Estimation Methods for Rasch Measures

Linacre, J. M. (1999) . Understanding Rasch measurement: estimation methods for Rasch Measures. Journal of Outcome Measurement, 3, 381-405 .

Rasch measurement is the only way to convert ordinal observations into linear measures (Fischer, 1995). These measures are represented as parameters in a Rasch model and are estimated from the data. The analyst, however, is rarely concerned about the estimation process, provided that it succeeds in obtaining reasonable values for the measures. An appreciation of the different methods of estimation, however, will better enable the analyst to evaluate what "success" means.

Estimation, Precision and Accuracy

When a new measurement situation is encountered, the Rasch measures are not known. The measures, the parameters of a Rasch model, must be inferred from data. This is accomplished by means of the method of "inverse probability", first addressed by Jacob Bernoulli (1713). The resultant estimates of the parameters are values obtained There is no "one best method" of estimation, nor is there "one best set" of Rasch estimates. Parameters are always estimated with imprecision and inaccuracy.

In this context, precision relates to the uncertainty in the estimated location of the parameter on the latent variable when it is specified that the data fit the Rasch model. The term "assumed" is avoided here, because the fit of the data to the model can be tested immediately. The term "assumed" is used when estimation conditions are introduced which cannot be tested immediately, and so are truly assumptions. When the data are specified to fit a Rasch model, then all unexpectedness in the data are deemed to be products of the stochastic process inherent in the model.

Precision can always be increased by collecting more relevant data or specifying rating scales with more categories, with the continuing condition that the data are specified to fit the model. Precision can be artificially improved by introducing constraints, often as assumptions, which reduce the location uncertainty. The most commonly introduced assumption is that one or more characteristics underlying the data are normally distributed.

Accuracy relates to the departure of the data from those values predicted by a Rasch model given the estimated locations of the parameters. No empirical data set fits the Rasch model perfectly, but "for problem solving purposes, we do not require an exact, but only an approximate, resemblance between theoretical results and experimental ones." (Laudan, 1977). Nevertheless, as the data depart ever further from meeting Rasch model expectations, doubt not only about the locations, but also about the meaning of parameter estimates increases. Accuracy can be increased by collecting more data that is likely to conform to a Rasch model, e.g., by avoiding administering items that are too trivial or too challenging, which are likely to provoke irrelevant behavior in respondents. Accuracy can also be increased by screening out responses deemed irrelevant for measurement purposes. Such responses may be highly diagnostic of idiosyncratic aspects of respondents, items, judges or the rating scale, but they do not contribute to constructing a generalizable measurement system.

In practice, some constraints must be introduced in order to make estimation possible, and some screening of the data must be performed, if the estimates are to be meaningful. In fact, a general assumption of Rasch estimation procedures is that the uncertainty in the data, at the response level, is, in some sense, normally distributed.

Due to the arbitrary nature of pass-fail decisions and the practical need to introduce determinacy into both norm-referenced and criterion-referenced reporting, Rasch estimates (as well as raw scores and other statistics) are usually treated as point-estimates of their underlying parameters. Thus estimates are commonly reported with more significant figures than either their precision or their accuracy supports. Since all estimation methods are approximate, the same estimation method under different conditions, or different estimation methods under the same conditions, may disagree numerically as to whether a subject, near to the pass-fail points, is a "pass" or a "fail". When the precision and accuracy of estimates are taken into account, even conspicuously different estimation methods agree (Wright, 1988).

nables the problemsSuch observations, however, are modeled to be stochastic and non-linear. Hence, the parameters of a Rasch model, the Rasch measures, are not deduced in closed form directly from the data, but are inferred in some approximate or iterative manner by means of the data moderated by other specifications and assumptions.

The Nature of the Rasch model

The basic unidimensional Rasch model postulates that the data are the outcome of a stochastic process governed by a linear combination of parameters. A general form of the model is:

$$\log\left(\frac{P_{nijk}}{P_{nij(k-1)}}\right) = B_n - D_i - C_j - F_k$$

(1)

where
B_n is the ability of subject n,
D_i is the difficulty of item i,
C_j is the severity of judge j,
F_k is the step calibration, the difficulty of the step up from category k-1 to category k, of the ordered rating scale, denoted by successive integers in the interval l,h.

This can also be expressed as:

$$P_{nijk} = \frac{e^{k(B_n - D_i - C_j) - \sum_{g=l}^{k} F_g}}{\sum_{t=l}^{h} e^{t(B_n - D_i - C_j) - \sum_{g=l}^{t} F_g}}$$

(2)

F_l, the step calibration into the bottom category, is a multiplier in every term, so cancels out. It is conventionally reported as 0. Nevertheless, it is a convenience to retain it in estimation because setting F_l to a large enough negative value prevents computational overflows on exponentiating. Such values may cause underflows, but these can be replaced by zeroes with minimal loss in estimation precision.

Were the model parameters to be known, then the probability of observing any particular datum would also be known. The probability of observing datum, X_{nij}, in category *x* is given by (2) above. The expected value of any datum, E_{nij}, is:

$$E_{nij} = \sum_{k=l}^{h} k \, P_{nijk}$$

This has a monotonically increasing ogival form when the rating scale structure, $\{F_k\}$, is held constant, and the sum of the other parameters is allowed to change. The local slope of the ogive is determined by the $\{F_k\}$. This ogive can be approximated by a logistic ogive with slope determined by the $\{F_k\}$.

For any parameter, e.g., $_n$, the marginal score, R_n is the sum of all observations modeled to be generated by $_n$:

$$R_n = \sum_{n} X_{nij}$$

Though this is discrete, it has an ogival form as B_n varies, and can also be approximated by a logistic ogive of suitable slope.

Following Fisher (1922), the likelihood of the data, L, is the product of the probabilities of the data points:

$$L = \prod_{n,i,j} P_{nijx}$$

An Overview of Estimation Methods

For many purposes, simple graphical techniques provide useable, if rough, estimates. Georg Rasch (1960, Ch. VI) demonstrates how plotting logistic transformations of success frequencies permits the drawing of trace lines by eye. These provide estimates of Rasch measures and support an investigation into measure accuracy. When empirical raw-score-based item characteristic curves are available, logistic transformations of both axes, i.e., log(right/wrong), yields plots equivalent to those produced by G. Rasch.

A numerical operation, equivalent to drawing plots, is the method of iterative proportional fitting applied to the cells of a data matrix (Kelderman and Steen, 1988). This is helpful when the logistic form of the Rasch model is reparameterized as a log-linear model, so that each cell contains the count of occurrences of a particular response string. Some estimation methods for log-linear versions of the Rasch model are addressed in Kelderman (1984).

The parameters of log-linear models can be estimated with standard statistical software. Thus such software can be used to estimate the parameters of Rasch models, but this approach has proved of limited utility. Rasch models, in log-linear form, often have hundreds of parameters and so overwhelm the software's computational capacity. Further, many of the cells contain incidental zeroes, because particular response strings did not happen to be observed. If missing data are permitted in response strings, estimation with standard statistical software becomes virtually impossible.

Most estimation methods employ some form of the method of maximum likelihood. The goal of this method, due to Fisher (1922), is to discover the parameter values which maximize the likelihood of the data, under whatever constraints the analyst imposes. An advantage of the method is that, in general, a second derivative of the likelihood function provides a standard error for the estimate.

The choice of constraints optimizes certain aspects of the estimation process or the estimates themselves, but always at a cost. For instance, there is the ideal of estimation consistency. A consistent estimation procedure produces estimates that asymptotically approach their latent values as the size of the data set increases. This might appear to be an essential feature of any estimation procedure, but it is not. First, estimation procedures which are consistent according to one method of increasing the data set, can be inconsistent according to another. Second, the inconsistency may be so small as to have no practical implications. Third, the inconsistency in any finite data set, termed "statistical bias", may be correctable. On the other hand, insisting on estimation consistency may prevent estimation under specific conditions, e.g., in the presence of missing data.

In nearly all estimation methods, extreme (zero and perfect) marginal scores imply infinite parameter values and so are inestimable. Accordingly, data corresponding to extreme scores must be eliminated before estimates are produced. There are separate estimation techniques for imputing reasonable measures to extreme scores, once the measures for non-extreme scores have been estimated.

Estimation methods are classified here according to several major considerations. (i) Is estimation conceptualized as proceeding datum by datum, or at the marginal (raw score per parameter) level? (ii) Are all parameters estimated or are some conditioned out of the estimation? (iii) Are parameters free or are they modeled as part of a distribution?

Estimation datum-by-datum

A Rasch model resembles a simple form of a "transition odds" or "adjacent logit" logistic (logit-linear) regression model. George Udny Yule (1925) and Joseph Berkson (1944) suggest methods for estimating the parameters of a logistic curve. Several methods of their are generally applicable. These methods are generally robust against missing data,

I. Gaussian least-squares.

This estimation method minimizes the sum of the squares of the differences between what is observed and what is expected across the data, D. The function to be minimized, F, is:

$$F = \sum_{D} (X_{nij} - E_{nij})^2$$

This must be minimized for all parameters simultaneously. From the perspective of a particular parameter, say B$_n$, the minimization occurs when:

$$\frac{dF}{dB_n} = \sum_{n} 2 (E_{nij} - X_{nij}) V_{nij} = 0$$

where V$_{nij}$ the model variance of an observed rating about its expectation, is

$$V_{nij} = \sum_{k=l}^{h} (k^2 P_{nijk}) - E_{nij}^2$$

The minimization condition for the rating scale parameters {F$_k$} is more complex, but tractable. From the Rasch measurement perspective, a drawback to this method is that different response strings with the same total raw score produce different measures.

II. Minimum chi-square.

In contrast to the previous method which minimizes numerical distances on the ordinal scale, the minimum chi-square method maximizes the fit of the data to the Rasch model. Consequently, outlying unexpected observations (such as coding errors) are more influential in the minimum chi-square approach and, again, the same marginal score can produce different estimates. The function, F, to be minimized for all parameters simultaneously is:

$$F = \sum_{D} \frac{(X_{nij} - E_{nij})^2}{V_{nij}}$$

III.Pairwise estimation.

Since the Rasch model is a log-odds model, an attractive approach is to use the relative frequencies of observations in the data to estimate the parameters. Suppose that two persons, *m* and *n*, are judged by the same judges on the same items. C$_{xy}$ is the number of times that subject *m* is rated in category *x* in the same circumstances that subject *n* is rated in category *y*. Similarly for C$_{yx}$. Then, an estimate of the difference in ability between *m* and *n* is given by the paired comparison

$$B_m - B_n = \frac{\log\left(\frac{C_{xy}}{C_{yx}} \right)}{x - y}$$

Following this approach, one data set can yield many different estimates of the relative ability of the same pair of subjects, and also many different estimates of the relative abilities of those subjects to other subjects. Further, some of these estimates may involve very few observations or even incidental zeroes.

The resolution of these contradictions is to combine the paired comparisons into a likelihood function (Wright & Masters, 1982) which is maximized when the parameter estimates simultaneously satisfy the relationship

$$\frac{dF}{dB_n} = \sum_{m} \sum_{x=l}^{h} \sum_{y=l}^{h} (y - x) \left(C_{xy} - (C_{xy} + C_{yx}) \left(1 + e^{(y-x)(B_n - B_m)} \right)^{-1} \right)$$

Here the summation over *m* includes every pairing in the data for every *m* for which C_{xy} and C_{yx} are both non-zero. In this summation, observations can be used multiple times. Consequently the standard errors provided by the second derivative are too small, roughly in proportion to the square-root of the average number of times each observation is used.

In this method, one set of parameters, here the subjects, are estimated. Then another set is estimated using the pairwise method and the two sets of estimates are aligned on one measurement continuum. Alternatively, the pairwise estimates are set as fixed values (anchors), and another method is used to estimate the other parameters.

Marginal Estimation without Distributional Assumptions

In marginal models, identical total raw scores, obtained under the same conditions, estimate identical Rasch measures, regardless of the specifics of the response string. This accords with Fisher's (1922) concept of sufficiency, but has been deemed counter-intuitive by empiricists. In general, however, any argument proposing that getting a hard item unexpectedly correct merits a higher measure can be offset by an equivalent argument that getting an easy item unexpectedly wrong merits a lower measure.

Item Response Theory (IRT) models generally require assumptions about the distribution of the latent parameters in order to be estimable. Rasch parameters, however, can be estimated with or without distributional assumptions regarding the parameters. There is one distributional specification, however, that is deemed to hold across these estimation methods. The unmodeled part of each datum, the residual difference between the observed and the expected values, is specified to be normally distributed, when the residual is standardized by its own model variance.

IV. Joint maximum likelihood estimation (JMLE).

Leading directly from Fisher sufficiency, and also from Gaussian least-squares, this method produces estimates for which the observed and expected marginal scores coincide. No parameters are conditioned out, so the method is also termed "unconditional." Though these estimates are independent of the computational details of the method used to obtain them, the usual approach is that of Newton-Raphson iteration. The estimation equation to produce a better estimate B_n' of the previous estimate B_n is:

$$B_{n'} = B_n + \frac{R_n - \sum_n E_{nij}}{\sum_n V_{nij}}$$

This estimation method has proved robust against missing data, and also allows easily the incorporation into one analysis of data generated by variants of the Rasch model (dichotomous, partial credit, rating scale, etc.)

A long-standing criticism of this method is that it is prone to noticeable estimation bias with short tests. For instance, if a two item dichotomous test were given to a sample of persons, the estimated difference between the item measures according to JMLE would be twice that estimated by the pairwise estimation method. In practice, however, this bias has few implications because the relative ordering and placement of the estimates is maintained. When JMLE is used to estimate measures from paired comparison data a correction factor of 0.5 removes the statistical estimation bias.

JMLE is amenable to pre-set (fixed, anchored) parameter estimates, so that it is often used to estimate those parameters which have been left unestimated by other estimation methods.

V.Conditional Maximum Likelihood Estimation (CMLE).

This methods capitalizes on the proposition that identical person raw scores produced under identical conditions imply identical measures, but avoids actually estimating those measures. This results in a method with minimal estimation bias and well-defined standard errors, but which is computationally intensive and generally intolerant of missing data.

The minimal estimation bias results from the very slight probability that a sample of respondents, whose measures correspond to the estimated parameters, would obtain an extreme score on a test item. If a large sample of respondents is obtained, then this probability is effectively zero.

In this method, the relevant probabilities that form the basis for the estimation process are conditional probabilities that build on the simple form of the Rasch model. First, the probability of every possible response string that generates a particular score must be estimated. The sum of these is the denominator. The numerator is the sum of the probabilities of those response strings (which produce the same raw score) in which the response to a specified item has a specified value. Since this division eliminates the parameter estimate corresponding to the raw score, the reference parameter estimate for the score group can be set to zero or any value convenient for computational purposes. Nevertheless, the long sums of exponentials that inevitably result can cause severe computational problems involving loss of computational precision and exponential overflow and underflow. Improvements in computer hardware and more sophisticated numerical methods have aided CMLE (Verhelst and Glas, 1995), but it is still impractical in most instances for long tests. An outline of the estimation method is shown in Wright & Masters (1982), Chapter 4.

Marginal Estimation with Distributional Assumptions

Distributional assumptions regarding some or all of the parameters can be usefully employed in a number of situations to simplify computation or even make estimation possible. If the distributional assumption seriously mismatches the latent parameter distribution, then severe estimation bias may be introduced.

VI.Marginal Maximum Likelihood Estimation (MMLE).

MMLE imposes a distribution function on the subject parameters. The simplest function is a normal distribution (paralleling IRT estimation). More sophisticated functions are also employed such as multivariate normal distributions based on demographic variables (Adams et al., 1977) and empirical-Bayesian distributions.

MMLE can surmount several obstacles at which other estimation methods balk. First, it permits the estimation of sample measure characteristics even when there is insufficient information to produce meaningful estimates for individuals within the samples. In particular, extreme scores, very short response strings and missing data can be easily managed. Second, it bypasses an analytic step when the intention is not to measure individuals, but to summarize estimates. Third, it supports forms of the Rasch model beyond the unidimensional (Wu et al., 1998).

MMLE produces estimates for the discrete parameters, usually corresponding to item difficulties and rating scale structures, such that their observed and expected marginal scores coincide, under the condition that the distribution of the other parameters has the required form.

VII. Normal Approximation Algorithm (PROX).

There is a convenient arithmetical relationship between the unit-normal ogive and the logistic ogive. Berkson (1944) takes advantage of it for bio-assay calculations. Cohen (1988) derives Rasch model estimation equations for dichotomous data from it. Wright & Stone (1987, Chap. 2) use Cohen's algorithm to demonstrate the estimation of Rasch measures by hand.

The relationship between the ogives is specified as:

$$\psi^{-1}(y) \approx 1.7\ \phi^{-1}(y)$$

where Ψ is the logistic function and φ is the normal cumulative function The standard equating value of 1.7 minimizes the maximum difference between the functions across their whole range (Camilli, 1994). Linacre (1997) suggests 1.65 as a better equating value for Rasch use.

When dichotomous data are complete and the parameters of each facet approximate a normal distribution, then non-iterative estimation equations are:

$$B_n = \sum_{i=1}^{L} D_i + X_B \log\left(\frac{R_n}{L - R_n}\right)$$

$$D_i = C_D - X_D \log\left(\frac{R_i}{N - R_i}\right)$$

By convention, $\Sigma D_i \equiv 0$ establishes the local origin of the measurement scale. In practice, C_D is chosen so that $\Sigma D_i = 0$. In principle, $C_D = \Sigma B_n$. Departure of C_D from ΣB_n indicates mismatch between the empirical and PROX-specified parameter distributions.

X_B and X_D are expansion factors to adjust for sample spread and test width.

$$X_B = \left(\frac{1 + S_D/2.89}{1 - S_D S_B/8.35}\right)^{\frac{1}{2}}$$

$$X_D = \left(\frac{1 + S_B/2.89}{1 - S_D S_B/8.35}\right)^{\frac{1}{2}}$$

where S_B and S_D are the population standard deviations given by

$$S_B = S.D.\left(\log\left(\frac{R_n}{L - R_n} \right) \right) \quad n = 1, N$$

$$S_D = S.D.\left(\log\left(\frac{R_i}{L - R_i} \right) \right) \quad i = 1, L$$

Linacre (1994) derives PROX estimation equations for missing data. Linacre (1995) extends PROX to polytomous data.

VIII. Items two-at-a-time.

When tests are short, many subjects obtain extreme scores. These introduce an unquantifiable amount of bias into summary statistics. The focus of measurement, however, may not be the subjects, but the samples to which they belong. When subjects are regarded as normally distributed, Wright (1998b) suggests estimation equations for the sample mean and standard deviation from the responses of subjects to pairs of items.

Table 1 about here

Imagine that a large sample of people have taken two dichotomous items, A and B, approximately as the Rasch model predicts. Table 1 is the tabulation of their scored responses. According to the Rasch model, the difference between the item difficulties is estimated directly by

$$D_A - D_B \approx \log\left(\frac{S_{01}}{S_{10}} \right) \quad with\ S.E. = \sqrt{\frac{S_{10} + S_{01}}{S_{01}\ S_{10}}}$$

If we assume that the sample is normally distributed, then we can estimate the sample mean and standard deviation. The sample mean ability is relative to the average difficulty of the two items. A simulation study reported in Wright (1998b) suggests the following estimator:

$$Sample\ mean \approx 1.864 \left[\log\left(TA\frac{1}{TA}0 \right) + \log\left(TB\frac{1}{TB}0 \right) \right]$$

$$+ 1.455\ \log\left(S\frac{00}{S}11 \right)$$

An estimator for sample standard deviation is:

$$
\begin{aligned}
&S.D. \approx 3.763 \\
&+\ 1.4 * \left[\log\left(S\frac{11}{T - S11} \right) + \log\left(S\frac{00}{T - S00} \right) \right] \\
&+\ 0.0101 * \log\left(S\frac{10}{S}01 \right)^2 \\
&+\ 0.081 \left[\log\left(TA\frac{1}{TA}0 \right)^2 + \log\left(TB\frac{1}{TB}0 \right)^2 \right]
\end{aligned}
$$

Estimating Extreme Scores

Under strict Rasch model conditions, extreme (zero and perfect) scores correspond to infinite measures. Here, infinite has the meaning of indefinite, any value outside the measurement range of the test. Consequently, under most estimation methods, the response vectors corresponding to extreme scores are dropped from the analysis. In many situations, however, measures must be reported for extreme scores, or the measures corresponding to extreme scores must be included in summary statistics.

There are two approaches to imputing measures for extreme scores. The first approach is to consider extreme scores to be part of a measure distribution. This requires an estimation method, such as MMLE, that estimates at the sample, rather than individual, level. The second approach is to apply some reasonable inference about the nature of the extreme score, and use this to estimate a measure.

Wright (1998a) suggests nine bases for choosing a measure corresponding to an extreme score. He concludes that, for dichotomous data, reasonable measures for extreme scores are between 1.0 and 1.2 logits more extreme than the measures for the most outlying non-extreme scores. For polytomous data, measures corresponding to scores between 0.25 and 0.5 score-points more central than the extreme scores can be usefully imputed as the extreme measures.

Estimation Error

A recurring theme in the literature of the Rasch model is estimation error. No estimation technique can guarantee to recover the measures of the generating parameters, even when the data fit the Rasch model. The difference between the estimate and the generators is termed estimation error. There are three main sources of estimation error, deficiencies in the theoretical properties of the estimates, deficiencies in the implementation of the estimation algorithm and mismatches between the distribution of the data and the assumptions of the estimation algorithm.

Some techniques could recover the generators, in theory, if they were provided infinite data of the right kind. For instance, the "two-at-a-time" and pairwise estimation techniques would recover the exact measure difference between items, given the responses of an infinite number of on-target persons under Rasch model conditions. Such estimation techniques are termed "consistent".

Though a desirable property, consistency is not of practical concern.

A theoretical deficiency in most estimation methods causes some degree of estimation bias, which can noticeably affect measures estimated from short tests or with small samples. Even then, the bias can usually be easily corrected (Wright, 1988). An example is the correction of bias in measures resulting from the use of JMLE for analyzing measures from paired-comparison observations (Linacre, 1984). Under Rasch model conditions, estimation bias is due to the inclusion of the possibility of extreme score vectors in the computations of the estimation algorithms, even though they must be eliminated from the data (or other arbitrary constraints introduced), because they produce infinite parameter estimates.

The bias in JMLE is chiefly caused by the likelihood of persons obtaining extreme scores. Linacre (1989) derives a JMLE-based estimation algorithm (XCON) which overcomes this deficiency, but there has been no demand, as yet, to implement it in a generally accessible way. CMLE is relatively bias free, because person extreme scores are eliminated from the estimation space, and there is only a remote possibility of an extreme score for an item.

Deficiencies in implementing estimation algorithms are most apparent with CMLE. Computations of the likelihoods of every possible response string that generates each observed raw score is required. This is a large computational load and, worse, involves the accumulation of many small numbers. Loss of numerical precision can result, leading to error in the estimates.

Mismatches between the distributional assumptions of the estimation algorithm and the data can skew MMLE and PROX estimates. PROX capitalizes on the normal distribution, so that good estimates will not be obtained with a highly skewed sample, such as those found in many clinical situations. MMLE can use more sophisticated methods to model the observed parameter distribution, but the match is always approximate.

Standard Errors of Measures

It is impossible to obtain point-estimates of Rasch parameters. Every Rasch measure is to some degree imprecise. This imprecision is usually reported as a standard error. For MMLE, it may be reported as a series of plausible values, intended to report a more complex error distribution, but, for practical purposes, even these can be summarized by a mean (corresponding to the estimate) and a standard deviation (corresponding to the standard error).

The algorithm to compute the standard error is derived from the properties of the estimates or is a by-product of the estimation method. The pairwise standard error is less well-defined than those of the other estimation methods because of the data-dependent reuse of observations in estimating observations. Correcting for the degree of data reuse results in serviceable standard errors.

All estimation methods produce estimates with standard errors of about the same size, because they are obtained from data containing the same information. In general, the more observations in which a parameter participates, the smaller the standard error of its estimate. The information in an individual observation is most influenced by the targeting of the parameters that generated the observation and the number of categories in the relevant rating scale. Covariance in the data reduces precision and so inflates the standard errors, but rarely to the extent that it would lead the analyst to a substantively different conclusion about the quality of the measures.

Regardless of the estimation method, there are four conventional ways of reporting Rasch standard errors (Wright, 1995). Standard errors can either be local or general. They can also be ideal or real. Local standard errors are computed relative to the estimate of some particular item on the test (usually the first one). This reference item has no standard error. Choice of a different reference item changes all the standard errors. This makes the standard errors difficult to interpret and awkward to transport to other contexts.

General standard errors are computed as though all other parameters are known, i.e., as though their estimates are point-estimates. Converting from general to local standard errors is merely a matter of choosing a reference item, and then computing joint standard errors between that reference item and all other items. The general standard errors have the virtue that they are easy to interpret and transport to other contexts.

Ideal standard errors are reflect the highest possible precision obtainable with data like those observed. These "best case" values are the smallest possible, estimated on the basis that the data fit the Rasch model. Any idiosyncracies in the data are regarded merely as evidence of the stochastic nature of the model. These "model" standard errors produce the highest possible estimates of test reliability.

Real standard errors reflect the most imprecision. These "worst case" values are obtained on the basis that all idiosyncracies in the data are contradictions to the Rasch model. These values will produce the lowest reasonable estimates of test reliability. As misfit in the data is brought under control, the real standard error approaches the ideal.

Implementations of the Estimation Methods

Rasch estimation methods are rarely implemented directly by the analyst, except perhaps for the estimation of person measures when item difficulties are known (Linacre, 1996, 1998). Instead, analysts rely on available computer programs.

To illustrate the similarities between the estimates obtained by different estimation approaches, five computer programs were employed. RUMM (Andrich et al., 1997) implements pairwise estimation. Quest (Adams & Toon, 1994) and Winsteps (Wright & Linacre, 1991) implement JMLE. ConQuest (Wu et al., 1998) implements MMLE. Lpcm-Win (Fischer, 1998) implements CMLE.

Though the intention was to analyze the same data set, representative of actual clinical data, with all 5 programs, this proved impossible with the versions of the programs available to the author. Instead, two data sets were used. One data set comprised 16 items and 156 persons. The items were polytomous with up to 4 categories. The data set included extreme scores and missing data. It was provided as a sample data set with the RUMM program. Measures were estimated from this data set with ConQuest, Quest, RUMM and Winsteps. A second data set was constructed from this data set. It comprised 15 items and 156 persons. There were no extreme scores nor missing data.

Measures were estimated from this data set with Lpcm-Win, ConQuest and Winsteps.

Each computer program was instructed to produce estimates in accordance with the Rasch partial credit model, but using the program's own default settings, as far as possible. Every estimation process was continued to convergence. Item, rating scale and person estimates were produced, to the extent each program allowed.

On inspection of program output, it was seen that item difficulties and rating scale (partial credit) estimates were reported in such different ways that simple comparison was not possible. It also emerged that there were two ways of reporting person measures, either case-by-case or for all possible non-extreme scores. The information provided by these two ways is combined for this discussion. Since most programs did not attempt to estimate measures corresponding to extreme scores, these are not considered here.

Figure 1 about here

Figure 1 depicts the person measures produced by four of the programs on the first data set. Though the programs themselves adopt different criteria for establishing the local origin of the measurement scale, all measures are equated to a common local origin in the Figure. Winsteps was run in its default mode which does not attempt to correct for JMLE estimation bias. This bias causes its estimates (represented by the diagonal) to be slightly wider (less central) than those of the other programs. It appears that Quest, also using JMLE, is correcting for estimation bias. The standard errors of the measures in this plot are .4 logits. All four programs, and so all four estimation methods, are producing substantively and statistically the same measures.

Figure 2 about here

Figure 2 plots person measures estimated from the second data set. At the lower end, the estimates coincide. For these estimates, standard errors are again 0.4 logits. At the upper end, differences are seen. Winsteps produced JMLE estimates without correction for estimation bias, represented by the diagonal line, the highest estimates. The correction for estimation bias, (Test length - 1)/ (Test length), was applied to the Winsteps measures. These are plotted as asterisks, which appear as the next lower (more central) estimates. The Lpcm-Win (CMLE) measures are next, plotted as X. The ConQuest (MMLE) measures are the most central, plotted as +. The range of estimates of the most extreme person in the top right of Figure 2 is .7 logits, but here the standard error is 1.0 logits. Again the estimates are statistically identical. Confusion might result, however, if measures from one program were interspersed with those from another.

Conclusion

Each Rasch estimation method has its strong points and its advocates in the professional community. Each also has its shortcomings. In practice, all methods produce statistically equivalent estimates, though precautions need to be taken when estimates produced by different computer programs or estimation methods are to be placed in one frame of reference. Preparation of this paper reinforced the perception that the structure of the data to be analyzed and the nature of the information required from the output of the analysis are primary in the selection of a Rasch estimation computer program. The theoretical properties of the estimates have been shown to be a minor consideration.

Acknowledgement

Andrew Stephanou of the Australian Council for Educational Research provided valuable computational assistance.

References

Adams, R. J., and Toon, K.S. (1994). *Quest: The Interactive Test Analysis System.* Melbourne, Australia: Australian Council for Educational Research.

Adams, R. J., Wilson, M. R., and Wu, M. L. (1997). Multilevel item response models: an approach to errors in variable regression. *Journal of Educational and Behavioral Statistics, 22* (1), 46-75.

Andrich, D. A., Lyne, A., Sheridan, B., Luo, G. (1997). *RUMM: Rasch Unidimensional Measurement Models.* Perth, Australia: RUMM Laboratory.

Berkson, J. (1944). Applications of the logistic function to bio-assay. *Journal of the American Statistical Society 39,* 357-365

Bernoulli, J. (1713). *Ars Conjectandi. Part 4.* Basel. Excerpted in *Rasch Measurement Transactions, 12* (1), 625. 1998.

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics 19* (3), 293-5.

Cohen, L. (1979). Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology 32* (1), 13-120.

Fischer, G. H. (1995). Derivations of the Rasch model. Chapter 2 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Fischer, G. H. (1998). *Lpcm-Win.* Minneapolis, Minnesota: Assessment Systems Corp.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Proceedings of the Royal Society 222,* 309-368.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

Kelderman, H., and Steen, R. (1988). *LOGIMO computer program for log-linear item response theory modelling.* Twente, the Netherlands: University of Twente.

Laudan, L. (1977). *Progress and its Problems*. Berkeley, CA: University of California Press.

Linacre, J. M. (1984). Paired comparisons with standard Rasch software. *Rasch Measurement Transactions 13* (1), 584-5.

Linacre, J. M. (1989). *Many-facet Rasch Measurement.* Chicago: MESA Press.

Linacre, J. M. (1994). PROX with missing data. *Rasch Measurement Transactions 8* (3), 378.

Linacre, J. M. (1995). PROX for polytomous data. *Rasch Measurement Transactions 8* (4), 400-401.

Linacre, J. M. (1996). Estimating measures with known item difficulties. *Rasch Measurement Transactions 10* (2), 499.

~~Linacre, J. M. (1997). The normal cumulative distribution and the logistic ogive. *Rasch*~~ *Measurement Transactions 11* (2), 569.

Linacre, J. M. (1998). Estimating measures with known polytomous item difficulties. *Rasch Measurement Transactions 12* (2), 638.

Rasch, G. (1960) Probabilistic Models for Some Intelligence and Attainment Tests. Chicago: University of Chicago Press. Reprinted, 1992. Chicago: MESA Press.

Verhelst, N. D., and Glas, C. A. W. (1995). The one parameter logistic model. Chapter 12 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction: comment on Jansen, Van den Wollenberg, and Wierda. *Applied Psychological Measurement 12*, 315-318.

Wright, B. D. (1995). Which standard error? *Rasch Measurement Transactions 9* (2), 436-7.

Wright, B. D., (1998a). Estimating measures for extreme scores. *Rasch Measurement Transactions 12* (2), 632-633.

Wright, B. D. (1998b). Two-item testing. *Rasch Measurement Transactions 12* (2), 627-8.

Wright, B. D., and Linacre, J. M. (1991). *Winsteps Rasch Measurement Computer Program.* Chicago: MESA Press.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B. D., and Stone, M. H. (1979). *Best Test Design.* Chicago: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R. (1998). *ConQuest: Generalised Item Response Modelling Software.* Melbourne, Australia: Australian Council for Educational Research.

Yule, G. U. (1925). The growth of population and the factors which control it. Presidential address. *Journal of the Royal Statistical Society 88,* 1-62.

Table 1. Counts on a Two-Item Test

		Item B		Totals:
		Right: 1	Wrong: 0	
Item A	Right: 1	S_{11}	S_{10}	T_{A1}
	Wrong: 0	S_{01}	S_{00}	T_{A0}
Totals:		T_{B1}	T_{B0}	T

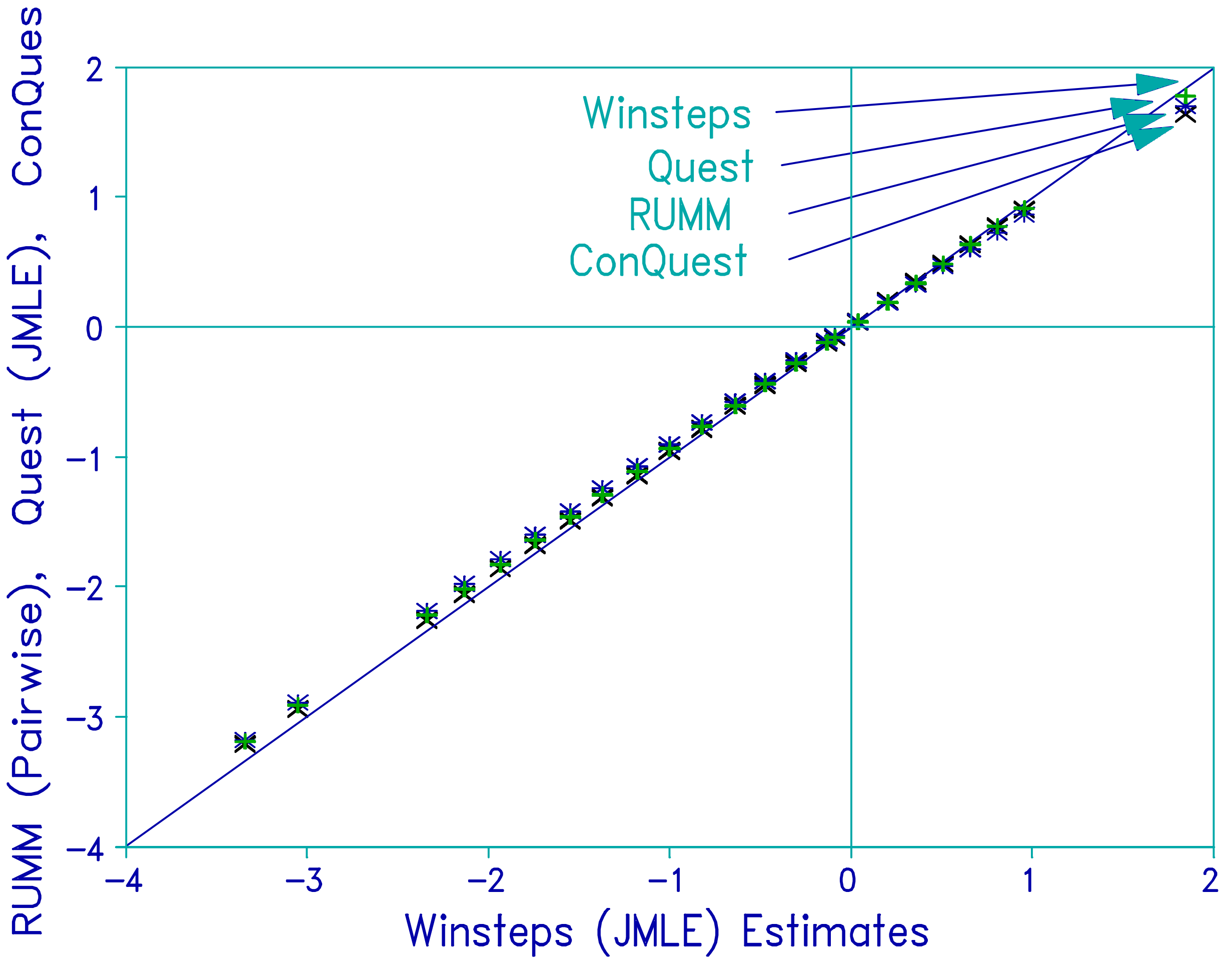

Figure 1. Person estimates from ConQuest, Quest, RUMM and Winsteps

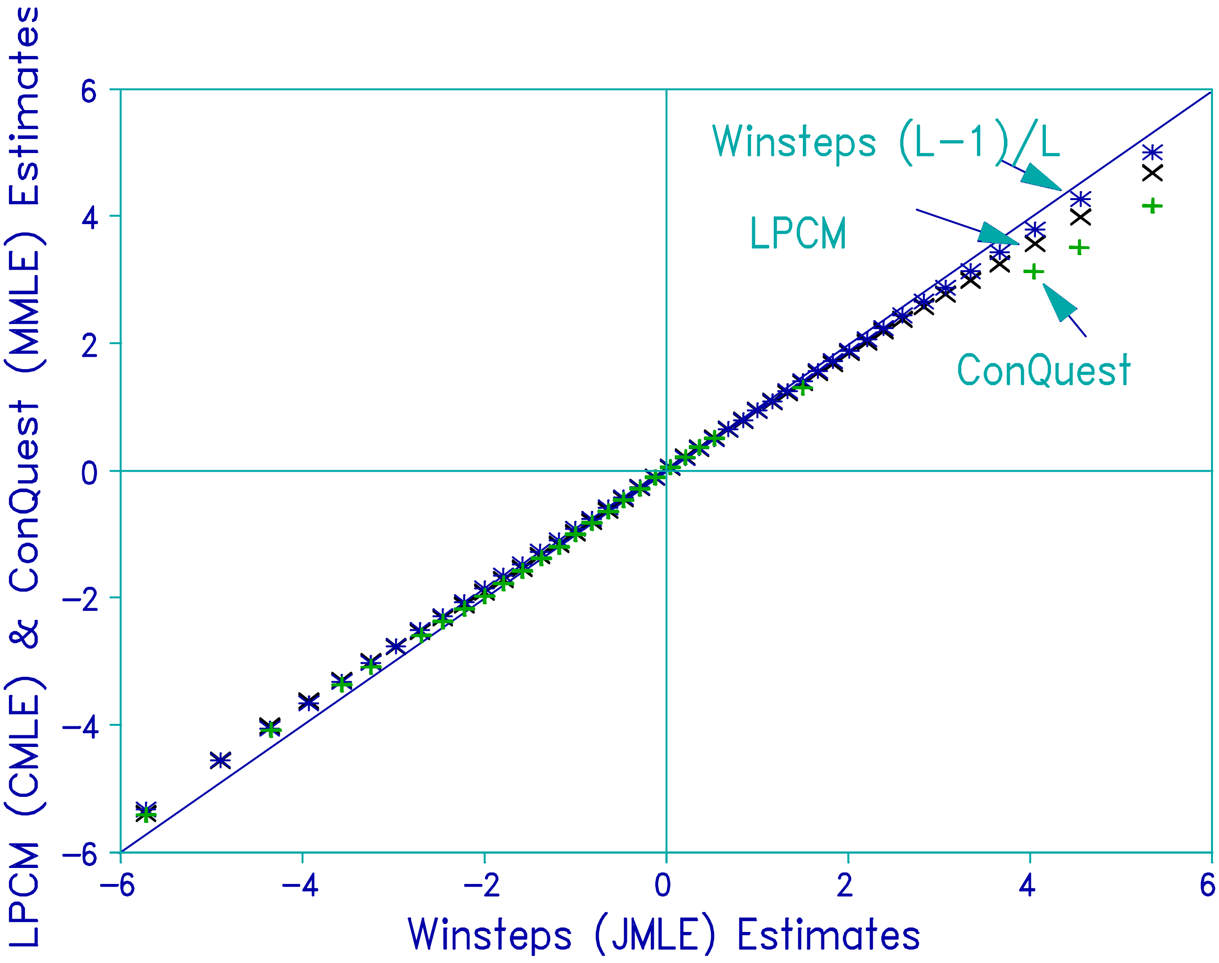

Figure 2. Person estimates from ConQuest, Lpcm-Win Winsteps