JOURNAL OF OUTCOME MEASUREMENT®, 3(4), 382–405

Understanding Rasch Measurement: Estimation Methods for Rasch Measures

John M. Linacre
University of Chicago

Rasch parameter estimation methods can be classified as non-interative and iterative. Non-iterative methods include the normal approximation algorithm (PROX) for complete dichotomous data. Iterative methods fall into 3 types. Datum-by-datum methods include Gaussian least-squares, minimum chi-square, and the pairwise (PAIR) method. Marginal methods without distributional assumptions include conditional maximum-likelihood estimation (CMLE), joint maximum-likelihood estimation (JMLE) and log-linear approaches. Marginal methods with distributional assumptions include marginal maximum-likelihood estimation (MMLE) and the normal approximation algorighm (PROX) for missing data. Estimates from all methods are characterized by standard errors and quality-control fit statistics. Standard errors can be local (defined relative to the measure of a particular item) or general (defined ralative to the abstract origin of the scale). They can also be ideal (as though the data fit the model) or inflated by the misfit to the model present in the data. Five computer programs, implementing different estimation methods, produce statistically equivalent estimates. Nevertheless, comparing estimates from different programs requires care.

Requests for reprints should be sent to John M. Linacre, MESA Psychometric Laboratory, University of Chicago, 5835 S. Kimbark Avenue, Chicago, IL 60637, e-mail: mesa@uchicago.edu.

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 ordinal data. The analyst, however, is rarely concerned about the estimation process, provided that reasonable values for the measures are obtained. The precision of the measures can be characterized by their standard errors, and their statistical validity by fit statistics. An appreciation of the different methods of estimation, however, enables the analyst to better evaluate what constitute reasonable measures.

When a new measurement situation is encountered, the Rasch measures, the parameters of a relevant Rasch model, must be inferred from data. This is accomplished by means of the method of inverse probability, first described by Jacob Bernoulli (1713). Inverse probability enables us to estimate values for the measures, but those values are always approximate to some degree. The fact that all measures (including Rasch measures) are approximate, is rarely of major concern, because "for problem solving purposes, we do not require an exact, but only an approximate, resemblance between theoretical results and experimental ones" (Laudan, 1977).

Estimation, Precision and Accuracy

Rasch estimates are always characterized by their precision and their accuracy. In this context, precision relates to the uncertainty in the measure, the estimated location of the parameter on the latent variable, when it is specified that the data fit the Rasch model. This precision is reported as the standard error of the measure. When the data are specified to fit a Rasch model, then all unexpectedness in the data are deemed to be products of the probabilistic processes required by Rasch models.

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. The degree of departure is summarized in fit statistics and other indica-

tors of conformity of the data to the Rasch model. No empirical data set fits the Rasch model perfectly. 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.

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 point, is a "pass" or a "fail".

The Nature of the Rasch model

Consider the basic Rasch model. This postulates that the data are the dichotomous outcomes of a probabilistic process governed by a linear combination of parameters, called here the person ability and the item difficulty. All estimation methods in common use for other Rasch models can also be applied to the basic model. This model is:

$$\log\left(\frac{P_{ni1}}{P_{ni0}}\right) \equiv B_n - D_i, \qquad (1)$$

where

B_n is the ability of subject n, where $n = 1,N$,
D_i is the difficulty of item i, where $i = 1,L$,
P_{ni1} is the probability that subject n will succeed on item i,
P_{ni0} is the probability of failure $1-P_{ni1}$.

P_{ni1} can be expressed as:

$$P_{ni1} = \frac{e^{(B_n - D_i)}}{1+e^{(B_n - D_i)}}. \quad (2)$$

Were the model parameters to be known, then the probability of observing any particular datum would also be known. Each data point, X_{ni}, has a value of x, which is 1 if person n succeeds on item i, and 0 otherwise. The expected value of the datum, E_{ni}, is P_{ni1}. For any parameter, e.g., n (or i), the marginal score, R_n (or R_i), the sum of all observations modeled to be generated by n (or i), is

$$R_n = \sum_n X_{ni} \approx \sum_n E_{ni} = \sum_n P_{ni1}. \quad (3)$$

Thus the frequency of successful responses in the data becomes the basis for inferring the probabilities of success, and so supports the estimation of Rasch measures.

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

$$L = \prod_{n,i} P_{nix}. \quad (4)$$

Non-iterative Estimation Methods

Rasch measures are additive, and so linear. Ordinal data are of unknown linearity. This means that Rasch estimates are non-linear transformations of data. Usually, estimation with non-linear functions requires an iterative approach, in which initial rough estimates are systematically improved until final estimates are obtained. There are, however, two estimation methods which do not require iteration.

Graphical methods

When all items are of equal difficulty, D, then B_n can be estimated in closed form:

$$B_n = \log\left(\frac{P_{i1}}{P_{i0}}\right) + D \approx \log\left(\frac{R_n}{L - R_n}\right) + D. \quad (5)$$

Similarly when all persons are of equal ability, B, then D_i can also be estimated in closed form:

$$D_i = B - \log\left(\frac{P_{i1}}{P_{i0}}\right) \approx B - \log\left(\frac{R_i}{N - R_i}\right). \quad (6)$$

Of course, both these conditions cannot hold simultaneously. Nevertheless, for rough approximations when precise measures are not required, these logistic transformations of raw scores provide the basis for a simple graphical method.

Georg Rasch (1960, Ch. VI) demonstrates how plotting logistic transformations of success frequencies permits the drawing of trace lines by eye. The persons are stratified by raw score, R, on the complete test. Each raw score is converted into a logit, log(R/(L-R)). For each score group, their percent success on each item is computed and converted into its logit value, log(success%/failure%). For each item, the success logits are plotted against the score logits.

Figure 1 shows the empirical jagged success-logits for each item and person score-group, together with the inferred parallel straight Rasch

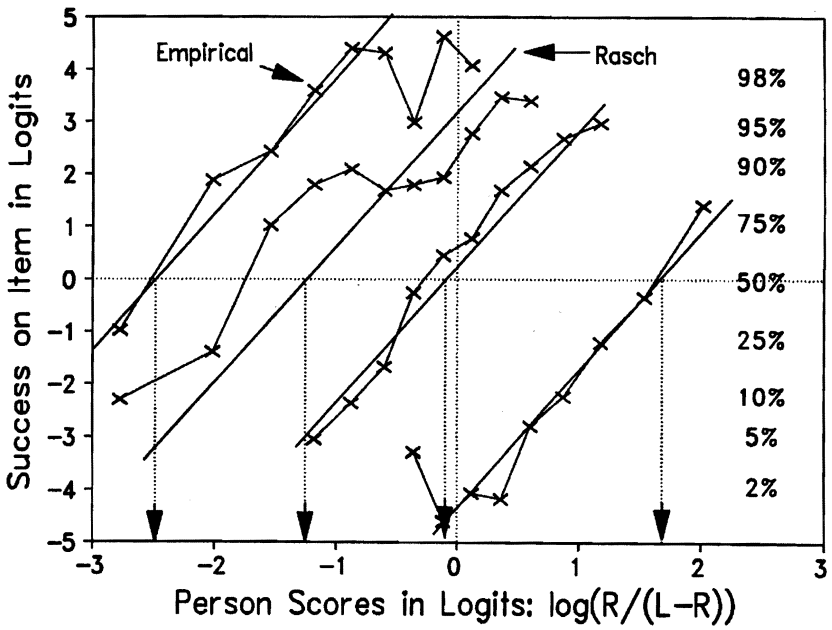

Figure 1. Graphical estimation of the difficulty measures of four items.

lines, drawn by eye. This Figure reports 4 items from G. Rasch's BPP test results. In the plot, the difficulty of an item is equal to the ability of people whose probability of success is 0.5. Though these estimates are somewhat compressed and distorted, they are not misleading as to the hierarchy and relative placement of persons and items on the latent variable. The plot also supports an investigation into measure accuracy in terms of the fit of the data to the Rasch model. Whenever empirical raw-score-based item characteristic curves are available, logistic transformation of both axes yields plots equivalent to those produced by G. Rasch.

Non-iterative Normal Approximation estimation (PROX)

The graphical method failed to allow for differences among person abilities and item difficulties. Leslie Cohen (1979) deduced a non-iterative method, PROX, for estimating Rasch measures, when the data are complete and both items and persons are approximately normally distributed. A procedure for performing PROX by hand is given in Wright and Stone (1979, Chap.2). Even when the distributional assumptions are not met, PROX provides useful starting values for the other estimation methods.

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. The relationship between the ogives is specified as:

$$\Psi^{-1}(y) \approx 1.7\Phi^{-1}(y), \tag{7}$$

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 (1997a) 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:

$$B_n = \sum_{i=1}^{L} D_i + X_D \log\left(\frac{R_n}{L - R_n}\right), and \tag{8}$$

$$D_i = C_D - X_B \log\left(\frac{R_i}{N - R_i}\right). \tag{9}$$

Since, by convention, $\Sigma D_i = 0$ establishes the local origin of the measurement scale, C_D is chosen so that $\Sigma D_i = 0$.

X_D and X_B are obtained from S_D and S_B, the population raw-score-based standard deviations. S_D the standard deviation of success on the test items is given by

$$S_D = S.D.\left(\log\left(\frac{R_i}{L - R_i}\right)\right), where\,(i = 1, L). \tag{10}$$

S_B the standard deviation of the success of the person sample is given by

$$S_B = S.D.\left(\log\left(\frac{R_n}{L - R_n}\right)\right), where\,(n = 1, N). \tag{11}$$

X_D adjusts for test width. The wider the spread of success on test items, S_D, the wider the spread of persons being measured, and so the bigger the measure difference between a low scoring and a high scoring person.

$$X_D = \left(\frac{1 + S_D / 2.89}{1 - S_D S_B / 8.35}\right)^{\frac{1}{2}} \tag{12}$$

X_B adjusts for sample spread. The wider the spread of success by the sample, S_B, the wider the range of item difficulties.

$$X_B = \left(\frac{1 + S_B / 2.89}{1 - S_D S_B / 8.35}\right)^{\frac{1}{2}} \tag{13}$$

Iterative Estimation Methods

Iterative estimation methods adopt initial rough starting values, e.g, zeroes, for the estimates. These estimates are used to obtain expected values for the data. A comparison is made between what was observed and what is expected. Then better estimates are produced which minimize discrepancies. This process is repeated, i.e., iterated, until the discrepancies are deemed inconsequential. At this point, the estimation process has converged.

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, "true", 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 out-of-range parameter values and so are inestimable.

Table 1

Iterative Estimation Methods

Type	Acronym	Name	Shortcomings	Software
Datum-by-datum		Gaussian Least-Squares	Many measures per score	
		Minimum Chi-Square	Many measures per score	
	PAIR	Pairwise	Many measures per score Correctable standard errors	RUMM
Marginal without distributional assumptions	CMLE CON FCON	Conditional Maximum Likelihood Estimation	Missing data intolerant Limited analysis size	Lpcm-Win
	JMLE UCON	Joint Maximum Likelihood Estimation	Correctable statistical bias	Facets Quest Winsteps
		Log-linear	Missing data intolerant	LOGIMO
Marginal with distributional assumptions	MMLE	Marginal Maximum Likelihood Estimation	Distributional mismatches	ConQuest (IRT)
	PROX	Normal Approximation Estimation	Distributional mismatches	Winsteps
		Items Two-at-a-time	Only for desperate situations	

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.

Iterative 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 of their methods are generally applicable. These methods are generally robust against missing data.

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. The function to be minimized, F, is:

$$F = \sum_{Data} (X_{ni} - E_{ni})^2. \qquad (14)$$

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

$$\frac{dF}{dB_n} = \sum_{n} 2(E_{ni} - X_{ni})V_{ni} = 0, \qquad (15)$$

where V_{ni} the model variance of an observed rating about its expectation, is

$$V_{ni} = (1 - E_{ni})^2 P_{ni1} + (0 - E_{ni})^2 P_{ni0} = P_{ni1} P_{ni0}. \qquad (16)$$

From the Rasch measurement perspective, a drawback to this and all other datum-by-datum methods is that different response strings with the same total raw score produce different measures.

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 influ-

ential 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:

$$F = \sum_{Data} \frac{(X_{ni} - E_{ni})^2}{V_{ni}}. \tag{17}$$

Pairwise estimation (PAIR). 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, n and m, respond to the same items. C_{10} is the number of times that person n succeeds on items that person m fails, and *vice-versa* for C_{01}. Then, an estimate of the difference in ability between n and m is given by the paired comparison

$$B_n - B_m \approx \log\left(\frac{C_{10}}{C_{01}}\right). \tag{18}$$

Following this approach, one data set yields estimates of the relative abilities of every pair of persons. These pairs of abilities, however, are likely to be somewhat contradictory. The resolution of these contradictions is to combine the paired comparisons into a likelihood function (Wright and Masters, 1982), which is maximized when the parameter estimates simultaneously satisfy the relationship

$$\frac{dF}{dB_n} = \sum_{m \neq n}^{N} \left(C_{10} - \frac{C_{10} + C_{01}}{1 + e^{(B_m - B_n)}} \right). \tag{19}$$

Within one estimation equation the same observation may be used many times. For instance, if person n is the only person to succeed on item 1, then that success is included in every C_{10} term for person n, and so adds N-1 to the total summation. This multiple use of observations means that standard errors derived from second derivatives 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, usually the item difficulties, 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 measures from them.

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.

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 is achieved by stratifying the person sample by raw score, and them estimating item difficulties within each raw score stratum. Estimation within stratum conditions out the person measures, resulting in estimates with minimal statistical bias and well-defined standard errors.

The minimal remaining estimation bias results from the very slight probability that a sample of respondents, whose measures correspond to the estimated parameters, would all simultaneously succeed or fail on a test item. If a large sample of respondents is obtained, then this probability is effectively zero, meaning that CMLE estimates become free of bias.

Estimation is conceptually simple, but challenging to implement. First, a set of rough starting values for the item difficulties is imputed. Then the likelihood of every possible response string that generates a particular score, r, is estimated. In this computation, a reference person of any computationally convenient ability can be used. All these likelihoods are summed for the particular score, r, becoming the likelihood of making that score in any way, Λ_r. The response strings for score r are then inspected for each item in turn. The likelihoods of all response strings with a success on item i are accumulated into a likelihood of observing a success on item i given a score of r, Λ_{ri}. In score stratum r, there are N_r persons. Thus the expected number of suc-

cess on item i in score stratum r is $N_r(\Lambda_{ri}/\Lambda_r)$. Finally, a revised estimate of the difficulty of item i is obtained by summing across all score strata and applying an estimation equation like

$$D_i' = D_i - \frac{R_i - \sum_{r=1}^{L-1} N_r \frac{\Lambda_{ri}}{\Lambda_r}}{(ModelVariance)}. \tag{20}$$

There is only a limited number of exponential terms that can be summed into Λ_r without loss of computational precision. This has restricted CMLE to short tests. Improvements in computer hardware and more sophisticated numerical methods have aided CMLE (Verhelst and Glas, 1995), but it is still impractical for long tests or test with many different patterns of missing data.

Joint Maximum Likelihood Estimation (JMLE). The Rasch measures for which the data are most likely to be observed are those for which the observed and expected scores coincide. Since raw scores are sufficient statistics for both items and persons measures, all measures can be estimated simultaneously. In JMLE, no parameters are conditioned out, so the method is also termed "unconditional" (UCON).

Since the marginal scores coincide with their expectations, JMLE estimates satisfy the optimal least squares criterion,

$$\left(R_n - \sum_{i=1|n}^{L} E_{ni} \right)^2 = 0. \tag{21}$$

As in all these methods, the final estimates are independent of the iterative path followed, but the usual approach follows Newton-Raphson. The estimation equation to produce a better estimate B_n' of the previous estimate B_n is:

$$B_n' = B_n + \frac{R_n - \sum_i E_{ni}}{\sum_i V_{ni}}, \tag{22}$$

where V_{ni} is defined in (16). This estimation method has proved robust against missing data, and also easily allows the incorporation into one analysis of data generated by variants of the Rasch model (dichotomous, partial credit, rating scale, Poisson, 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 (Linacre,1997b).

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.

Log-linear methods. Logit-linear (logistic) models, including Rasch models, can be reparameterized as log-linear models and applied to frequency tables. In the frequency table, there is a cell for each observed, non-extreme, pattern of responses to the items. The cell contains a count of the number of times the pattern is observed. Then, for a particular response string s with frequency F_s,

$$\log(F_s) = -\sum_s X_{si} D_i + \gamma, \qquad (23)$$

where $X_{si} = 1$ if the response to item i in string s is 1, and $X_{si} = 0$ otherwise. The item difficulties, D_i, are estimated, and γ is chosen such that $\Sigma D_i = 0$. Person abilities can be estimated by another method, after anchoring the item difficulties.

The parameters of log-linear models can be estimated with standard statistical computer programs, but these have proved of limited utility. Rasch models, in log-linear form, can have hundreds of parameters and millions of cells. These overwhelm the computational capacity of most statistical software. Further, sometimes the design of the estimation algorithm requires a cell for every possible response string. Then many of the cells will contain incidental zeroes, because particular response strings did not happen to be observed. Further, if data are missing from response strings, estimation with standard statistical software becomes virtually impossible.

Kelderman (1984) has devised an estimation approach specifically for Rasch log-linear models. This is implemented in Kelderman and Steen (1988). It can handle long tests, but is still intolerant of missing data.

Marginal Estimation with Distributional Assumptions

Distributional assumptions regarding some or all of the parameters can be usefully employed to simplify computation or even make estimation possible. If the distributional assumptions seriously mismatch the latent parameter distributions, then severe estimation bias may be introduced.

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, Guttman patterns and missing data can be easily managed. Second, when the intention is not to measure individuals, but to summarize estimates, it bypasses an analytic step. Third, it supports generalized multidimensional forms of the Rasch model (Wu, et al., 1998).

MMLE produces estimates for the discrete parameters, usually corresponding to item difficulties, such that their observed and expected marginal scores coincide, under the condition that the distribution of the other parameters has the required form. This requires a two-stage estimation approach, such as the E-M, Expectation-Maximization, algorithm (Bock and Aitken, 1981).

The MMLE method is ubiquitous in the estimation of the two- and three-parameter Item Response Theory (IRT)models. When those models are constrained to take the form of Rasch models, then Rasch MMLE estimates are obtained.

Normal Approximation Algorithm (PROX). The PROX algorithm has an iterative form which can accommodate missing data. The estimation equations are resubscripted to indicate that only those instances when person n actually responded to item i are to be considered. For instance,

$$B_n = \sum_{i \in n}^{L_n} D_i + X_{D_n} \log\left(\frac{R_n}{L_n - R_n}\right), \tag{24}$$

where ΣD_i applies only to those L_n items encountered by person n. X_{Dn} refers to the spread of those items. Linacre (1994) derives iterative PROX estimation equations for missing data. Linacre (1995) extends PROX to polytomous data.

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.

Imagine that a large sample of people have taken two dichotomous items, A and B, approximately as the Rasch model predicts. Table 2 is the tabulation of their scored responses. According to the Rasch model, the difference between the item difficulties is estimated directly by

$$D_A - D_B \approx \log\left(\frac{N_{01}}{N_{10}}\right), \quad with\ S.E. \approx \sqrt{\frac{N_{10}+N_{01}}{N_{01}N_{10}}}. \tag{25}$$

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:

$$Sample\ Mean \approx 1.864\left[\log\left(\frac{T_{A1}}{T_{A0}}\right)+\log\left(\frac{T_{B1}}{T_{B0}}\right)\right]+1.455\log\left(\frac{N_{00}}{N_{11}}\right). \tag{26}$$

Table 2

Counts on a Two-Item Test

		Item B		
		Right: 1	Wrong: 0	Totals:
Item A	Right :1	N_{11}	N_{10}	T_{a1}
	Wrong: 0	N_{01}	N_{00}	T_{A0}
Totals:		T_{B1}	T_{B0}	T

An estimator for sample standard deviation is:

$$S.D. \approx 3.763 + 1.4 * \left[\log\left(\frac{N_{11}}{T - N_{11}}\right) + \log\left(\frac{N_{00}}{T - N_{00}}\right) \right] + 0.0101 * \log\left(\frac{N_{10}}{N_{01}}\right)^2$$
$$+0.081 \left[\log\left(\frac{T_{A1}}{T_{A0}}\right)^2 + \log\left(\frac{T_{B1}}{T_{B0}}\right)^2 \right] \quad (27)$$

Estimating by Other Methods and for Other Models

Estimation methods for dichotomous data are further discussed in Molenaar (1995), and Hoijtink and Boomsma (1995), generally in a context of short tests with no missing data. Andrich (1988, Chap. 5) provides worked examples of CMLE, JMLE and PAIR in a broader context.

Most of estimation methods have been broadened to cover polytomous and other extended Rasch models. The characteristics of the estimation methods remain the same. Wright and Masters (1982) provide algorithms for CMLE, JMLE and PAIR. Andersen (1995) addresses CMLE and MMLE.

Estimating Extreme Scores

Under strict Rasch model conditions, extreme (zero and perfect) scores correspond to indefinite measures, and can take 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 reproduce the exact measures of the generating parameters, even when the data fit the Rasch model. The difference between the estimates 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 re-

sponse 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. Like all other measures, 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. Adjustment for covariance 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.

Most Rasch estimation programs report local, ideal standard errors. JMLE estimates are usually characterized with general standard errors.

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 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 data 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 used. RUMM (Andrich, et al., 1997) implements pairwise estimation. Quest (Adams and Toon, 1994) and Winsteps (Wright and 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 with no missing data. 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 2 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 0.4 logits. All four programs, and so all four estimation methods, are producing substantively and statistically the same measures.

Figure 3 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 Lpcm-Win (CMLE) measures are next most central, plotted as X. The ConQuest (MMLE) measures

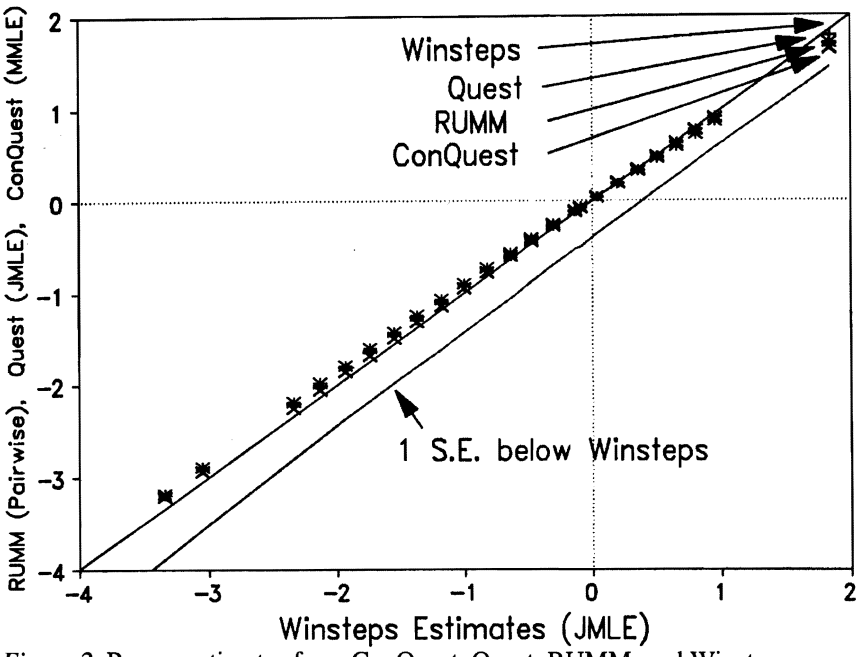

Figure 2. Person estimates from ConQuest, Quest, RUMM, and Winsteps.

are the most central, plotted as +. The range of estimates of the most extreme person in the top right of the Figure is 0.7 logits. A line one standard error below the Winsteps estimates is also plotted. Again it is seen that 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. Nevertheless, when the precision and accuracy of estimates are taken into account (Wright, 1988), all methods produce statistically equivalent estimates. Care needs to be taken, however, when estimates produced by different computer programs or estimation methods are to be compared or placed on a common measurement continuum.

Acknowledgment

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

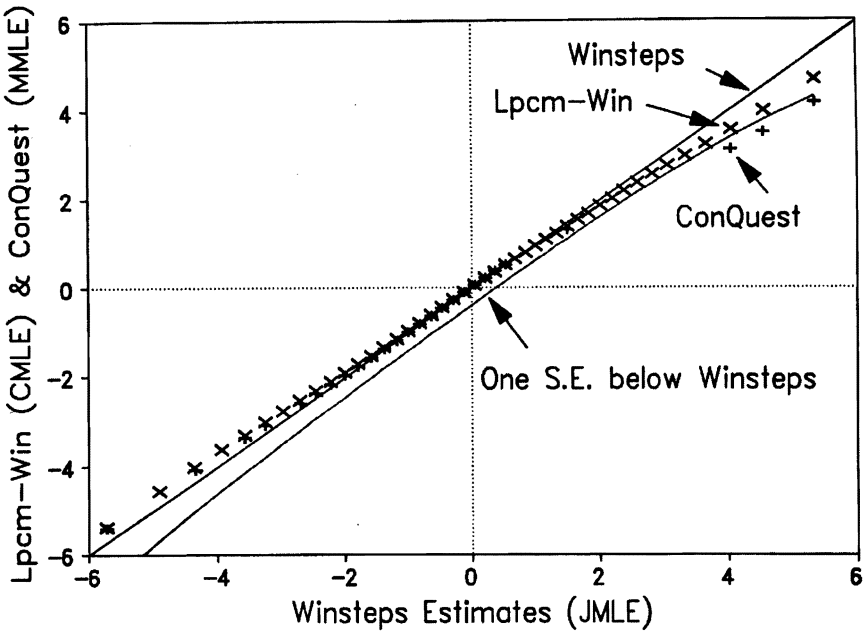

Figure 3. Person estimates from ConQuest, Lpcm-Win and Winsteps.

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.

Andersen, E. B. (1995) Polytomous Rasch models and their estimation. Chapter 15 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Andrich, D. A. (1988). Rasch Models for Measurement. Newbury Park, CA: Sage Publications.

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.

Bock, R. D. and Aitken, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46,* 443-459.

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, MN: Assessment Systems Corp.

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

Hoijtink, H., and Boomsma, A.(1995). On person parameter estimation in the dichotomous Rasch model. Chapter 4 in G. H. Fischer and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

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

Linacre , J. M. (1997b) Paired comparisons with standard Rasch software. *Rasch Measurement Transactions, 11*(3), 584-5.

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

Molenaar, I. W. (1995). Estimation of item parameters. Chapter 3 in G. H. Fischer

and I. W. Molenaar, *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

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, 92)*, 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.