Rasch Measurement in Language Education Part 2:

Measurement Scales and Invariance

by James Sick, Ed.D. (J. F. Oberlin University, Tokyo)

Part 1 of this series presented an overview of Rasch measurement theory (RMT), particularly in comparison to classical test theory (CTT) and item response theory (IRT). This installment will focus on measurement scales and the property of invariance, a feature that Rasch theorists argue is fundamental to all measurement. Readers wishing to pose questions for discussion or request further elaboration are invited to email jimsick@iname.com. Let us start off by responding to a question from the previous article.

Q: In the first installment you talked about Rasch measures having an "arbitrary origin." I understand what zero means when it refers to a raw score: the test taker missed every single item. But on a logit scale, it seems that zero can refer to anything. Surely, the numbers on a scale must mean something.

A: Thank you for requesting clarification on this crucial point. Before discussing Rasch measurement and the property of invariance, it is worth reviewing measurement scales in general. If you are unfamiliar with Stevens' (1946) classification of measurement scales, have a look at Brown (1988, pp. 20-28) or one of the many online resources such as Becker (1999). First, let's draw a distinction between *measuring* and *counting*. We can count the number of teacups on a table and report that there are precisely ten, no more or no less. To report the *weight* of the ten teacups, however, requires that we employ an instrument such as a balance or digital scales that has been calibrated to some benchmark, such as a gram or a pound. Moreover, the comparison to the benchmark standard is bounded by a limit of precision. By definition, all measurement is approximate, reported within a stated or unstated band of error. Measurement precision is limited not only by the accuracy of the measuring device but, some would argue, by the eyesight or fastidiousness of the human being who observes and reports it.

Raw scores are essentially counts. Unlike measures, they can be reported with absolute precision. In CTT, raw score counts are viewed as reasonable approximations of continuous, interval scale measures, their precision estimated by comparison to a hypothesized "true score" that would emerge if the test were taken repeatedly. Rasch theorists, in contrast, argue that raw scores should be regarded as rankings only. That is, they are ordinal rather than interval scales. When conducting a Rasch analysis, we are attempting to move beyond raw score rankings by using the item difficulties to estimate the *distance* between rankings, thus constructing true, continuous, interval scale measures.

"Rasch theorists . . . argue that raw scores should be regarded as rankings only."

Interval scales, according to Stevens (1946), do not have a natural, absolute zero point. Temperature, for example, is measured in relation to a designated benchmark such as the freezing point of water for the Celsius scale, or the lowest temperature that Daniel Fahrenheit was able to produce in his laboratory in 1724 for the eponymous Fahrenheit scale. Some scales, labeled *ratio scales* in Stevens classification, appear to have a natural zero. Weight or height are often given as examples. Steven's classification, however, is not without challengers (see Velleman & Wilkinson, 1993; Wright, 1997). One might philosophically question whether it is meaningful to say that an object has a height or weight of zero. It is usually more meaningful to measure the height of a mountain in relation to the sea, or the height of a person in relation to the surface he or she is standing on. At any rate, psychological attributes, like temperature, must usually be scaled in reference to something else. In Rasch measurement, we generally use the center of the range of item difficulties as a default origin, but zero can be anything we like, so long as we make that clear to ourselves and other end users.

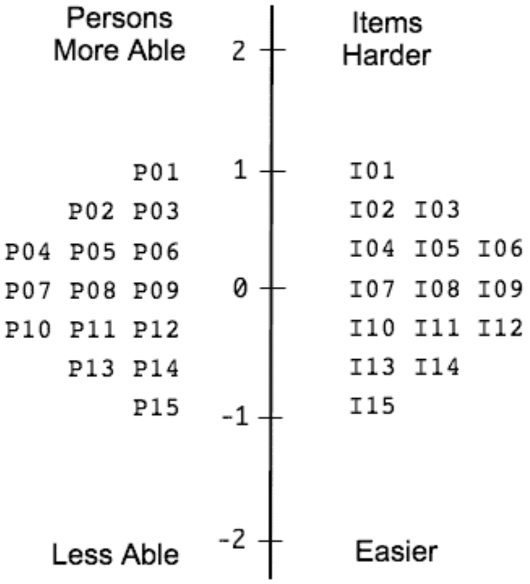
That said, your point is well taken. Making a measurement scale meaningful to end users presents a challenge to the Rasch practitioner. One method is to avoid zero and negative measures by applying a linear transformation to the logit scale. A common transformation sets the item or person mean at 50 and a logit unit to 10 or 15 points. Such a scale typically ranges from about 20 to 80 and resembles a more familiar percentage scale. We could also link our scale to some historical benchmark, such as "60 is equivalent to the average proficiency of students who entered our institution in 2005." Ultimately, it is shared qualitative associations that make any scale meaningful. We know from past experience that 40 degrees Celsius is uncomfortably hot, and that a person scoring 500 on a TOEFL® IPT is reasonably prepared for study abroad. Raw score counts might be meaningful to the person who created the test, but are of limited use in inferring general ability or making comparisons to another test. For that we need a benchmarked scale that is invariant across samples and occasions.

What is invariance?

In principle, the property of invariance means that the ability of the test takers is independent of the test they take, and the difficulty of the test items is independent of the test takers. We like to believe that the attribute we are measuring is a property in an individual's brain, something that the test taker brings to the test. Our measurement of this attribute should not depend on the sample of items we have selected to measure it. Likewise, we hope that the difficulty of the items is an inherent property of the construct, something we can qualitatively predict from theory and experience. In practice, invariance means that once measurement instruments have been calibrated to a common scale, estimates of person ability and item difficulty should not vary across tests and person samples by more than the error that is a consequence of measurement precision.

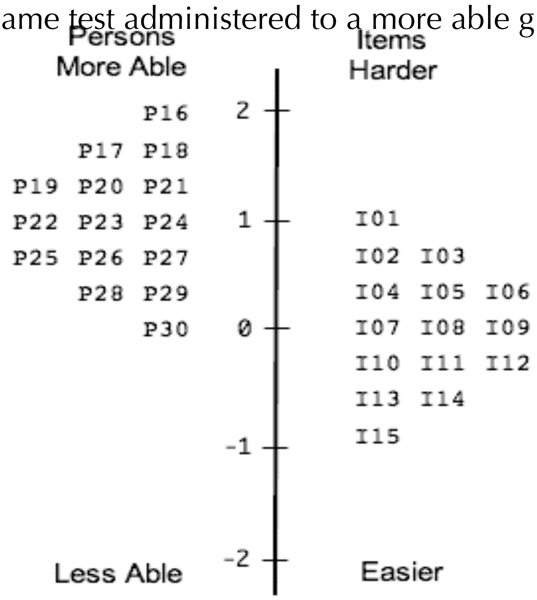
To envision how this works, imagine a short test administered to a sample of test takers who are well-matched to the test. That is, the mean item difficulty closely matches the mean person ability. We use the customary default and set the zero point to the mean item difficulty. This is illustrated in Figure 1, where we can see that Items 7, 8, and 9, and Persons 7, 8, and 9 have measures near the arbitrary zero point.

Figure 1. A short test administered to a well-targeted group (illustrative).



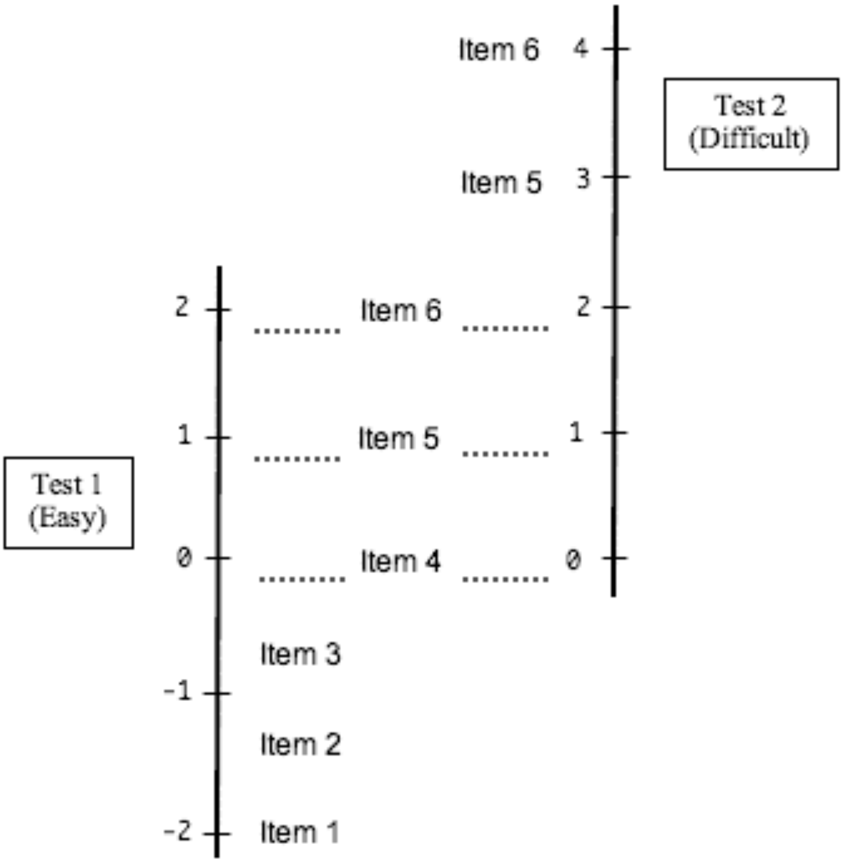
What would happen if we were to administer these same 15 items to a group of test takers of somewhat higher ability? Would that affect our estimates of item difficulty? In CTT, where item difficulties are expressed as the percentage of test takers who answered correctly, a more able group of test takers would produce higher difficulty values. In a Rasch analysis, discounting measurement error, the difficulty estimates should be exactly the same, provided the origin is again set at the mean item difficulty. The mass of persons would move up the scale, however, making their scores comparable to the first group, as illustrated in Figure 2.

Figure 2. The same test administered to a more able group (illustrative).



But what if the first group where administered a different test? How could person and item invariance be maintained in that case? In order to link different tests, it is necessary to have some common items that serve as benchmarks, similar to how the freezing and boiling points of water serve to define the Celsius temperature scale. Figure 3 illustrates how two tests with three common items can be linked to a common scale. In Test 1, Item 4 is at the center of the range and has been set at the default value of zero logits. In Test 2, the easier items have been dropped and more difficult ones added. To construct a common scale, we merely declare that Items 4, 5, and 6 will retain their original values of 0, 1, and 2 logits, respectively. All new item and person measures will be estimated in relation to the common items, generally referred to as anchor items. As we can see in Figure 3, Item 4 is now the *easiest* item rather than an average item, and the origin of our scale has been shifted from the item mean down to the easiest item. What would be the result if a person took both tests? Assuming her ability had not changed between tests, her logit measure should be the same. If her ability were 1 logit on our common scale, we would expect her to answer about two-thirds of the items on Test 1 correctly, but only one third of the items on Test 2. Even though her raw scores have changed, her ability expressed in logits is the same on both the easier and the harder test.

Figure 3. Two tests linked by common items.



Q: Are item characteristics really so dependable that we can expect them to generate the same Rasch calibrations in all conceivable testing contexts?

> *"Rasch item calibrations are population parameters estimated from sample data and should always be regarded as approximate."*

A: Alas, no. Screening items for invariance and selecting ones suitable to serve as anchor items is a complex process that is beyond the scope of this article. To begin with, Rasch item calibrations are population parameters estimated from sample data and should always be regarded as approximate. As such, each item calibration has a standard error of estimate, and invariance must be evaluated in relation to measurement precision. We generally consider an item to be invariant when its difficulty calibration has not changed by more than two standard errors across occasions.

Apart from variation due to measurement error, item characteristics can be influenced by various external factors. Explicitly teaching the content of a single test item, for example, might make that item easier, relative to other items, than it was previously, in effect changing its rank in the item hierarchy. Sometimes social events affect an item's difficulty. An interesting example is counting backwards from 10 to 0. Apparently, in the 1950s this was a cognitively challenging arithmetic item, but with the arrival of televised rocket launches its relative difficulty decreased (Linacre, 2000). When an item's difficulty drifts considerably, we must treat it as if it were a new item and re-calibrate it rather than anchor it to a previous value. We might say that while the item's form has remained the same, the inferences we can justifiable draw from it have changed. Thus, in an inferential system such as RMT, we regard it as a different item. As a final comment, I should remind readers that invariance is a property of the Rasch model and for the property to hold, the items must show reasonably good fit to the Rasch model. Before using RMT to link tests, the practitioner must confirm that the items have a unidimensional hierarchical structure and prune any items that deviate from this structure substantially.

More information on testing for invariance and selecting anchor items can be found in Wolfe and Chui (1997; 1999), Wright (1996), and Sick (2007, pp. 119-134). The next installment of this series will examine the family of Rasch models in greater detail and discuss how they can be applied in various measurement situations.

References

Becker, L. (1999). *Scales of measurement.* Retrieved April 5, 2008 from http://web.uccs.edu/lbecker/SPSS/scalemeas.htm#1.%20Overview

Brown, J. D. (1988). *Understanding research in second language learning: A teacher's guide to statistics and research design.* Cambridge, England: Cambridge University Press.

Linacre, J. M. (2000). Computer-adaptive testing CAT: A methodology whose time has come [Electronic Version]. *MESA Memorandum 69*. Retrieved April 5, 2009 from http://www.rasch.org/memo69.pdf.

Sick, J. R. (2007). The learner's contribution: Individual differences in language learning in a Japanese high school. (Doctoral Dissertation, Temple University Japan, 2007). Available through ProQuest Dissertation Abstracts International, UMI: 3255147.

Smith, E., Jr. (2000). Metric development and score reporting in Rasch measurement. Journal of *Applied Measurement, 1*(3), 303-326.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science,* 103, 677-680.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*(1), 65-72. Retrieved April, 2008 from http://www.spss.com/research/wilkinson/Publications/Stevens.pdf.

Wolfe, E. W., & Chiu, C. (1997). Measuring change over time with a Rasch rating scale model. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC document reproduction service No. ED 408 325).

Wolfe, E. W., & Chiu, C. (1999). Measuring change across multiple occasions using the Rasch rating scale model. *Journal of Outcome Measurement,* 3(4), 360-381.

Wright, B. D. (1996). Time 1 to time 2 comparison [Electronic Version]. *Rasch Measurement Transactions,* 10 (1), 478-479. Retrieved April 1, 2008 from http://www.rasch.org/rmt/rmt101f.htm.

Wright, B. D. (1997). S. S. Stevens revisited [Electronic Version]. Rasch Measurement Transactions, 11 (1), 552-553. Retrieved April 1, 2008 from http://www.rasch.org/rmt/rmt111n.htm.

Wright, B. D., & Stone, M. (1979). Best test design: Rasch measurement (Chapter 8: Choosing a Scale). Chicago: Mesa Press. 191-209.

Last Next

http://jalt.org/test/sic_2.htm (HTML)

http://jalt.org/test/PDF/Sick2.htm (PDF)