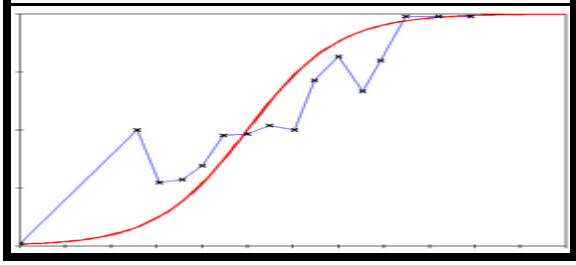
Model & Empirical ICCs



RASCH MEASUREMENT

Transactions of the Rasch Measurement SIG
American Educational Research Association

Vol. 20 No. 2 Autumn 2006 ISSN 1051-0796

Item Discrimination and Rasch-Andrich Thresholds Revisited

Linacre (2006) concludes that it might be an advantage to have thresholds reversed relative to their natural order because the expected value curve is then steeper and more discriminating than if the items are in their natural order. He uses the case of four discrete dichotomous items being summed to form a polytomous item with a maximum score of 4 to illustrate his point. This note demonstrates that the apparent improved discrimination in such a case is artificial. It is directly analogous to over discrimination in the case of dichotomous items. This point is explained below.

The Rasch model for more than two ordered categories can be expressed as:

$$\Pr\{X_{ni} = x\} = \exp(x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_{ik}) / \gamma_{ni} \quad (1)$$

where $x \in \{0,1,2,..m_i\}$,

$\gamma_{ni} = \sum_{x=0}^{m_i} \exp(x(\beta_n - \delta_i) - \sum_{k=0}^{x} \tau_{ik})$ is a normalizing factor, β_n is the location of person *n*, δ_i is the summary location of item *i*, and the $\tau_{ik}, k = 1,2,...m_i$ are thresholds with $\sum_{k=0}^{m_i} \tau_{ik} = 0$. The special case of a response to a dichotomous item, $x \in \{0,1\}$, is

$$\Pr\{X_{ni} = x\} = \exp(x(\beta_n - \delta_i)) / \gamma_{ni}. \quad (2)$$

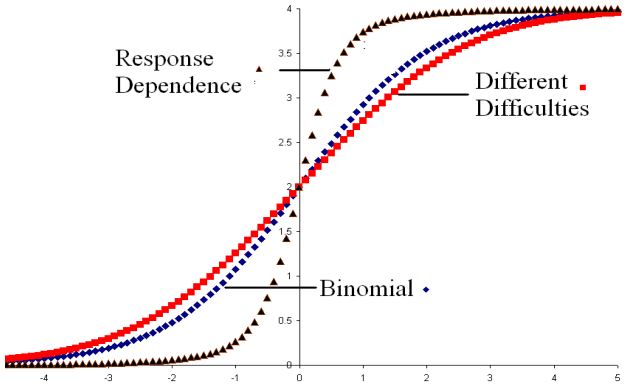


Figure 1: ICCs for three subtests

Suppose we sum the responses of four discrete and statistically *independent* dichotomous items of *equal* difficulty and these are analyzed according to Eq. (1) in which $m_i = 4$. Let us call the item so formed a *subtest*.

First, because the items have equal difficulty and are statistically independent, the total score of the subtest results in a binomial distribution. In that case the model of Eq. (1) specializes to

$$\Pr\{X_{ni} = x\} = \exp(x(\beta_n - \delta_i) + \ln\binom{m_i}{x}) / \gamma_{ni} \quad (3)$$

and the thresholds take on very specific values. In the case of four items, these values are -1.39, -0.41, 0.41 and 1.39 (Andrich, 1985). Notice that they are symmetrical, as would be expected, and properly ordered. These values have nothing to do with the distribution of persons, or the relative difficulties of the items – they characterize the response structure among scores, given the location of the person and the item.

The two requirements of a binomial distribution, equal difficulty of items and statistical independence, may be violated in two simple ways. First, the items may be of different difficulty, and second, the responses may not be independent. If the items are of different difficulty but independent, then the distribution of scores in the subtest for any β_n and δ_i will regress to the central scores and there will be a greater proportion of frequencies in the middle scores than in the binomial. This difference in difficulty will not violate the general Rasch model, however.

Second, if the items are of the same difficulty but there is response dependence among the items, then a score of 1 in one item will give a greater probability of 1 for a

Table of Contents

CAT standard errors and stopping rules 1052
Commercial measurement and research 1058
Gas law and a Rasch reading law 1059
Item discrimination and thresholds Revisited.....1055
Measurement of psychological value 1060
SAS Rasch simulation 1061

dependent item than if there were no dependence, and correspondingly a score of 0 in one item will give a greater probability of 0 in the dependent item than if there were no dependence. Thus if the items are of the same difficulty but the responses are dependent, then the distribution of scores in the subtest for any β_n and δ_i will diverge to the extremes scores and there will be a greater proportion of frequencies in the extreme scores than in the binomial. This is opposite to the effect of items having items of different difficulty in the presence of statistical independence.

Table 1 shows the probabilities of each score of a subtest composed of four discrete items under three conditions for the case where $\beta_n = \delta_i = 0$. In Case 1 the items are of equal difficulty, and statistically independent and give a binomial distribution, in Case 2 items have different relative difficulties, $-1.5, -0.5, 0.5, 1.5$, but statistical independence holds, and in Case 3 the items are of equal difficulty 0, but all items show statistical dependence on the response of the first item. The dependence is constructed in the following way: for a value of $d = 2$, if the response to item 1 is $X_{n1} = 1$, then the probability of a response $X_{ni} = 1, i = 2,3,4$ is

$$\Pr\{X_{ni} = 1 \mid X_{n1} = 1\} = \exp(x(\beta_n + d - \delta_i)/\gamma_{ni},$$

i=2,3,4; and if the response on item 1 is $X_{n1} = 0$, then the probability of a response $X_{ni} = 1, i = 2,3,4$ for subsequent items is

$$\Pr\{X_{ni} = 1 \mid X_{n1} = 0\} = \exp(x(\beta_n + d - \delta_i)/\gamma_{ni},$$

i=2,3,4. That is, if the response to the first item is 1, the probability of a response of 1 for subsequent items is greater than if there was no dependence, and if the response to item 1 is 0, then the probability of a response of 1 is less than if there was no dependence. This may occur in a set of four items all of which belong to one reading stem, and the correct response to the first item gives clues to the correct responses to the other items.

Table 1
Distributions of total scores for the binomial, for items of different difficulties, and for items with responses dependent on the first item

Score X	Binomial Pr{X}	Different difficulties, independence Pr{X}	Equal difficulties, dependence Pr{X}
0	0.0625	0.03505	0.34166
1	0.2500	0.24395	0.13956
2	0.3750	0.44199	0.03756
3	0.2500	0.24395	0.13956
4	0.0625	0.03505	0.34166
Sum	1	1	1

It is clear from Table 1 that when the items are of different difficulty, that the probability of the middle score of 2 in the subtest is greater than in the binomial distribution, and when items 2, 3 and 4 have response dependence on item 1, the probability of extreme score in the subtest is greater than in the binomial. An extreme case of dependence would be where the three items were totally dependent on the first, in which case all responses would be the same, and all scores would be 0 or 4. Table 2 shows the threshold values which correspond to the frequencies shown in Table 1.

Table 2
Thresholds for the binomial distribution and one with items of different difficulty and one with items with response dependence on the first item

Threshold k	Binomial	Different difficulties, independence	Equal difficulties, dependence
1	-1.39	-1.94	0.90
2	-0.41	-0.59	1.31
3	0.41	0.59	-1.31
4	1.39	1.94	-0.90

In particular, in the subtest where the items are of different difficulty and the responses are independent, the thresholds are further apart than in the binomial distribution; in the subtest where the items are of the same difficulty but the responses of items 2, 3 and 4 are dependent on the response to item 1, then the thresholds are closer together than the binomial, and even reversed relative to the natural order.

Figure 1 shows the Item Characteristic Curves (ICCs) for these three subtests. It is evident that the slope of the ICC, that is the discrimination, for the subtest with response dependence is the greatest, and that for the subtest with items of different difficulty it is the least. The subtest with a binomial response structure and which provides the frame of reference for the analysis and interpretation., has a slope between the other two subtests.

However, it should be evident that these curves cannot be interpreted simply in terms of the relative discrimination of the items, with the conclusion that the greater the discrimination the better. In particular, the ICC of the subtest with dependence, the information is not equivalent to that from four independent items. Instead the information available from these four items is less than would be obtained from the same number of i statistically independent items. The model for ordered categories follows the data in obtaining the threshold estimates, and accounts for the dependence among the items, but the information available is less than if the items were statistically independent.

Indeed the effect shown here is well known in traditional test theory (TTT). In TTT, the reliability index is a key statistic to evaluate a test. And the greater the correlation among items, the greater the reliability. However, TTT

adherents observed that the reliability is the highest when all items are identical, and understood that they then effectively have only one item. In that case the validity of the test was that of just one item. They called this reduction in validity with such an increase in reliability as the *attenuation paradox*.

The Rasch model, and the above analysis provides an understanding of this paradox, and the binomial distribution provides a criterion to assess if there is dependence among subsets of items. Because differences in item difficulty and dependence in responses have opposite effects on the distribution of total scores of a subtest of each person, the binomial thresholds can be used as a conservative criterion to identify dependence. Specifically, because items will generally have some difference in difficulty, if the thresholds of a subtest are closer together than those of the binomial (or reversed), then it follows that there must be response dependence. If the items have different difficulty, then there must be even more response dependence than if the items are of the same difficulty.

This analysis and specific values for different maximum scores of a subtest that provide a conservative criterion for evidence of dependence is provided in detail in Andrich (1985). The application of the Rasch model in understanding the attenuation paradox of TTT is also described in Andrich (1988). In summary, the dichotomous Rasch model fixes the test discrimination as a kind of average of the discrimination of all items, and under discrimination of any items relative to this average suggests multidimensionality, while over discrimination relative to this average suggests response dependence. This is one basic difference between the perspectives of TTT and a Rasch model analysis. In TTT, the greater the discrimination of an item the better, though there is an awareness that items may over discriminate and not add to the validity of the test; in a Rasch model analysis there is an explicit criterion of over discrimination relative to the test as a whole as an indicator of possible response dependence. The above example in which a subtest is composed of discrete items, shows that the idea of *over* discrimination in dichotomous items can also manifest itself with a polytomously scored item. In this case the over discrimination in some sense accounts for dependence among the discrete items, but it does not add to the information, instead, in accounting for the dependence it shows that there is less information in the subtest than if the items were independent.

Thus there is a lesson to be learned from the above analysis. The information is in the data, and information cannot be contrived which might show some better index without understanding how the model interacts with the data and how it manifests properties of the data, including properties that are dysfunctional in the data. In our case of four discrete items above, the very high discrimination reflects statistical dependence and less information than if the items were statistically independent. That is, there is no greater information in the data of the subtest just because it has been rearranged and analyzed in a different way resulting in a manifest high discrimination in the subtest. In the extreme case we mentioned above, if the responses to the items were identical reflecting total dependence, all responses would be 0 or 4, and the thresholds would be so reversed that the ICC for the subtest would be vertical at the subtest difficulty. This would give the impression that at this point there is infinite information! Clearly, this is not case - the information is just that of a single item.

David Andrich, Murdoch University

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. Chap. 9 in S. Embretson (Ed.), *Test design: Contributions from psychology, education and psychometrics.* Academic Press, New York

Andrich, D. (1988). Rasch models for measurement. Sage Publications.

Linacre, J.M. (2006) Item Discrimination and Rasch-Andrich thresholds. Rasch Measurement Transactions, 20, 1, 1054.

Rasch-related Coming Events

Dec 2006 - Dec 2008 3-day Rasch courses, Leeds, UK home.btconnect.com/Psylab_at_Leeds

Dec. 8, 2006, Fri. Midwest Objective Measurement Seminar - MOMS, Chicago Mary Lunz www.measurementresearch.com

Dec. 11-13, 2006, Mon.-Wed. Objective Measurement in the Social Sciences - ACSPRI Conference, Australia www.acspri.org

Jan. 15-19, 2007, Mon.-Fri. Measurement in the Psychosocial Sciences: from raw scores to Rasch measures (Andrew Stephanou), Australia www.acspri.org

Feb. 12-16, 2007, Mon.-Fri. Item Response Modeling With ConQuest (Ray Adams & Margaret Wu), Australia www.edfac.unimelb.edu.au

Feb. 16 - Mar. 16, 2007, Fri.-Fri. Practical Rasch Measurement (Winsteps) online course (Mike Linacre) www.statistics.com

March 26-27, 2007, Mon.-Tues. Introduction to IRT/Rasch Measurement Using Winsteps (Conrad & Bezruczko), Chicago www.winsteps.com

Apr. 7-8, 2007, Sat.-Sun. Introduction to Rasch Measurement: Theory and Applications, Chicago IL (Smith & Smith) www.jampress.org

Apr. 9-13, 2007, Mon.-Fri. AERA Annual Meeting, Chicago www.aera.net

May 4 - June 1, 2007, Fri.-Fri. Facets online course (Mike Linacre) www.statistics.com

Commercial Measurement and Academic Research

Are precision instrumentation and technology the products, by-products, or spin-offs of scientific research that is conducted in, by, or through academic institutions? No! "Historically the arrow of causality is largely from the technology to the science" (Price, 1986, p. 240).

Scientific discoveries and theories do not lead to technological innovation. Rather, technological innovations spring up from within the framework of existing engineering problems in industrial and commercial contexts. Wallace (1972, p. 239) in his classic study of the Industrial Revolution discovers that economic pressures drive the development of new technologies more often than new scientific discoveries calling for application. Kuhn (1977, p. 90) makes this point in an example from the history of energy conversion processes, showing that the vast majority of the pioneers who had some degree of success in quantifying conversion processes were engineers who actually worked with engines.

Rabkin (1992, p. 66) makes the same point again with regard to the sequence of events assumed in a 1965 US National Academy of Sciences report. Rabkin says that the usual "scheme seems to be at variance with much of the evidence in the history of science. It has been shown that the integration of instruments has been rarely due to the demand on the part of the researcher. Rather it occurs through vigorous supply of advanced instruments on the part of the industry."

And so it is said that "thermodynamics owes much more to the steam engine than ever the steam engine owed to thermodynamics" and that "the chemical revolution resulted much more from the technique of the electric battery than from the careful measurements or new theories of Lavoisier" (Price, 1986, pp. 240, 248).

Thus, contrary to the popular perception of technology as a product of academic research science, it often, if not usually, happens that widespread commercial applications of a new technology precede the science based on that technology.

Many academic researchers believe that their measurement-theoretic quantitative tests and tools have a practical capacity to achieve results that are not accessible by other methods' lax standards. But where is the cutting edge in precision test- or survey-based measurement? What research publications set the pace and establish the standard, and how do they compare with the measures employed at the big educational and psychological test publishers?

To take the handiest example, I contend that Stenner *et al.* (2006) describe the state of the art in measurement applications in reading education. These applications currently involve about 20 million US students, 100,000 books, tens of millions of magazine articles, in English and Spanish, and every major children's book, elementary and secondary textbook, and reading test publisher. All of the work was performed by MetaMetrics, Inc. and its business partners (some of it with funding from the NIH's *Small Business Innovation Research* program). Nothing of comparable precision or validity, not to speak of widespread application, has yet been accomplished in academic research on reading measurement. Stenner's Lexile Framework for Reading is the living embodiment of the "vigorous supply of advanced instruments on the part of industry," as Rabkin puts it.

The bottom line is that there is considerable truth and value to be found in Ernest Rutherford's comment that, if you cannot understand the results of your experiment without doing a statistical analysis, then you should have done a better experiment (quoted in Wise, 1995, p. 11). Proper measurement in a clean experimental design obviates the need for complex and difficult statistical manipulations. Universal uniform reference standard metrics go the further distance of obviating the need for meta-analytic syntheses of different experiments, since everyone everywhere is able to see their results expressed in the same unit.

But if the history of science is to be believed, we're not going to have that kind of common language on any appreciable scale in outcomes research in education and health care until they can be made commercially viable. Alternative approaches that go against the historical grain might be worth considering, but the odds would seem to favor a new iteration of the old pattern.

William P. Fisher, Jr.

Kuhn, T. S. (1977). The essential tension: Selected studies in scientific tradition and change. Chicago, Illinois: University of Chicago Press.

Price, D. J. de Solla. (1986). Of sealing wax and string. In Little Science, Big Science--and Beyond (pp. 237-253). New York: Columbia University Press.

Rabkin, Y. M. (1992). Rediscovering the instrument: Research, industry, and education. In R. Bud & S. E. Cozzens (Eds.), Invisible connections: Instruments, institutions, and science (pp. 57-82). Bellingham, Washington: SPIE Optical Engineering Press.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2006). How accurate are Lexile text measures? Journal of Applied Measurement, 7(3), 307-22.

Wallace, A. F. C. (1972). Rockdale: The growth of an American village in the early Industrial Revolution (Technical drawings by Robert Howard). New York: W. W. Norton & Company.

Wise, M. N. (Ed.). (1995). The values of precision. Princeton, New Jersey: Princeton University Press.

The Combined Gas Law and a Rasch Reading Law

Many physical laws are expressed as universal conditionals among variable triplets. Newton's second law, for example, formalizes the relationship between mass and acceleration when holding force constant (i.e., conditioning on) as F = MA. Similarly, the combined gas law specifies the relationship between volume and temperature conditioning on pressure. After transformation (e.g., log pressure + log volume - log temperature = constant, given a frame of reference specified by the number of molecules), each of these laws can be abstracted to a common form (a + b – c = constant). Note that these laws permit causal claims expressible as counterfactual conditionals. If we have 20 Liters of a gas at 2000° K under 20 atmospheres of pressure and we cool the gas to 1000° K, we will observe a decrease in the pressure to 10 atmospheres.

The value of such laws may lie more in the explicit causal organization of key constructs than in accuracy of prediction in the real world. Cartwright (1983) made a useful distinction between "the tidy and simple mathematical equations of abstract theory, and the intricate and messy descriptions, in either words or formulae, which express our knowledge of what happens in real systems made of real materials" (p. 128). This distinction led Cartwright to the view that "fundamental equations do not govern objects in reality; they govern only objects in models [i.e., idealizations]" (p. 129).

The human sciences, for the most part, lack laws such as those stated above and consequently lack causal stories that are universal in application: "Lacking a 'complete (causal) theory' of what influences what, and how much, we simply cannot compute expected numerical changes in stochastic dependencies when moving from one population or setting to another" (Meehl, 1978, p. 814, emphasis in original). In this note we build on the abstracted formalism derived above and imagine the form of a Rasch Reading Law.

Table 1
Comprehension Rates for Readers of Different Ability with Texts of the Same Readability or How Reader Ability and Comprehension Rate Relate Under Constant Text Readability

Reader Ability	*Sports Illustrated* Readability	Comprehension Rate
500L	1000L	25%
750L	1000L	50%
1000L	1000L	75%
1250L	1000L	90%
1500L	1000L	96%

Contemporary reading theory recognizes three related constructs: reader ability (a stable attribute of persons), text readability (a stable attribute of text), and comprehension (the rate at which a particular reader makes meaning from a particular text). As a result of 25 years of ongoing research, we know that comprehension is a function of the difference between reader ability and text readability (Stenner & Burdick, 1997). Table 1 illustrates the relationship between reader ability and comprehension rate with text readability held constant. With increasing reader ability, the model forecasts increasing comprehension rate conditioning on text readability. This description of the relationship between reader, text, and comprehension echoes the description of the combined gas law (Table 2).

Table 2
How Temperature and Pressure Relate Under Constant Volume

Temperature	Volume	Pressure
2000 k	20 L	20.0 atm
1000 k	20 L	10.0 atm
500 k	20 L	5.0 atm
250 k	20 L	2.5 atm
125 k	20 L	1.25 atm

In fact, logit transformed comprehension rate + text measure - reader measure = the constant 1.1 (given a frame of reference that specifies 75% comprehension whenever text measure = reader measure). Therefore, a + b – c = constant holds as the common abstracted form of both the combined gas law and the Rasch Reading Law as well as many other physical laws. Below are several causal corollaries of the Rasch Reading Law.

(1) For any reader (and thus for all readers), an increase in text measure causes a decrease in comprehension.

(2) For any reader (and thus for all readers), a decrease in text measure causes an increase in comprehension.

(3) For any text (and thus for all texts), an increase in reader ability causes an increase in comprehension.

(4) For any text (and thus for all texts), a decrease in reader ability causes a decrease in comprehension.

Corollaries such as those above are consequences of the highly abstracted a + b - c = constant, holding in a domain of enquiry. Tables 1 and 2 concretize this abstraction for the gas law and reading law. The Rasch model, in concert with a substantive theory, is a powerful tool for discovering and testing the adequacy of such formulations. Note, however, that the fact that data fit a Rasch model says nothing about causality. Rasch models are associational rather than causal. Substantive theory provides the causal story for the variation detected by a measurement procedure. Specification equations formalize these causal stories and allow precise predictions.

These causal explanations have truth built into them. When I infer from an effect to a cause, I am asking what made the effect occur, what brought it about. No explanation of that sort explains at all unless it does present a cause, and in accepting such an explanation, I

am accepting not only that it explains in the sense of organizing and making plain, but also that it presents me with a cause. (Cartwright, 1983, p. 91)

If one of our children cannot summarize what he just read in his fifth grade science text, we explain this by pointing out that he is a 580L reader and the text book is at 830L. The equation that models comprehension rate as a function of the difference between reader measure and text measure produces an expected comprehension rate below 50%. We hypothesize that the child's failure to produce a good summary has a cause: low comprehension. Suppose that we go to the Web and find a 600L article on the same science topic, and the child reads the article and produces a coherent summary of the text. We conclude that, indeed, low comprehension was the cause of poor summarization. Manipulating the reader-text match caused an increase in comprehension, which in turn caused a change in summary performance. Clearly I am inferring from effect to probable cause. Note that this explanation is unintelligible "without the direct implication that there are [readers, texts and comprehension rates]" (Cartwright, 1983, p. 92).

We wonder how many other variable triplets in the human sciences can be abstracted to the form a + b – c = constant. The implications of this kind of law-making for construct validity should be evident (see Borsboom, 2005).

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

Borsboom, D. (2005). Measuring the mind. Cambridge: Cambridge University Press.

Cartwright, N. (1983). How the laws of physics lie. New York: Oxford Press.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Stenner, A. J., & Burdick, D. S. (1997). The objective measurement of reading comprehension: In response to technical questions raised by the California Department of Education technical study group. Unpublished manuscript. www.lexile.com/lexilearticles/objective-measurement-reading-response.pdf

The Measurement of Psychological Value

Insights from: L. L.Thurstone: "The Measurement of Psychological Value." In Thomas Vernor Smith and William Kelley Wright (eds), Essays in Philosophy by Seventeen Doctors of Philosophy of the University of Chicago. Chicago: Open Court (1929): 157-174. at spartan.ac.brocku.ca/~lward/inventory.html

"Some of the postulates underlying physical measurement are so obvious and so common that ordinarily they need not be stated. But when these same postulates are used for psychological measurement they need explicit formulation. One of these postulates is that a *measurement describes only one attribute of the object. ....* you cannot be completely described in a single measurement any more than a table can be completely described by merely counting the number of its legs. No matter what the object of measurement may be, the measurement describes only one attribute of the object." *[Emphasis: Thurstone's]*(p. 158)

"Another postulate that underlies all measurement is that *the measured attribute is always uni-dimensional. ...* If a series of landscape pictures is arranged by a group of judges in order of estimated excellence or artistic merit, it is tacitly assumed that it is possible to allocate all of the pictures to as many points in a single continuum of excellence, no matter how much the judges might object to so direct a statement of what they are doing." (p. 159)

"This leads to another fundamental consideration. It is possible to describe the attitude of an individual toward an object by allocating him to a point on the affective continuum. But it is also possible to describe the object by allocating it to the same point on the same continuum. Here we see that the measurement of the attitude of people toward an object or idea, and the measurement of the psychological value of the object are identical operations. If the measurement. is used as a description of a person or of a group of people, then it is a measurement of attitude. But if the same measurement is used as a description of the object or idea, then it is a measurement of the psychological value of the object. These two concepts, attitude and psychological value, as here defined, are quantitatively identical. They differ only in the purposes to which they are put. They are the two faces of the same thing." (p. 163)

"The criterion of internal consistency, the additive criterion, now demands that the distance between any two points on this line should agree with the experimental determination of the separation between these two points. This condition must be satisfied within the errors of measurement for all the possible pairs of stimuli in the series. No quantitative description of anything can be called a measurement except in so far as this additive criterion is satisfied. It is so obvious in physical measurement that it need rarely be stated. ...The uni-dimensionality of the scale values of the stimuli is demonstrated when this additive criterion is satisfied." (pp. 171, 174)

A SAS Solution to Simulate a Rasch Computerized Adaptive Test

As far as we know, few or any software programs are currently available to simulate adaptive testing. One of them is POSTSIM distributed by Assessment Systems Corporation. This software package is mostly limited by its inability to allow users to modify the program or to add new routines. In fact, scientists exploring the characteristics of adaptive tests usually develop these programs themselves and for themselves. Consequently, the research community's access to them is difficult. Also, when available, they are not versatile, neither is the source code that goes with them in order to favor adaptations.

To palliate this situation and support adaptive testing research, SIMCAT 1.0, a SAS solution, was proposed (Raîche and Blais, 2006c). A preliminary version of this program had been used by Raîche and Blais (2001) to study the sampling distribution of the proficiency level in adaptive testing according to two stopping rules: taking into account the number of administered items and the standard error of the estimated proficiency level. The new version gives access to improvements as regards to the program versatility and to new proficiency level estimation methods. The Rasch dichotomous response model is the one retained.

Expected a posteriori proficiency level estimation method (EAP) is applied to compute estimated provisory and final proficiency level. The new proficiency level estimation methods are all adaptive modifications brought to the EAP method (Raîche and Blais, 2001, 2006b). These methods are all adaptive in the a priori proficiency level estimation, the proficiency level estimation bias correction, the integration interval, or a combination of them. The use of these adaptive EAP estimation methods diminishes considerably the shrinking, and so biasing, effect of the estimated a priori proficiency level encountered when this a priori is fixed at a constant value independently of the previously computed value of the proficiency level.

Another of the program's peculiarity is its feasibility to compare theoretical values of the standard error, skewness, and kurtosis of the estimated proficiency level with the empirical values of these statistics. This according to predetermined sampling distributions of the estimated proficiency level.

The program, a 20 pages manual describing how to use it with a sample of results and the source code are available from by email: *raiche.gilles -at- uqam.ca*

Gilles Raîche, Université du Québec à Montréal
Jean-Guy Blais, Université de Montréal
Martin Riopel, Université du Québec à Montréal

Blais, J.-G. and Raîche, G. (2006a). Features of the estimated sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules. In M. Garner, G. Engelhard, M. Wilson and W. Fisher (Eds.): Advances in Rasch Measurement. Volume 1. Maple Grove, MN: JAM Press.

Raîche, G. and Blais, J.-G. (2006b). Considerations about expected a posteriori estimation in adaptive testing : adaptive a priori, adaptive correction for bias, and adaptive integration interval. In M. Garner, G. Engelhard, M. Wilson and W. Fisher (Eds.): Advances in Rasch Measurement. Volume 1. Maple Grove, MN: JAM Press.

Raîche, G. and Blais, J.-G. (2006c). SIMCAT 1.0 - A SAS Computer Program for Simulating Computer Adaptive Testing Applied Psychological Measurement. Applied Psychological Measurement, 30(1), 60-61.

Raîche, G., and Blais, J.-G. (2001). [Study of the sampling distribution of the estimated proficiency level in Rasch-based adaptive testing according to two stopping rules]. Mesure et évaluation en éducation 14(2-3), 23-39.

Computer Adaptive Tests (CAT), Standard Errors and Stopping Rules

The standard error of measurement (S.E.) is widely used for stopping a computer-adaptive test. For instance, if the current measure estimate is more than 1.96 S.E.s from the pass-fail measure, then there is 95% confidence in the pass-fail decision. Or 2.58 S.E.s for 99% confidence. But how many items are needed to reach a desired S.E.?

If a person has probability, P, of succeeding on a dichotomous item (such as a multiple-choice question), then the statistical information in the response is P*(1-P). The standard error of the estimated measure is

$$S.E. = 1/\sqrt{\text{information}} = 1/\sqrt{\sum P(1-P)}$$

The largest information, and so the smallest standard error, occurs when P=0.5, i.e., when the CAT items are targeted exactly on the persons. But this can produce an unsatisfactory testing experience for the examinee so higher probabilities of success are targeted, such as .7 (for 70% success) and .8 (for 80% success). Here is a Table showing the targeting, standard error, and minimum number of items administered for a specific S.E.:

Minimum number of CAT Items Administered						
Targeting Probability	**S.E. (Logits)**					
	0.5	**0.4**	**0.3**	**0.2**	**0.15**	**0.1**
P=0.5	16	25	45	100	178	400
0.6	17	27	47	105	186	417
0.7	20	30	53	120	212	477
0.8	25	40	70	157	278	625
0.9	45	70	124	278	494	1112

It is seen that the penalty for going from P=0.5 to P=0.6 targeting is the administration of about 5% more items. From P=0.5 to P=0.7 is about 20% more items. From P=0.5 to P=0.8 is 60% more items. P=0.9 almost triples the test length. An S.E. of 0.15 logits requires about 10 times as many items as an S.E. of 0.5 logits.

Minimum Number of Items for 95% Confidence (|t|>=1.96) in Pass-Fail Decision										
Targeting Probability	**Logit Distance of Ability Estimate from Pass-Fail Point**									
	1	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**	**0.4**	**0.3**	**0.2**	**0.1**
P=0.5	16	19	25	32	43	62	97	171	385	1537
0.6	17	20	26	33	45	65	101	178	401	1601
0.7	19	23	29	38	51	74	115	204	458	1830
0.8	25	30	38	49	67	97	151	267	601	2401
0.9	43	53	67	88	119	171	267	475	1068	4269

John Michael Linacre

Pre-Post Gain on an Item

Question: "My sample were tested on the same items (more or less) before the intervention and after it. How can I compute gain on each item for the sample?"

Answer: In raw score terms, the movement is the average rating on an item at post-intervention minus the average rating at pre-intervention. This is probably good enough provided the data are reasonably complete. In measurement terms, you could do a "stacked" analysis of the pairs of pre- and post- records. Then the measured gain on an item is (the average overall ability of the post-sample minus the average overall ability of the pre-sample) + (the item's pre-post item DIF measure difference). So that if the overall sample has gained 2 logits, and the item's pre-post DIF indicates 1 logit easier at post-, then the sample has gained 3 logits on that item.

Deceptive Percentages

The Wall Street Journal Editorial Page, July 25, 2006:
"If the real difference between two groups, measured as it should be with means and standard deviations, remains constant, ... you can generate a curve that predicts how the point gap will change as tests are made easier or harder or as students become more or less competent."

"*Question:* Doesn't this mean that the same set of scores could be made to show a rising or falling group [percentage] difference just by changing the definition of a passing score?

Answer: Yes. At stake is not some arcane statistical nuance. The US federal government is doling out rewards and penalties to school systems across the country based on changes in pass percentages. It is an uninformative measure for many reasons, but, when it comes to measuring one of the central outcomes sought by No Child Left Behind, the closure of the achievement gap that separates poor students from rich, Latino from white, and black from white, the [percentage] measure is beyond uninformative. It is deceptive. "

Charles Murray
W.H. Brady Scholar at the American Enterprise Institute
www.opinionjournal.com/editorial/feature.html

Mathematics and Empirical Science

"Those who firmly believe that rigorous science must consist largely of mathematics and statistics have something to unlearn. Such a belief implies emasculating science of its basic substantive nature. Mathematics is contentless, and hence - by itself - not empirical science. As will be seen, rather rigorous treatment of content or subject matter is needed before some mathematics can be thought of as a possibly useful (but limited) partner for empirical science."

Louis Guttman in S. Levy (Ed.), Louis Guttman on theory and methodology: Selected writings (p. 82). Brookfield, VT: Dartmouth Publishing Company. Courtesy of William P. Fisher