +
             |T
             |   86A
             |
             |   79A
             |
 3           +   88A
             |
             |   51A     75A     94A
             |
             |   23A     61A     74A     85B     90A     95A
           T |   47A     6B      87A     89A
 2     XX    +
       XX   |S 27A
       X    |   59B
       XXXXX |   18A     97B     102A    104A    107B    110A
       XXXX  |
       XXXXXXX |  22B    53A     66B     98B
 1     XXXXXXX +  19A    24B     60B     69B     83C     93B
             |   9B
       XXXXX S|  52B     84B
       XXX   |   15C     41B     4B      70B     81B     82B
     XXXXXXXXXXX |  14C   56C     80C
       XXXXX |   43C
    XXXXXXXXXXXX |  34C   62C
 0     XXXXXXXXX +M 33B   40C     54C     7C      96C
     XXXXXXXXXXX |  48C   58C
       XXXXXXX M|  32C    49B
       XXXXXXXX |  36C    42C     73C
       XXX   |   2D      31C     101C    105C    103C    109D
       X     |   25D     5C      72D     91D
 -1    XXXXXXXX +
       XX    |   37D     68D
       XXXX  |   100D    17D     65D
     XXXXXXXX |   13D     16D     64D     71D
       XX   S|   10D     39D     63D     106E    108D    111D
       XX   |S 12E       50D     76E     77D
 -2    XXXXX +   11E     67E
       XXXX  |   35E     57D     78E     8E
       XXX   |   29E     44D
       XXXX  |   20E     21E     3E
       XXX   |   38E
       X    T|
 -3    XX    +   1E
             |   26E     28E
       X     |   45E
       XX   |
             |   30E     55E
            |T
 -4          +
```

Figure 1: Difficulty order of items and their judge-based corresponding levels