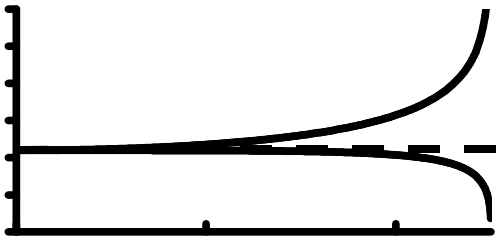
Combining Sub-Test Measures



RASCH MEASUREMENT

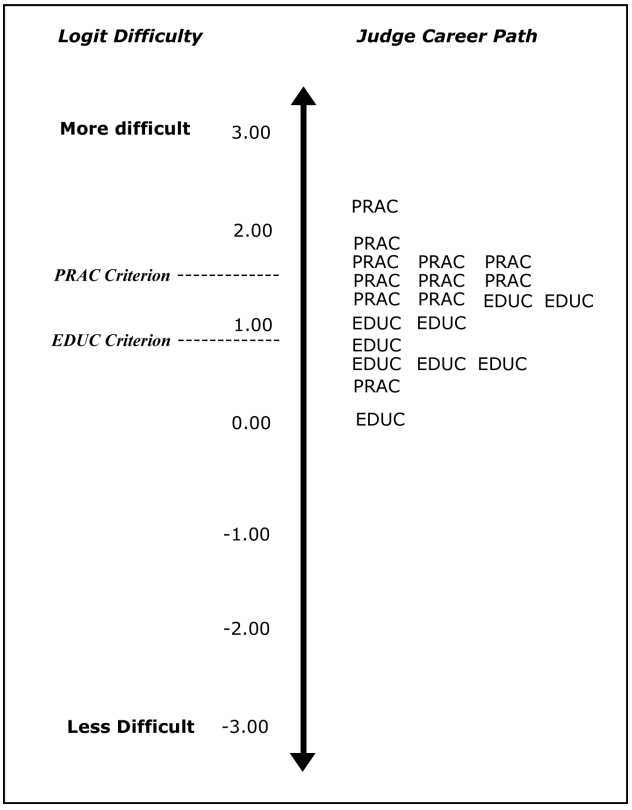
Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 17 No. 2 Autumn 2003 ISSN 1051-0796

A Standard of Importance

The establishment of passing standards is a critical component of a successful examination program. The models available for setting standards vary greatly in their methodological frameworks, yet each, whether acknowledged or not, is ultimately an evaluative process that includes the use of some form of measurement or statistical assistance, but is not defined by it. As with any human endeavor, the sample of participants used greatly influences the outcome. In the context of standards this suggests that **who sets standards for passing examinations is as important to the outcome as is the choice of standard setting methodology itself.** A recent study in the field of high-stakes medical examinations reveals this phenomenon quite well.

Figure One: Distribution of Criterion Points - Practitioners (PRAC) versus Educators (EDUC)



The study was conducted with a national medical board in charge of a high-stakes certification testing program. The board employed the Rasch-derived Objective Standard Setting model to set the passing standard for the examination. The board consisted of 20 members. Of these members, 10 considered themselves to be primarily practitioners (PRAC) of medicine, while the remaining 10 considered their primary occupation to be that of an educator (EDUC) at a university or hospital training program.

Participants in the exercise began to define their criterion in the traditional Objective manner. After an extensive group discussion about the meaning of minimal competence and the essentiality of items, each member was presented with a complete, previously calibrated examination. The members individually reviewed each item and assessed the content and taxonomic conveyance included. Members would then decide for themselves whether the content as presented in each item was essential for an entry-level practicing physician to understand. Ultimately individual sets of core items were defined whose mean item difficulties represented the quantification of the content selected by each member participant.

An inspection of the criteria proved interesting. There is a statistically significant difference that is apparent even on simple visual inspection of Figure 1. The practitioners are noticeably stratified above the educators. There is an obvious gap between the criterion (mean = 1.52 logits) established by the practitioner members and the criterion

Table of Contents

(mean = 0.94 logits) established by the educator members.

High-stakes testing plays a critical role in the career of hopeful students. It also provides a measure of safety for our society. The selection of participant members on high-stakes boards must be carefully considered. In our case the question became, whose standard should be adopted? Practitioners are clearly closer to patient care, but educators may sometimes have a broader curricular focus. Should boards require a certain mixture?

While the use of a multi-faceted approach would account for differences in rater severity, it would not eliminate the more fundamental question of legitimate definitional differences. Indeed, while standard setters debate and discuss the merits of methodology, they cannot afford to ignore that most basic of confounding variables – the sample of participants selected.

Gregory E. Stone
The University of Toledo

Note: Wright & Grosse (RMT 7:3, 315) point out that "failing the possibly incompetent" requires a higher standard than "passing the probably competent" . Perhaps in Figure One, practitioners are subconsciously relatively more concerned with protecting patient well-being, while educators are relatively more concerned with enhancing student careers.

Rasch Measurement SIG Chair and Secretary Call for Nominations

A new SIG *Chair* and *Secretary* commence their 2 year terms at the Spring 2004 AERA Meeting. They must be AERA and SIG members. The **Chair** oversees SIG activities, represents the SIG to AERA and chairs the Annual Business Meeting at AERA. The **Secretary** oversees the SIG mailing list and bank account.

Appointed SIG Officers are the **Program Chair**, who manages the selection process for papers to be presented at AERA, and the **Editor of Rasch Measurement Transactions**, who compiles the SIG's quarterly publication.

Please email *Chair* and *Secretary* nominations to the address below before December 1, 2003. Please include a paragraph about your nominee to be published in *RMT*. Self-nomination is welcomed. Balloting will be by email in December-January.

Trevor Bond
Rasch Measurement SIG Chair
Trevor.Bond@jcu.edu.au

2nd International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch and Unfolding Models Perth, Australia, 2004

www.education.murdoch.edu.au/educ_RaschJanuary2004.html

January 5 - 9	Introductory five day course on Rasch measurement. Weekend of January 10 -11 free. Includes use of the program RUMM2020.
January 10	Course Barbecue.
January 12 - 16	Advanced five day course in Rasch measurement and unfolding models. Weekend of January 17 - 18 free. Includes use of the programs RUMM2020, RATEFOLD.
January 19 9.00 a.m. - 4.00 p.m.	One day workshop focusing on using RUMM2020.
January 19 5.00 p.m. - 7.00 p.m.	Conference Registration.
January 20 - 22	**Conference** papers on applications of Rasch and related measurement models in any substantive field of application - education, psychology, health care and rehabilitation, marketing, etc. Papers on theory and history of the Rasch measurement are also welcome.
January 21	Conference dinner at the Nedlands Golf Club, located two miles from the city of Perth, and overlooking the Swan River - native home of black swans.

IOMW-XII: An Australian Contribution

June 30th, Wednesday – July 2nd, Friday, 2004
Cairns, Queensland, Australia

The Twelfth International Objective Measurement Workshop will focus on developments in four professional strands as well as the usual more general Rasch-based presentations.

Education Chair:	**Psychology** Chair:	**Business** Chair:	**Health Care** Chair:
Juho Looveer	Karen M. Schmidt	Thomas Salzberger	Robert W. Massof

Conference Registration Fee:
$75.00 US international delegates; $75.00 AUD for Australian & New Zealand delegates
$50 AUD for full-time students; $20.00 AUD for Teachers' Day

Abstracts deadline: 31 January 2004 Acceptances notified: 28 February 2004.
Registration process & proposal submission will be conducted online:
Website: www.soe.jcu.edu.au/iomw2004/ email: iomw@jcu.edu.au

Delegates requiring early acceptance in order to commit funds, seek employer support / approval, etc. should email: iomw@jcu.edu.au

Proposals after the due date may be accepted, subject to program / accommodation availability.

IOMW Pre-Conference Workshops:
Learn the features of the software from the people who wrote it:
Winsteps & Facets - June 28 Monday afternoon - Linacre
RUMM - June 29 Tuesday morning - Sheridan / Andrich
ConQuest - June 29 Tuesday afternoon - Adams / Wu / Wilson

Teachers' Workshops: July 2, Friday

Accommodation: Just 15 minutes drive north of Cairns, Quest Marlin Cove Resort is situated in the lush tropical surrounds of Cairns' favorite beach, Trinity Beach. The Resort comprises of a combination of one, two and three bedroom apartments and offers the convenience of Hotel Service with the flexibility and value of apartment living. Special rates are offered for IOMW delegates. Please book now. Email jindorato@questapartments.com.au for details (mention **IOMW** - request shared accommodation, if you wish). Website: www.questapartments.com.au – Far North Queensland.

Transport: The Resort is 18 km (11 miles) from Cairns International Airport (CNS) served by Qantas (codesharing with American Airlines). Rental cars available (Australia drives on the left). Much more about local transport and amenities at www.cairns.aust.com/about/trinity.htm

Weather: Rain: very little. Sunshine: 7+ hours per day. Temperature: 17ºC - 26ºC daily (63ºF – 78ºF). Humidity: low.

IOMW-XII Chair: Trevor Bond

IOMW II is supported by the School of Education, James Cook University

Psychometric Entities

Since Thales (6th century BC) first proposed substituting empirical ideas for spiritual explanation, rigid thinkers have attempted to restrict and control scientific thinking. For instance, consider the rule: "When theories and facts are in conflict, the theories must yield" (Simon, 1989). Perhaps this is true ultimately, but certainly not immediately. Such a rule could reduce science to merely summarizing the current empirical data, decrying as automatically invalid any theory that can be contradicted in any way by "hard facts".

The reoccurring attempt to impose dogma on empirical methods and its persistent rejection by scientists are summarized by Feyerabend below:

> *Not a single rule, ... however firmly grounded in epistemology, ... is not violated at some time or other .... atomism in antiquity, the Copernican Revolution, the rise of modern atomism (kinetic theory; dispersion theory; stereochemistry; quantum theory), and the gradual emergence of wave theory occurred only because some thinkers either* ***decided*** *not to be bound by certain 'obvious' methodological rules, or because they* ***unwittingly broke*** *them. (Feyerabend, 1978, p. 230-231)*

The best defense of scientific inference is a fundamental understanding of its empirical implications. This is as true of Rasch measurement as of any other theory. Michell claims that "no psychological attribute can ever be shown to be measurable" (2002). I concur with Graves (2003) in his emphasis that the more important issue is "how well these putative measurements .... relate to other behaviors and other theories" (p. 915). This more constructive approach to advancing social science aims at better understanding the influence of measurement on "psychometric knowledge", and specifically on psychometric and conceptual entities.

Conceptual entities are fictions or "working hypotheses" invented by scientists to explain dynamic regularities, and they are typically expressed as mathematical abstractions. When they lead to empirical predictions, both entities and abstractions acquire material status and become scientifically important. Molecular movement, for example, was mere speculation until the empirical demonstration of Brownian motion.

In psychometrics, researchers may fail to realize that entities offer conceptual foundations for addressing a whole host of scientific questions concerning meaningfulness and theory, many far more important than mere measurability criteria. For example, any group of items showing reasonable Rasch model fit can be claimed to exhibit not only empirical measurement properties, but also material status as a mathematical object. That is the point of calibrating items and measuring persons, because they provide empirical foundations for something "real" such as an ability or a psychological trait.

An item-person map represents an ability or psychological entity that was inferred by a mathematical abstraction and, when data fit, it reveals a reproducible aspect of experience. In this context, an item parameter value is not a transient sample artifact, but a quantitative object with exact and tangible material properties that are reaffirmed whenever item responses are analyzed. The material significance of this conceptual entity, now specifically a "psychometric entity", increases as framework invariance is verified and extended.

Physical theory provides many examples that reinforce the importance of conceptual entities for constructing scientific knowledge. In fact, the importance of conceptual entities in scientific theory is difficult to overstate (Maxwell, 1999). Electrons, gravity, and planetary orbits are material entities that are central to understanding a wide range of physical observations. (Prominent historical failures are phlogiston, aether, and humors -- among many others.) While originally the conceptual entities had no reality or ontological status, but were only conjectures within physical theory, scientists have expressed them in mathematical abstractions and linked them to empirical "reality".

Using linear measuring instruments, their empirical implications have provided foundations for an enormously successful body of scientific knowledge, and scientific advances have led to mathematical consolidations such as the reduction of chemistry to physics. (Consider Newton's inverse square law which governs movement of both planets and electrons.) This achievement is remarkable because scientists have never "seen" an electron, gravity, or planetary orbits but only highly predictable empirical effects.

In contrast, the ontological status of psychometric entities in education and psychology are scientifically eccentric because raw score rank order reliability and correlation are virtually their only claim to material status which, not surprisingly, has severely inhibited their maturation as scientific disciplines. Raw score structures provide only fleeting glimpses of a reality that is dependent on particular item sets and samples. A logical consequence is fragmented and discontinuous constructs that are virtually impossible to consolidate or integrate into an overall body

Rasch Measurement Transactions
P.O. Box 811322, Chicago IL 60681-1322
Tel. & FAX (312) 264-2352
rmt@rasch.org www.rasch.org/rmt/
Editor: John Michael Linacre
Copyright © 2003 Rasch Measurement SIG
Permission to copy is granted.
SIG Chair: Trevor Bond SIG Secretary: Ed Wolfe

of scientific propositions.

The mathematical object offered by Rasch measurement preserves the rank-order achievements of 20th century psychometrics but offers much more. First it reveals meaningfulness that is inherent in a transitive numerical item structure (probabilistic additivity insures axiomatic transitivity). Second, it presents a common quantitative framework which facilitates consolidation of disparate data sets. The ontological significance of this framework gains scientific importance as measures with an explicit metric are related to other variables in mathematical functions – not just rank-order correlations. When mathematical functions are subsumed under comprehensive explanatory theories, they demonstrate science's explicit intention to reduce and unify knowledge.

While ontology and entities may seem obscure and irrelevant to researchers and practitioners, in fact, they determine the "kind" of science that we practice. Psychometric entities with additive, linear measurement properties profoundly improve what we "know" from test scores. They open a window on scientific knowledge which is certainly more important than preoccupations about measurability. In the contemporary psychometric climate, a coherent understanding of why we should do more than simply compute test score reliability is an unusual opportunity to advance social science.

Nikolaus Bezruczko

Cohen, B. I. (1985) *Revolution in Science.* Cambridge, Massachusetts.: Harvard University Press.

Feyerabend, P. (1978). *Against method.* London: Verso.

Graves, R. E. (2003). In pursuit of Rasch measurement: Explorations following Michell. *RMT, 17*, 1, 914-915.

Maxwell, G. (1999). Theoretical entities. In Robert Klee (Ed.) *Scientific Inquiry: Readings in the Philosophy of Science*. Oxford: Oxford University Press.

Michell, J. (2002). *Conjoint measurement and the Rasch model: Quantitative versus ordinal structure.* Paper at IOMW, New Orleans, 2002.

Simon, H. (1989) Remark in W. Sichel (Ed.) *The State of Economic Science.* Upjohn.

COMET and IOM Chicago Chapter

at Rehabilitation Institute of Chicago

3:30 PM, Thursday, September 25, 2003
Matthew Schulz: Using domain scores to describe mathematics achievement in National Assessment of Educational Progress

3:30 PM, Thursday, October 23
Theresa Pape: Measurement, Treatment Effectiveness and Outcomes Post Severe Brain Injury

Thursday, November 20, 2003
Ken Conrad: Equating Items from Instruments with Identical Constructs

Thursday, December 18, 2003
Patrick Fisher: When Should I Use FACETS? Always!

***C**hicago **O**bjective **ME**asurement **T**able*
Nikolaus Bezruczko and Patrick Fisher, coordinators.
pfisher@measurementresearch.com

SAS Macros for Rasch

Danish statistician, Karl Bang Christensen, and I have developed some SAS macros for Rasch modeling. This was part of Karl's Ph.D. study and he has laid down the statistical foundation with some help from me on programming. The macros can handle:

- the standard Rasch model and
- the Partial Credit model for polytomous items
- the Martin-Löf test for multidimensionality
- log-linear Rasch models (that can model differential item functioning)
- latent regression models (where the outcome is measured by a Rasch model or a log-linear Rasch model)
- Poisson or logistic regression models (where a covariate is measured by Rasch model or a log-linear Rasch model)

The macros are discussed in a technical report in which you can also find the link to the macros themselves:
Christensen, K. B., & Bjorner, J. B. (2003). SAS macros for Rasch based latent variable modeling (Tech. Rep. No. 03/13). Department of Biostatistics, University of Copenhagen. Available from
www.biostat.ku.dk/publ-e.htm

We would be very happy if someone would use the macros in their analytic work. We also hope that the macros can illustrate how these models can be fitted with a standard statistical package. However, for long scales the macros will be slow and you are probably better off using another program. Note also, that you would use the macros at your own risk and that we cannot commit ourselves to providing any support. For comments on the macros, please write either Karl kbc@ami.dk or myself.

Jakob Bue Bjorner, MD, PhD
Deputy Chief Science Officer
QualityMetric Incorporated
jbjorner@qualitymetric.com
www.qualitymetric.com
www.sf-36.org

Unobserved Categories: Estimating and Anchoring

Estimating measures from data containing rating scale, partial credit, polytomous and other ordinal structures, with categories that are not observed, is awkward.

Structural and incidental zeroes

Unobserved categories can be of two types: structural zeroes and incidental or sampling zeroes. Structural zero means that the category is unobserved because it is defined not to exist. For instance, a rating scale may be defined to consist of the 4 categories 10, 20, 30, 40. Categories 11, 12, 13, ... cannot be observed. They are merely an artifact of the category numbering system. For analysis purposes, 10, 20, 30, 40 represent categories 1, 2, 3, 4. Consequently either by direct recoding, or automatically within the software, this numerical transformation must be performed. If it is necessary to report adjusted raw scores in the original numbering system, a reverse transformation is required.

Incidental zeroes correspond to categories that are defined to exist and are expected to be observed with some samples, but are not observed with this particular sample. They correspond to performance levels that are within the range of this sample, but do not happen to have been manifested. This is likely to occur with long rating scales (such as percentages) used with small samples, but may happen, by chance, with any ordinal scale with any sample. In this case, unobserved categories correspond to actual performance levels. They must be maintained in order for the ordinal structure to keep its integrity, despite their lack of observations.

Structural and incidental zeroes can occur simultaneously. If a 4 category rating scale has categories numbered 10, 20, 30, 40, perhaps only categories 10, 30, 40 are observed this time. Then to eliminate structural zeroes, the scale is renumbered, for estimation purposes, 1, 2, 3, 4. In the original data, 20 was not observed, so, for the purposes of estimation, the rating scale becomes, 1, 3, 4 with 2 an incidental zero.

Remedying structural zeroes

Rasch polytomous analysis (rating scale, partial credit, Poisson, etc.) proceeds on the basis that each advance of one qualitative level up the polytomy is represented by a one-point ordinal advance. This requires that structural zeroes be eliminated, and the qualitative levels renumbered cardinally. Renumbering may be performed automatically by software or may require explicit data recoding, particularly if the software demands that the cardinal numbers start at 0.

Remedying incidental zeroes: approximate, but effective

Incidental zeroes correspond to categories that could be observed, but aren't. A consequence will be that comparison with other analyses, in which all or different categories are observed, will be difficult. **The simplest remedy is to include some dummy data records which include those unobserved categories.** For instance, suppose that one category of a 7 category rating scale is not observed in this data set. Construct and include a reasonable data record which contains the missing category. If it is the lowest category that is missing, the data record would have next-but-lowest categories for all items except the hardest, which would have the lowest, otherwise unobserved, category. For an intermediate category, a mid-difficulty item would have the unobserved category. Easier items the next higher category. Harder items the next lower category. In general, a few dummy data records would have minimal direct impact on fit statistics or summary statistics. If necessary, the dummy data records can be used to produce rating scale anchor values, then the rating scales can be anchored and the dummy records dropped for final reporting.

Incidental extreme zeroes: no exact remedy

Consider unobserved extreme (high or low) categories. These correspond to performance levels outside that of the current sample. This sample provides no information to estimate their probability of occurrence. Accordingly unobserved extreme high and low categories are ignored for exact estimation based only on this sample. Thus a rating scale may be defined with categories 1, 2, 3, 4, 5, 6 but if categories 1, 2, and 6 are not observed in this dataset, then analysis proceeds as though the rating scale is defined to be 3, 4, 5. The advancing integers correspond to advancing performance levels. The choice of initial integer, 3 in this case, makes no difference to the estimation, but may be constrained by software considerations to 0. In which case, 3, 4, 5 must be renumbered 0, 1, 2.

Incidental intermediate zeroes: an exact remedy

Suppose that intermediate category z is not observed in this dataset. The items are grouped by rating scale structure using subscript *g*. If this is the Andrich rating scale model, then one instance of *g* includes all items. If this is the Masters' Partial Credit model, then there is one instance of *g* for each item. If different groups of items share different scale structures, then there is an instance of *g* for each group. For each group, the observable categories are numbered cardinally from 0 to m_g, the highest observed category for the group. Then

$$P_{ni(x \neq z)} = \frac{\exp \sum_{j=0}^{x} (B_n - D_{gi} - F_{gj})}{\sum_{k=0, \neq z}^{m_g} \exp \sum_{j=0}^{k} (B_n - D_{gi} - F_{gj})}$$

where $P_{ni(x \neq z)}$ is the probability that person *n* on item *i* is observed in category *x,* which cannot be *z,* the unobserved

category, for which $P_{niz} = 0$. B_n is the ability of person *n*. D_{gi} is the difficulty of item *i,* a member of group *g*. F_{gj} is the calibration of category *j* relative to category *j-1* in the ordinal scale for group *g*. F_{g0} =0 and F_{gz}=0 for computational purposes. $\Sigma F_{gj} = 0$ summed for all m_g+1 categories in group *g*.

Incidental zeroes:
an almost exact remedy, useful for anchoring

Dropping extreme unobserved categories and flagging intermediate unobserved categories is awkward, perhaps impossible if not supported by software. It is also not transportable, in terms of anchor values, to other analyses in which those categories are observed. Accordingly, the rating scale structure from the "exact remedy" can be modified to include the unobserved categories directly. Unobserved categories have not been observed and so the inference is that they must have a very low probability of being observed.

For intermediate categories, this corresponds to a very high value of F_{gj} for the unobserved category, and a very low value of $F_{g(j+1)}$ for the next category. In practice, "very high" means "add 40 logits", very low means "subtract 40 logits". Applied to the 4, 2, 0, 4 example above, the parameter estimates become: $F_1 = \log(2)$, $F_2 = 40$, $F_3 = -40-\log(2)$, with no category flagged or dropped.

Multiple incidentally unobserved categories can be anchored using the same approach of adding 40 at the low end and subtracting 40 at the high end. If the frequencies of categories 0-6 are 12, 2, 0, 0, 0, 16, 5, then $F_1 = 1.6$, $F_2 = 40$, $F_3 = 0$, $F_4 = 0$, $F_5 = -42.6$, $F_6 = 1.0$. The basis for inference becomes weaker the more unobserved categories there are.

For an extreme unobserved bottom category, consider category frequencies 0, 1, 2, 1. Then $F_1 = -40$, $F_2 = 20-\log(2)$, $F_3 = 20+\log(2)$. The first observed category has a very low parameter estimate, but the relationship between other estimates is unchanged, and their overall sum remains zero.

For an extreme unobserved top category, consider category frequencies 1, 2, 1, 0. Then $F_1 = -20-\log(2)$, $F_2 = -20+\log(2)$, $F_3 = 40$. The unobserved top category has a very high parameter estimate, but the relationship between other estimates is unchanged, and their overall sum remains zero.

Incidental zeroes: a curve-fitting approach

The Guttman-component technique (D. Andrich & G. Luo, 2003, Conditional pairwise..., Journal of Applied Measurement 4:3, 205-221) is one technique that can bridge over unobserved categories by modeling all categories to be part of a smooth process. This is particularly powerful for long rating scales, such as percentages, with many incidental-zero categories .

John Michael Linacre

Summing Multiple Measures

Question: I have a multi-stage testing procedure using items drawn from a pre-calibrated item bank. Each stage gives each person a measure and standard error. Can I combine these measures to give each person an overall measure?

Answer: The most exact answer would be to use all of a person's responses to all stages to estimate a new measure – but this may not be practical. Typically the first-stage is a short, wide subtest. The later stages may be narrower, longer subtests. So raw score performance and measurement precision can vary greatly between the subtests. RMT 8:3, p. 376, suggests that a useful combined measure is obtained by standard error weighting the subtest measures:

$$\text{Test Measure} = \frac{\sum_{\text{subtests}} \frac{\text{Measure}_{\text{subtest}}}{SE_{\text{subtest}}}}{\sum_{\text{subtests}} \frac{1}{SE_{\text{subtest}}}}$$

If it is reasonable to think of all subtests as measuring this same fixed effect, then its standard error approximates

$$1 / \sqrt{\Sigma (1 / SE_{\text{subtest}}^2)}.$$

IRT-Lab Version 2

Using software to enhance research and pedagogy in the fields of Rasch measurement and IRT.

IRT-Lab performs a variety of functions that can be used by researchers, teachers, and students of Rasch measurement and item response theory (IRT) to facilitate the understanding of these nonlinear measurement models.

- Graph item characteristic curves associated with the dichotomous Rasch and IRT models, the partial credit model, the generalized partial credit model, and the graded response model.

- View the log-likelihood functions associated with person and item parameter estimation

- View the information functions and corresponding standard error functions associated with person parameter estimation.

- Simulate data for tests having items following a variety of Rasch and IRT models.

A copy of IRT-Lab can be obtained by providing a full mailing address to Randy Penfield at

penfield@coe.ufl.edu

What is Item Response Theory, IRT? A Tentative Taxonomy

Debate about IRT sometimes verges on the nonsensical, and certainly on the irascible, because protagonists are using the term in very different senses. Scanning the psychometric literature reveals at least three tentative definitions:

(1) **IRT encompasses any model** "relating the probability of an examinee's response to a test item to an underlying ability." (HMIRT, p. v). This definition is so broad that it includes everything from Classical Test Theory (CTT) to non-parametric Mokken scaling.

(2) **IRT** encompasses any mathematical model which attempts to **predict observations from locations on a latent variable.** This is also called "Latent Trait Theory". It includes logistic models of all types, normal ogive models, log-log models, etc.

(3) **IRT** centers on the particular **models advocated by Frederic M. Lord**, particularly 2-PL and 3-PL, but also 1-PL, Normal Ogive, and recently by extension, Generalized Partial Credit models.

What is disconcerting is that deconstruction is needed in order to determine the IRT definition intended by an author and precisely what models lie within that definition.

A Taxonomic Adventure

Consider the typical, apparently straightforward, statement that

> **"IRT item parameters are not dependent on the sample used to generate the parameters, and are assumed to be invariant (within a linear transformation)** across divergent groups within a research population and across populations" (Reeve, 2002).

This is not true of CTT, because item p-values are highly sample dependent, and linear transformations are meaningless. So this statement does not apply, in general, to (1) above.

Consider (3). Is this true of conventional 1-PL? Not exactly. Under 1-PL, the person sample mean is set at 0. Consequently the item parameter estimates change depending on the person sample ability distribution. But perhaps this is what is intended by **"linear transformation"**. So let us grant this.

Note: 1-PL is an approximation to the normal ogive model, expressed in logit terms with fixed discrimination and no guessing. Usually the person mean is set at 0, i.e., it is norm -referenced. The Rasch dichotomous model is a derivation from measurement axioms. It has nothing to do with the normal ogive model. The item mean is set at 0, i.e., it is criterion-referenced. By an accidental coincidence, 1-PL and the Rasch dichotomous model are, in principle, algebraically equivalent.

Is the statement true of 3-PL? We encounter another nicety. What does **"IRT item parameters"** mean? Does it mean their true values as expressed in the mathematical model, or their estimated values from data? If it means their true values, then the statement is a tautology because the true values are, by definition, not dependent on any sample. So does it mean "parameter estimates"? This takes us into another level of complexity.

"... not dependent on the sample used ..." implies that any reasonable sample of the same kind of subjects with any mean, variance, skewness, kurtosis, modality, discreteness, etc. yields statistically equivalent parameter estimates. The description that Lord (1980, p. 180) gives

Journal of Applied Measurement Volume 4, Number 3. Fall 2003

Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories using Principal Components. *David Andrich and Guanzhong Luo, p. 205-221*

Reliability and True-score Measures of Binary Items as a Function of Their Rasch Difficulty Parameter. *Dimiter M. Dimitrov, p. 222-233.*

Using Logistic Regression to Detect Item-level Non-response Bias in Surveys. *Edward W. Wolfe, p. 234-248.*

Rasch Measurement in the Assessment of Amytropic Lateral Sclerosis Patients. *Josephine M. Norquist, Ray Fitzpatrick, and Crispin Jenkinson, p. 249-257.*

Measuring Client Satisfaction with Public Education II: Comparing Schools with State Benchmarks. *Trevor G. Bond and John A. King, p. 258-268.*

The Recovery of the Density Scale using a Stochastic Quasi-Realization of Additive Conjoint Measurement. *Timothy W. Pelton and C. Victor Bunderson, p. 269-281.*

Understanding Rasch Measurement: Substantive Scale Construction. *Mark H. Stone, p. 282-297.*

This issue is the third issue in an expanded format. There is a larger page size and a double column format to make it easier to read, as well as seven articles per issue. Sample copies are available from the Editor. *Recommend JAM to your librarian!*

Richard M. Smith, Editor
Journal of Applied Measurement
P.O. Box 1283, Maple Grove, MN 55311
JAM web site: ***www.jampress.org***

of his estimation procedure implies this is true. But that procedure cannot work. It diverges. Constraints must be placed on the sample distribution and on other parameter values. These constraints compress and expand the latent variable so that it loses its intended linear form and becomes a local description. But perhaps this is what the statement means by **"assumed"**.

If analysts and decision-makers are prepared to assume that the constraints on sample distribution, etc., will always match their empirical data, then they are justified in assuming that their parameter estimates will not depend on the sample used. But their assumptions will always be insecure. Once the assumption, or rather assertion, that the sample has any particular distribution is imposed on the estimation process, that process will yield a sample distribution that matches the assumption. The estimation process becomes a self-fulfilling prophecy. Surely this is not what the statement intends.

We now see that a statement intended to characterize all IRT models, in fact characterizes only a limited set, and not even all the ones in (3) that Fred Lord advocated.

Wright's Bifurcation

Consider (2) above. Ben Wright (e.g., 1984) divides it into two sub-classes according to the axiomatic basis of the psychometric models.

In one sub-class are what Wright labels **"IRT models."** These accord with Lord (1980, p. 14), where he writes, "The reader may ask for *a priori* justification of [3-PL]. No convincing *a priori* justification exists .... The model must be justified on the basis of the results obtained, not on *a priori* grounds." Here is Martha Stocking's summary of Lord's statistical methodology: "Building statistical models is just like this. You take a real situation with real data, messy as this is, and build a model that works to explain the behavior of real data." *(New York Times, 2-10-2000).* In other words, if the model doesn't fit a particular data set, change the model!

In the other sub-class are what Wright labels **"fundamental measurement models,"** based on measurement axioms. These include Rasch models. Such models embody mathematical ideals (analogous to parallel lines and Pythagorean triangles) that can never be realized empirically. They can only be approximated. But a good approximation is all that is required for utility. Accordingly, if the data don't approximate the desired model, the data are not immediately useful for measurement, and so must be changed or replaced.

The Reeve (2002) statement applies most exactly to this "fundamental measurement" sub-class of definition (2), and not, in general, to (1), (3) or the sub-class of (2) that Wright labels "IRT".

Those new to IRT have good reason to be confused.

Now it's your turn ....

Here is another statement:

> "Rasch scaling transforms the ordinal items to **the logit scale and, thus, to interval-level measurement.** It should be noted that **this metric is characteristic of all IRT models**, not just the Rasch model" (Cook et al., 2003).

Please deconstruct this statement. How does it relate to the three tentative definitions of IRT? Is it robust against Wright's bifurcation?

John Michael Linacre

Cook K.C., Monahan P.O., McHorney C.A. (2003) Delicate balance between theory and practice: Health status assessment and Item Response Theory. Medical Care 41:5, 571-4.

HMIRT: Van der Linden, W. J., Hambleton, R. K. (eds.) (1997). Handbook of Modern Item Response Theory. New York: Springer Verlag.

Lord F.M. (1980) Applications of Item Response Theory to Practical Testing Problems. Hillsdale NJ: Erlbaum.

Reeve, B.B. (2002) An Introduction to Modern Measurement Theory. National Cancer Inst.

Wright B.D. (1984) Despair and hope for educational measurement. Contemporary Education Review, 3(1), 281-288.

Metrology, Law and Providence

"One of the central ideas of ancient metrology is explained by the much read and little understood *Fifth Book of Aristotle's Ethics,* in which the idea of justice is explained by referring to money and to the price structure. This book explains why money [in Greek] is called by the same name that applies to civil law and to natural law *(nomos)* and why this term is synonymous with *arithmos* [(number)]. **Metrology first developed as an attempt to assure justice in the contract of sale by mathematizing the relation."**

"The origins of the art of legislation and of legal science are to be found in the lists that state how many measures of a given commodity would correspond to a measure of another commodity. Once one takes this practical outlook, one can see how the idea of Divine Providence is linked with the methods used in the rationing of food, of which Greek inscriptions provide the most abundant evidence. Once one keeps in mind the metrological aspects of the idea of Providence, one can see the meaning of the word *epiousios* [daily] in the Lord's Prayer, a word on the interpretation of which an entire library has been written. One must keep in mind the ethical aspects of metrology to see in the Gospels the metrological reasons for the two miracles of the multiplication of the bread, the Feeding of the Four Thousand and the Feeding of the Five Thousand. **In metrology, one must steadily shift from metaphysical and ethical presuppositions to practical aspects."**

from "A History of Measures" by Livio C. Stecchini (ca. 1960).

The Hierarchical Rater Model from a Rasch Perspective

Modeling rater behavior is problematic. Are the rater's intended to be acting as **locally-independent experts**, each with a unique perspective of what is the "true' rating? If so, each rater provides new information about the person being rated. The raters have the same status as test items, and a many-facet Rasch model is indicated. In general, more ratings by more raters of the same person-item interaction produce more measurement information.

Are the raters merely **human scoring machines**, all expected to produce the same one, "true" rating? If so, then the same type of quality control that would be applied to optical scanning equipment is indicated. More ratings by more raters of the same person-item interaction produce no additional measurement information, nor more information about the "true" rating.

In practice, however, the situation is ambivalent. Raters are told to use their expertise, but are also **instructed to conform** with other raters in awarding "true" ratings. More ratings by more raters of the same person-item interaction produce more information about the "true" rating, but not otherwise more measurement information about the performance.

So how is this asymmetry in the rating process to be modeled? **The Hierarchical Rater Model** (HRM) is one approach.

HRM (Patz et al., 2000, a variant is Donoghue & Hombo, 2003) uses a two-level approach. At the first level is modeling person performance. HRM uses a Rasch Partial

$$\log\left(\frac{P_{nij}}{P_{ni(j-1)}}\right) = B_n - D_i - F_{ij}$$

Credit Model with persons and items, but the estimates are based on idealized "true" (not empirical) ratings.
where "j" represents "true", not empirical, ratings.

At the second level are the idealized "true" ratings. HRM models each rater's empirical ratings to follow a normal distribution on a "raw rating" variable. Somewhere on this treated-as-linear variable is the "ideal" or "true category" rating, i.e., the rating that would have been awarded by a perfect rating machine to a particular person on a particular item.

Each empirical rating, however, is displaced from its corresponding ideal by
(a) its rater r's leniency, μ_r, expressed as a fractional-raw-score rating adjustment, and
(b) its rater r's unreliability, expressed as the fractional-raw-score standard deviation, σ_r, of a normal distribution around the rater's severity.

$$P_{j|nir} \propto e^{\frac{-1}{2\sigma_r^2}[k_r - (j+\mu_r)]^2}$$

where j is the ideal "true" rating of person *n* on item *i* and *k* is the empirical rating observed for rater r.

Donoghue & Hombo differ from Patz et al. in using the generalized partial credit model (i.e., the Rasch partial credit model with an item discrimination parameter) and a "fixed effect" rating model (not completely specified in their paper).

From a Rasch perspective, using the "partial credit" model is impeccable. The "ideal" rating model, however, is deficient. The "raw rating" variable is definitely not interval, it is ordinal, and may only be dichotomous. For a very lenient rater on a long rating scale, the most probable rating, according to HRM, could be a category above the top of the scale. This is impossible, so an adjustment must be made. Most obviously, the probability of awarding categories above the top category should be added to the probability of the top category. But this does not appear to have been done. Instead, out-of-range categories are merely ignored. The effect of this is that lenient raters are estimated to be even more lenient, and vice-versa for severe raters.

This suggests that an immediate improvement to the HRM model would be to express the "idealized rating" model in logistic terms, e.g., most simply,

log ((x-"bottom")/("top"-x)),

where "bottom" and "top" are the extreme categories. The probability of observing any particular category then becomes the integral of the probabilities of the rating occurring within 0.5 rating-points of that category on the logistic rating variable. A further improvement (perhaps already made by Patz or Donoghue) would be to bring into the "partial credit" model not merely the "idealized rating" for each person-item confluence, but the set of all possible ratings, and the probability that each one is ideal.

This area of research is at an early stage. Here is an opportunity for a Rasch-oriented doctoral student to formulate a truly measurement-based HRM model.

John Michael Linacre

Patz R.J., Junker B.W., Johnson M.S. (2000) The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. Revised AERA Paper.

Donoghue J.R., Hombo C.M. (2003) An Extension of the Hierarchical Raters' Model to Polytomous Items. NCME Paper.