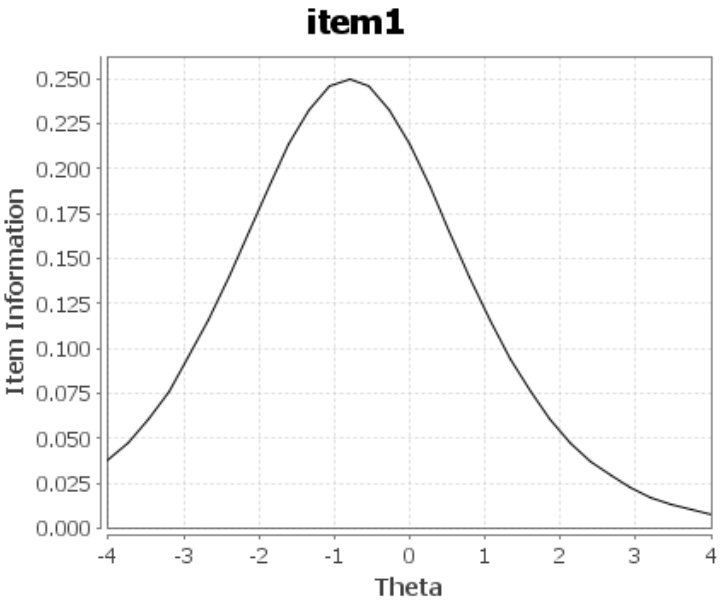
Center for Response to Intervention in Early Childhood
University of Minnesota
Michael C. Rodriguez
2013

Rasch Score Precision and Reliability

In the Rasch model, score precision is a function of the information available to estimate a given score – a given ability level. Information is a direct function of the number of items located at that ability level: more items yield more information.

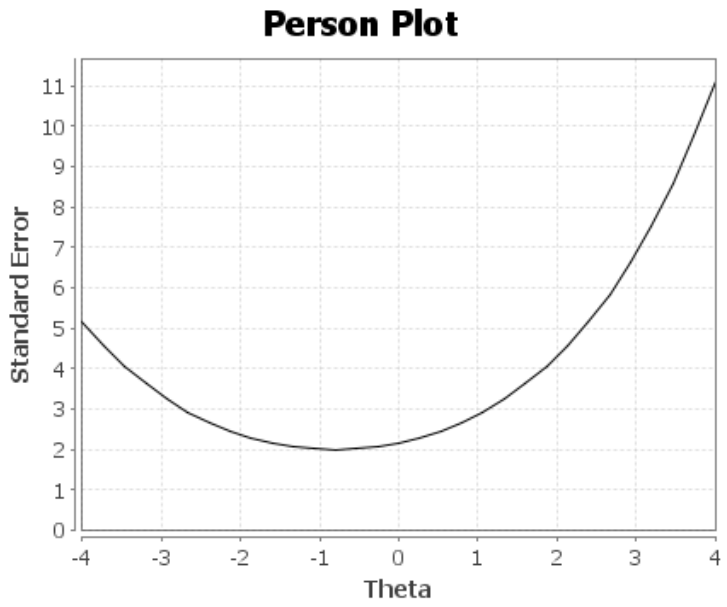
Information is the inverse of the standard error of the estimated ability. The standard error is thus also a direct function of the number of items (item responses from a person) available to estimate ability. Because a given item provides different levels of information across the ability continuum, with maximum information at the location (difficulty) of the item, each ability level has a different level of precision.

For a single item, the item information might look something like the following figure:



Notice that this item is located near theta or ability of -0.8, so the information is at its maximum at this point on the ability scale. Also note that in the Rasch model, the location of an item is the ability level required to have a 50% chance of correctly responding to the item.

The standard error of ability estimates given this item is a function of the inverse of information, something like the following:



Here we see that the standard error is smallest near theta of -0.8. However, the standard errors are large in this case because we are considering only one item. With 15 items, the information functions are summed across the ability continuum and the standard error is again an inverse function of the sum of the information functions – thus the standard error will be much smaller with many more items. Test information, $I(\theta)$, is the sum of item information curves and is defined at a given ability (θ).

The standard error is the square root of the inverse of information.

$SE(\theta) = \sqrt{\frac{1}{I(\theta)}}$ is the standard error of a test at θ, the precision of the ability estimate.

Some have attempted to answer the request for information about score reliability more globally. This can be done by squaring each of the SE values for a set of scores (making them error variances), computing their average, and computing the ratio of true-score variance to observed-score variance:

In Classical Test Theory (CTT), reliability is conceived as the ratio of true-score to observed-score variance (Observed-score variance = True-score variance + Error-score variance).

$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2}$. In the Rasch model, error variance, σ_e^2, can be estimated by using the average person error variance (mean squared measurement error). The observed-score variance is estimated directly as the variance of Rasch scores in the sample.

Score Reliability

A group estimate of reliability can be estimated from person *SE*s. These *SE*s can be squared and summed to compute the average error variance for the sample. The average squared person *SE* is the MSE_p, the mean squared error across persons (error-score variance). This can be subtracted from the observed variance SD^2 to estimate of true variance SA^2:

$$MSE_p = \sum_{n=1}^{N} \frac{s_n^2}{N}$$

This is the mean square error, the average error variance across N persons.

$$SA_p^2 = SD_p^2 - MSE_p$$

This is the estimated true variance (total variance minus error variance).

$$R_p = \frac{SA_p^2}{SD_p^2} = 1 - \frac{MSE_p}{SD_p^2}$$

This is the estimated person score reliability, based on the ratio of true variance to observed variance.

Cautionary Note in Applying Internal Consistency Estimates to Criterion-Referenced Measures

Estimates of internal consistency reliability depend on score variability. The optimal condition for internal consistency is when applied to norm-referenced tests, as the goal is to spread scores out to facilitate norm-referenced interpretations.

In criterion-referenced testing (CRT), the goal is to enable inferences about what individuals know or can do. The test and resulting score scale are not specifically designed to spread people out to estimate individual differences. CRTs often result in narrow score ranges, particularly when tailored to test takers at a particular level of ability (minimum skills tests, mastery tests, etc.). Such narrowing of the score scale makes internal consistency estimates of reliability significantly less useful for CRTs.

Additional models of reliability are generally used for CRT scores, including indices of decision consistency and generalizability.

For more information, see:

Schumacker, R.E., & Smith, E.V. (2007). Reliability: A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394-409.