JOURNAL OF APPLIED MEASUREMENT, *5*(1), 95–110

Rasch Model Estimation: Further Topics

John M. Linacre
University of the Sunshine Coast
Australia

Building on Wright and Masters (1982), several Rasch estimation methods are briefly described, including Marginal Maximum Likelihood Estimation (MMLE) and minimum chi-square methods. General attributes of Rasch estimation algorithms are discussed, including the handling of missing data, precision and accuracy, estimate consistency, bias and symmetry. Reasons for, and the implications of, measure misestimation are explained, including the effect of loose convergence criteria, and failure of Newton-Raphson iteration to converge. Alternative parameterizations of rating scales broaden the scope of Rasch measurement methodology.

Requests for reprints should be sent to John M. Linacre, P. O. Box 811322, Chicago, IL 60681-1322, e-mail: ML@winsteps.com

Introduction

Rasch models have the algebraic form of logit-linear models. Techniques for estimating the parameter values of such models have a long history in modern statistics (Yule, 1925). Several of the estimation methods currently in wide use are described in Wright and Masters (1982). This paper documents some of the other estimation methods now in use, and some current variants of earlier methods. Some statistical properties of all estimation methods are also discussed.

The estimation methods described in Wright and Master (1982) are the "Normal Approximation Algorithm" (PROX), "Pairwise Conditional Estimation" (PAIR), "Unconditional Maximum Likelihood Estimation" (UCON), also called "Joint Maximum Likelihood Estimation" (JMLE), and "Fully Conditional Estimation" (FCON, CON), also called "Conditional Maximum Likelihood Estimation" (CMLE).

Additional methods to be described here include "Marginal Maximum Likelihood Estimation" (MMLE), "Extra-Conditional Maximum Likelihood" (XMLE), "Minimum Chi-Square Estimation", and loglinear estimation.

Refinements to estimation procedures include alternative rating scale parameterizations, and alternatives to Newton-Raphson iteration.

Statistical properties to be examined include consistency, bias and the effects of misestimation.

A Rasch Rating Scale Model

Following the notation in Wright and Masters (1982), a Rating Scale model (Andrich, 1978) which defines the probability, π_{nix}, of person *n* of ability β_n on the latent variable being observed in category *x* of item *i* with difficulty δ_i as:

$$\pi_{nix} = \frac{e^{\sum\limits_{j=0}^{x}\left[\beta_n - (\delta_i + \tau_j)\right]}}{\sum\limits_{k=0}^{m} e^{\sum\limits_{j=0}^{k}\left[\beta_n - (\delta_i + \tau_j)\right]}} \qquad (1)$$

where the categories are ordered from 0 to *m*, and $\{\tau_j\}$ are the rating scale structure parameters ("step difficulties", "Rasch thresholds"). τ_0 is introduced for mathematical symmetry. Algebraically, it cancels out. It is usually set at 0. But, choosing a large negative value for τ_0 enables the computation of probabilities for a very wide range of ability-difficulty differences. These would otherwise cause exponential overflow in a computer's math processor. Computers are generally accommodating of floating-point underflow, treating it as equivalent to an arithmetical zero, but they report computational failure on floating-point overflow.

Sufficient statistics for the person and item parameters are the sums of the scored observations, i.e., the raw scores, associated with those parameters. For instance, R_n is the raw score associated with β_n, and S_i with δ_i. For each of the scale parameters, τ_j, j=1, *m*, a sufficient statistic is the count of observations in the associated category. Ronald Fisher (1922) writes of a sufficient statistic "that the statistic chosen should summarize the whole of the relevant information supplied by the sample." A Rasch model goes further, specifying that the irrelevant information be random noise with a certain distribution. Quality-control fit statistics report the extent to which data meet this specification.

Equation (1) is more conveniently expressed as:

$$\log\left(\frac{\pi_{nix}}{\pi_{ni(x-1)}}\right) = \beta_n - (\delta_i + \tau_x) \qquad (2)$$

Here, $\pi_{ni(x-1)}$ is the probability of person *n* being observed in category *x-1* of item *i*. This expression of the Rasch model emphasizes that the probabilistic, log-odds, structure of the data, on the left, is modeled to be the manifestation of an additive combination of latent parameters on the right.

Estimation Methods

This section discusses estimation methods omitted from Wright and Masters (1982). It largely overlaps material in Fischer and Molenaar (1995) and Linacre (1989) to which the reader is referred for a more technical exposition.

Linacre (1999) compared current implementations of several Rasch estimation algorithms, and concluded that, for practical purposes, "all methods produce statistically equivalent estimates" (p. 402). Rasch measurement has not yet progressed to the point that the slight differences between the estimates produced by different algorithms have any systematic substantive impact. Of course they can have accidental consequences. For instance, a barely statistically significant difference between two measures, might be computed to be "just significant" when the measures are estimated using one method, but "just falling short of significance" when a different estimation method is employed. This merely indicates the insecure nature of hairline decisions, whether of significance or of pass-fail decisions relative to a criterion point.

Novel estimation methods continue to be proposed, each with its own particular virtues. For instance, Karabatsos (2001) proposes a nonparametric method, perhaps immune to scaling distortions which Nickerson and McClelland (1984) perceive to be undetectable by numerical methods. Such methods have yet to reach production software, so their substantive impact is not yet known. The Rasch analyst, however, should continue to exercise caution with respect to claims such as "we suggest that our [Rasch estimation] method is superior to all others currently available." (Sheng and Carrière, 2002).

Marginal Maximum Likelihood Estimation (MMLE)

MMLE is implemented in Item Response Theory (IRT) software, such as BILOG (Mislevy and Bock, 1996), and Rasch-specific software, such as ConQuest (Wu, Adams and Wilson, 1998). Its advantage is that parameter estimates for very large samples and very long tests can be obtained. Its disadvantage is that assertions must be made about the sample distribution. A unique feature is that this sample distribution is usually modeled to include persons with extreme (zero and perfect) scores.

Under MMLE, the sample distribution is imagined to conform to some convenient mathematical fiction. A frequent choice is the normal distribution. Then the $\{\beta_n\}$ can be replaced by a normal distribution, parameterized with θ, of mean μ and standard deviation σ, i.e., of form $N(\mu,\sigma)$.

Equation (1) then becomes, for any person n in the sample,

$$\pi_{nix} = \int_{-\infty}^{+\infty} \frac{e^{\sum_{j=0}^{x}\left[\theta - (\delta_i + \tau_j)\right]}}{\sum_{k=0}^{m} e^{\sum_{j=0}^{k}\left[\theta - (\delta_i + \tau_j)\right]}} f(\theta)\, d\theta \qquad (3)$$

where

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\theta-\mu)^2}{2\sigma^2}} \qquad (4)$$

Numerical estimates can be obtained by Newton-Raphson iteration or other techniques, see Fischer and Molenaar (1995). A constraint must be introduced to make the estimates unique. In ConQuest, the constraint is that the mean of the item difficulties is zero. In many IRT programs, the mean of the person distribution, μ, is set to zero.

The PROX algorithm follows the distributional train of thought even further, applying normal distributions to both persons and items. It takes advantage of an approximate identity between the Cumulative Normal Φ and Logistic Ψ distributions, which is (Camilli, 1994)

$$\int_{-\infty}^{+\infty} f(\theta)\, d\theta = \Psi(x) \approx \Phi(x / 1.702) \qquad (5)$$

In fact, Joseph Berkson (1944), who coined the term "logit", proclaims that this approximation is indeed good enough for practical work. It greatly simplifies and speeds computation, because the integral now has a closed form solution. This usually removes the need for integration by numerical quadrature, which itself is a source of estimation error. In view of the fact that the asserted person distribution is, at best, only an approximate match to the empirical distribution, it is surprising that MMLE software does not routinely take advantage of this equivalence.

As usually implemented, MMLE provides item and rating scale structure parameter estimates, but only summary estimates of the person measures. Individual person measures can be obtained by estimating the measure corresponding to each raw score given the item and structure parameters, as in the JMLE method (Wright and Masters, 1982, p. 77). But this will not summarize to the same person measure distribution as the MMLE estimates. In particular, under the JMLE method, extreme scores do not yield estimable measures.

Another approach is to consider the raw score of person n, R_n, and to compute the probability, $P_{n\theta}$ that the person had a particular ability, θ. This computation is performed over the entire MMLE ability distribution. The MMLE person ability is

$$\theta_n = \int_{-\infty}^{+\infty} \theta \pi_{\theta|R_n} d\theta \qquad (6)$$

Estimates like these will necessarily summarize to the asserted MMLE distribution, no matter what it is. Accordingly it is essential that they be characterized with fit statistics if any attempt is to be made validate the reasonableness of the asserted distribution.

MMLE, and its close relative, PROX, can mislead the unwary. An obvious drawback to these approaches is that they violate Rasch's precept that "it ought to be possible to compare stimuli belonging to the same class—'measuring the same thing'—independent of which particular individuals within a class considered were instrumental for the comparison." (Rasch, 1980, p. xx). Indeed on the page, Rasch has reported criticisms of "group-centered" statistics. Practical considerations, however, can override idealistic ambitions. MMLE and PROX were developed for the analysis of dichotomous educational tests in which it is reasonable to assume a unimodal, reasonably symmetric person distribution. Misestimation at the tales of the distribution, geniuses and dunces, would not lead to incorrect inferences.

In clinical situations, however, the distributions and inferences can be far different. If a typical clinical instrument, e.g., a quality of life assessment, is applied to a random sample of the adult population, then the results can be expected to be highly skewed. The crucial clinical cases are likely to be in the long lower tail. If analysis combines clinical and non-clinical samples, then the person distribution may be bimodal. Even in educational testing bimodal distributions may occur (Lee, 1991). In clinical applications, misestimation may lead to incorrect inferences about the clinical state, or the clinical improvement, of patients who are the focus of the assessment.

Accordingly, care must be taken to verify that the asserted person distribution is a reasonable match to the empirical one. This can be done by inspection of the raw score distribution, or, for analyses involving missing data, comparison of the MMLE person distribution with that produced by JMLE or PAIR

Extra-Conditional Maximum Likelihood Estimation (XMLE)

The familiar objection to the convenient JMLE method is that under certain conditions its estimates are statistically inconsistent and, for short tests or small samples, noticeably statistically biased (Haberman, 1977; Wright 1988). The JMLE bias produces estimates that are too dispersed. For a two-item dichotomous test, the JMLE item estimates are twice as dispersed as the CMLE ones (Andersen, 1973). Wright (1988) discusses a simple correction which effectively eliminates JMLE bias.

An analytic attempt to remedy the estimation-bias defect in JMLE, while maintaining flexibility of data designs, is XMLE, "extra-conditional MLE", originally XCON (Linacre, 1989), implemented in WINSTEPS (Linacre, 2002c). The source of the bias in JMLE estimation is that it acts as though it can estimate finite measures for extreme score vectors, even though it can't. XMLE adjusts for this by removing the possibility of extreme score vectors from the estimation space.

Equation (1) specifies the probability that $X_{ni}=x$ and this is the probability used in JMLE.

But the probability that X_{ni} is part of a zero score vector for person n is

$$\pi_{n\{0\}} = \prod_{i=1}^{L} \pi_{ni0} \tag{7}$$

Similarly for the perfect score vector, $\pi_{n\{m\}}$, and the extreme item score vectors $\pi_{\{0\}i}$ and $\pi_{\{m\}i}$. Accordingly, in XMLE, these four probabilities are computed for each X_{ni}, and then the probabilities used for estimation become:

$$\begin{aligned} P_{nim} &= \left(\pi_{nim} - \pi_{n\{m\}} - \pi_{i\{m\}} + \pi_{n\{m\}}\pi_{i\{0\}}\right) / \gamma \\ P_{nix} &= \pi_{nix} / \gamma \qquad x = 1, m-1 \\ P_{ni0} &= \left(\pi_{ni0} - \pi_{n\{0\}} - \pi_{i\{0\}} + \pi_{n\{0\}}\pi_{i\{0\}}\right) / \gamma \end{aligned} \tag{8}$$

where γ is a local normalizing value such that $\Sigma\, P_{nix}$, x=0, m, is 0.

JMLE estimation is then executed with these adjusted probabilities. For instance, Wright and Masters (1982, p. 76) equation 4.4.5 retains its form but with adjusted probabilities.

$$\frac{\partial \lambda}{\partial \beta_n} = R_n - \sum_{i=1}^{L} \sum_{k=0}^{m} k P_{nik} \tag{9}$$

where λ is the log-likelihood of the the data. R_n is the raw score of person n on a test of L items each with categories numbered from 0 to m.

This adjustment has the effect of reducing the probabilities of the extreme categories. Consequently the XMLE estimate corresponding to a score vector for an item or person, is more central than the JMLE estimate. Thus XMLE essentially corrects the bias and inconsistency problems, but, as always, raising other concerns which are addressed later in this chapter.

Minimum Chi-Square Estimation

This estimation method is older than any of the others listed here or in Wright and Masters (1982). It is applied to logit-linear models in Yule (1925). It produces parameter estimates which maximize the fit of the data to the model. The expectation, E_{ni}, corresponding to observation, X_{ni}, is:

$$E_{ni} = \sum_{k=0}^{m} k \pi_{nik} \tag{10}$$

The multinomial variance, W_{ni}, of the observation about its expectation is given by:

$$W_{ni} = \sum_{k=0}^{m} (k - E_{ni})^2 \pi_{nik} \tag{11}$$

From these can be obtained a set of standardized residuals, $\{Z_{ni}\}$, whose distribution approximates N(0,1):

$$Z_{ni} = \frac{X_{ni} - E_{ni}}{\sqrt{W_{ni}}} \tag{12}$$

Accumulating these, for any person n or item i, yields an approximate chi-square statistic:

$$\chi^2_n = \sum_{i=1}^{L} Z^2_{ni} = \sum_{i=1}^{L} \frac{(X_{ni} - E_{ni})^2}{W_{ni}} \quad d.f. \approx L-1 \tag{13}$$

where L is the length of the test taken by person n.

To minimize this for β_n, i.e., to find the value of β_n which produces the best local fit of the data to the model, we use:

$$\frac{\partial P_{nik}}{\partial \beta_n} = (k - E_{ni}) P_{nik}, \qquad \frac{\partial E_{ni}}{\partial \beta_n} = W_{ni} \tag{14}$$

which yields relationships similar to:

$$\frac{\partial \chi^2_n}{\partial \beta_n} = \sum_{i=1}^{L} \left[2(X_{ni} - E_{ni}) - \frac{(X_{ni} - E_{ni})^2}{W^2_{ni}} \sum_{k=0}^{m} (k - E_{ni})^3 P_{nik} \right] \tag{15}$$

for which estimates can be obtained by Newton-Raphson iteration.

The first term of (15) sums to zero when the observed person raw score matches the expected score, but the sum of the second term exists if there is any imbalance in the residuals. Figure 1 illustrates this with a 6-item dichotomous test. The item difficulties are symmetrically distributed so that a raw score of 3 on the test, places the MLE person ability estimate in the center of the items. If the respondent had succeeded on the three easier items, and failed on the three harder, this would also have been the minimum chi-square ability estimate. This respondent, however, failed on an easier item, but succeeded on the most difficult item. The chi-squares for this response string for different ability levels are plotted. The minimum chi-square, and so the ability estimate according to this method, is now at 0.33 logits, noticeably above the center of the test.

Using the minimum chi-square method, persons with different response patterns, but with the same raw score, obtain different measures characterized by different standard errors, and similarly for the items. This is usually considered unacceptable for reporting purposes, even though PAIR, as implemented in RUMM2010 (Andrich, et al., 2000), reports different estimates for dichotomous items with the same raw score, and IRT programs, such as MULTILOG (Thissen, 1991), routinely report different estimates for persons with the same raw score.

Log-Linear Estimation

Strictly speaking, this isn't a different method of estimation, but rather a different way of formulating the Rasch model. Typically, this formulation is used so that Rasch item parameters can be estimated with generally available statistical software, though LOGIMO (Kelderman and Steen, 1988) also takes advantage of this conceptualization.

The log-linear version of a Rasch model is based on cell frequencies in a contingency table. The cell identification corresponds to the person response string. Thus the probability of observing a particular response string, *S*, which sums to raw score R_s, for a person of ability β_n is given by:

$$Prob_n\left(\{S_{si}\}_{i=1,L}\right)=\prod_{i=1}^{L}\pi_{niX_{si}} \qquad (16)$$

where X_{si} is the response to item *i* in string S which consists of *L*, the test length, responses.

Then expected frequency of this cell for the sample of *N* persons is thus:

$$Freq(S)=\sum_{n=1}^{N}\prod_{i=1}^{L}\pi_{niX_{si}} \qquad (17)$$

So that, from (1),

$$\log\left(Freq(S)\right)=-\sum_{i=1}^{L}X_{si}\,\delta_i\;-\;\sum_{k=1}^{m}\;\sum_{i=1\|X_{si}\geq k}^{L}\tau_k\;+\;L_{R_s} \qquad (18)$$

where δ_i is the difficulty of item *i* and τ_k is the

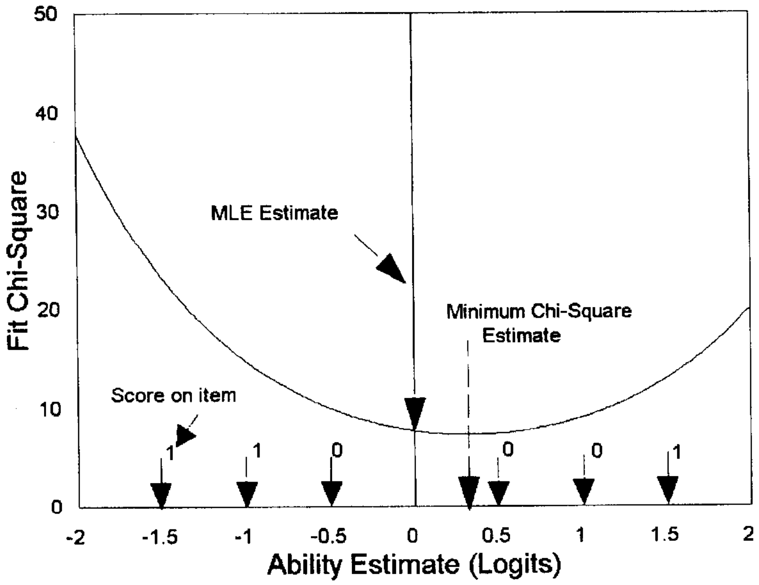

Figure 1. Minimum chi-square estimation.

Rasch threshold parameter at which categories k-1 and k are equally probably. L_{R_s} is a term dependent on the raw score, R_s, the person distribution and the item distribution, but not on the particular response string. L_{R_s} is constant across all response strings with the same raw score as S.

Item and scale structure parameters in this formulation can be estimated using standard log-linear estimation methods.

Technicalities of Estimation

Precision and Accuracy

Precision is reproducibility, i.e., the extent to which a measuring instrument agrees with itself. Accuracy is the extent to which an instrument reports the "truth". In thermometry, these are seen to be clearly different, and sometimes opposing, attributes (National Physical Laboratory, 1955). Rasch software routinely reports precision as estimate standard errors, in a manner familiar to statisticians and metrologists. End users, however, are rarely able to capitalize on this information, and tend to regard Rasch measures in the same way they have always regarded raw scores, i.e., as point estimates.

Accuracy of Rasch estimates, in terms of an external reference standard, such as a "standard meter", is not known or reported. Rasch measurement has not yet advanced that far. Internal accuracy is reported as the fit of the data to a Rasch model, but it is largely unclear what level of accuracy is needed for any particular application. In practice, however, the alternatives to Rasch measurement, be they raw scores, descriptive IRT models, or qualitative approaches, are even less secure. Consequently, once blatant inaccuracies, such as wild guessing, miskeyed items and misoriented rating scales, have been eliminated from the analysis, the resultant Rasch estimates are "accurate enough for government work."

Missing Data

For conventional analyses, complete data exist when there is a scored observation for every person on every item. Missing data can occur when items are not administered to persons, as in adaptive testing and test linking, or when persons do not respond to items. Typically, if a non-response is scored, it is keyed as "wrong" for a multiple-choice test, and as "don't know" or "neutral" for an attitude survey. But this depends on the purpose of the analysis. In this present discussion, "missing data" means data for which no scored responses are available.

In principle, missing data are of no concern to the Rasch model. Equation (1) is an observation-level model. Estimation methods which rely on sufficient statistics, such as JMLE and PAIR, merely summarize the non-missing observations that are relevant to each parameter, and compare them with their expectations. Thus, for JMLE or XMLE, Equation (9) becomes

$$\frac{\partial\lambda}{\partial\beta_n} = R_n - \sum_{i=1|X_{ni}\geq 0}^{L} \sum_{k=0}^{m} k P_{nik}$$
$$R_n = \sum_{i=1|X_{ni}\geq 0}^{L} X_{ni} \qquad (19)$$

Some estimation methods require complete data, or, a little more flexibly, data consisting of complete subsets. These methods include CMLE and log-linear approaches. Since complete data permits optimization of estimation routines, some implementations of PROX, JMLE and MMLE also require complete data. When complete data are required, either casewise deletion or imputation of missing values is performed. If only a few observations are missing in the data set, casewise deletion is indicated. When performed, imputation of missing values also requires correction of measure standard errors and fit statistics for the influence of the imputed data.

Estimate Consistency

Consistency is the property that, given an infinite amount of data which fit a statistical model, the estimation procedure would recover the values of the parameters used to generate those data. Consistency differs from accuracy. Accuracy is the extent to which the estimates meet external criteria of exactness. In Rasch analysis, these are usually quality-control fit criteria.

Consistency can never be observed directly. It can only be inferred from the mathematical properties of the estimation method. It is usually assumed to be asymptotic, so that, as the amount of data becomes large, estimates are expected to approach their generating values, but this is not necessarily true. For instance, a Monte Carlo method which guessed at random would finally guess the correct estimate, but the last guess of a large number of guesses might not be any better than the first.

Considerable effort has been expended criticizing JMLE for its lack of consistency, e.g., Jansen, et al., (1988). But consistency is a slippery property. Consistency requires not only an infinite amount of data, but data that is infinite in particular ways. Thus Haberman (1977) demonstrates that JMLE can be consistent if both the number of persons and number of items become infinite together. Estimation methods which eliminate, in some way, the individual person parameters from estimation of the item parameters, such as CMLE, PAIR and MMLE, are consistent if the number of persons becomes infinite (Pfanzagl, 1994; Zwinderman 1995). CMLE, as usually implemented, is not consistent if the number of persons remains finite, but the number of items becomes infinite.

The usual reason for inconsistency is the possibility of extreme score vectors (zero and perfect raw scores). These imply that the corresponding parameter is infinite or undefined. How able is a person who succeeds on every item? How easy is an item on which everyone succeeds? Consistency requires that, as the amount of data become infinite, the probability of an extreme score vector in the estimation space becomes zero.

For CMLE, consistency occurs in two stages. Extreme person vectors are explicitly excluded from the estimation space. Only non-extreme person score vectors contribute to the likelihood of the data, which is to be maximized. The likelihood of extreme item vectors is reduced to zero by expanding the person sample to infinity, so that it becomes certain that there is at least one success and one failure on every item.

Consistency is a theoretical property. Of practical concern is the extent to which estimates based on finite data differ from their true values. This is termed estimation bias.

Estimate Bias

An estimation algorithm is unlikely to recover the true value of a parameter from finite data. Instead, for each parameter an estimate is reported and its precision, in the form of a standard error of that estimate around its unknown true value. Since the true value is unknown, the estimate is usually treated as the true value, and standard error is applied to the estimate. Since error distributions are not always of a simple form, measurement imprecision may be reported in different forms, e.g., as plausible values in ConQuest. Nevertheless, it is assumed that the true value is somehow central in the error distribution.

Estimation bias occurs when the estimated values are, on average, higher or lower than the true values, and perhaps even outside the reported precision. Though JMLE estimation bias has been thoroughly discussed (Wright, 1988), that of other estimation algorithms has been generally ignored. For example, CMLE is generally much less biased than JMLE for short tests, but not totally unbiased.

Consider a 2-item dichotomous test administered to 3 persons. The 64 possible data matrices are shown in Table 1. Only the 6 matrices shown in roman bold contain no extreme score vectors for either persons or items. When computing the likelihoods used in making its estimates, JMLE includes all 64 data matrices, despite the fact that 58 of them contain inestimable response strings. CMLE explicitly excludes data matrices with inestimable person response strings from its likelihood calculations. This eliminates 56 of the 64 data matrices, keeping the 6 bolded matrices. CMLE also keeps the two data matrices with inestimable item response strings which are shown in italics at the diagonal corners of Table 1. These introduce bias into the CMLE estimates. According to JMLE, the logit distance between the items is 2log(2) = 1.39 logits. According to CMLE, the distance is log(2) = 0.69 logits. In fact, the exact MLE estimate for these data is that the items are infinitely far apart!

In this boundary case, the JMLE estimates are actually more accurate than the CMLE ones, though both are infinitely wrong. XMLE estimation diverges on this analysis, which, though correct, is not helpful. Here is a case where incorrect, but finite, estimates are more useful than correct, infinite ones.

If we double the sample size to 6 persons, keeping the same success ratio, there are now 4096 possible data matrices, of which 62 are estimable. CMLE again includes the likelihood for 2 inestimable matrices and JMLE includes all 4096 possible matrices, The CMLE item estimates remain 0.69 logits apart, and the JMLE estimate remains 1.39 logits. The exact MLE estimate is now .89 logits apart. For 9 persons, the exact MLE estimate is .74 logits apart. As sample size increases, the bias in the CMLE estimates rapidly disappears. For JMLE it never does.

The most extreme example of estimation bias is that of JMLE estimates obtained from a two-item dichotomous test which we have just examined. The reported dispersion of the two items is twice that of the CMLE values. This motivated Wright and Douglas (1977) to propose an $(L\text{-}1)/L$ bias correction for short tests, which, according to Jansen, et al., (1988), largely eliminates the bias for tests of over 10 items. But, for long tests, or short tests comprising rating scales with many categories, the JMLE estimation bias also becomes negligibly small, even without explicit correction.

The small sample estimation bias for PAIR and MMLE depends on technical details of the implementation.

Estimation Symmetry

As Rasch measurement is applied in ever more diverse fields, the problem of estimate symmetry is coming more into prominence. In the educational and health care fields, it is generally obvious what are the objects of measurement, the students or patients to be measured, and what are the agents of measurement, the test items designed to perform the probing. But, even from the earliest days, there are cases where this is uncertain.

In Georg Rasch's (1969) analysis of traffic accidents, he considered the number of fatal accidents in specific short road segments over specific periods of time. With current Rasch methodology, this situation could be easily modeled using the rating scale, 0=no fatalities, 1=1 fatal accident, 2=more than one fatal accident. But what is the object of measurement and what is the agent? Is the object of the investigation to measure "danger inherent in different road segments" or "danger inherent in driving at different times of day"? The choice of what are the

Table 1

All 64 possible data matrices, each comprising 2 dichotomous items (rows)administered to 3 persons (columns)

000	000	000	000	000	000	000	***000***
000	001	010	011	100	101	110	***111***
100	100	100	**100**	100	100	100	100
000	001	010	**011**	100	101	110	111
010	010	010	010	010	**010**	010	010
000	001	010	011	100	**101**	110	111
110	**110**	110	110	110	110	110	110
000	**001**	010	011	100	101	110	111
001	001	001	001	001	001	**001**	001
000	001	010	011	100	101	**110**	111
101	101	**101**	101	101	101	101	101
000	001	**010**	011	100	101	110	111
011	011	011	011	**011**	011	011	011
000	001	010	011	**100**	101	110	111
111	111	111	111	111	111	111	111
000	001	010	011	100	101	110	111

objects of measurement and what are the agents is arbitrary.

This may seem to be an academic discussion until the data matrix is constructed, and estimation undertaken. Though Rasch models, like (1), are symmetric in the way objects and agents enter into the analysis, estimation software rarely is. Most software provides a more detailed analysis for items than for persons. More perplexing to the user, however, is the fact that transposing the data matrix can produce non-equivalent sets of estimates. JMLE and PROX estimates are symmetric. Perforce, MMLE must be asymmetric because its distributional specification is one-sided. CMLE and PAIR estimates are asymmetric, but the extent of the mismatch depends on the methods used to obtain object estimates from the agent estimates.

Newton-Raphson: A Cautionary Tale

Except for the PROX method applied to complete data (Cohen, 1979), the estimation of Rasch measures requires iteration. An initial estimate of a measure is made. The implications of this measure are compared with the data. A correction is made to the initial measure in an attempt to bring its implications into closer conformity with the data. Many commonly used iterative methods are based on the Newton-Raphson approach. These methods provide adjustments based on the sizes of discrepancies between the observed and the expected, and the local slopes of relevant mathematical functions. Newton-Raphson is particularly well suited to the Rasch model because of the mathematically well-behaved nature of the logistic ogive. But there can be problems.

If there is only one estimate to be obtained, e.g., the measure of a person on a set of items of known difficulty, then Newton-Raphson can work smoothly and quickly. In order to avoid overflow or loss of precision during computation, it is advisable to choose an initial estimate more central than the final estimate, and also to limit changes in the estimate to one logit per iteration.

Conceptually, Newton-Raphson is operating on a plane, a two-dimensional space. Under most circumstances, however, many parameters are estimated simultaneously. For instance, for a dichotomous test of *L* items, *L*-1 free parameters are usually estimated and one constraint imposed. This means that Newton-Raphson is operating in an *L*-dimensional space, but suggesting changes for each estimate in terms of a local 2-dimensional space. This can lead to the situation shown in Figure 2.

In Figure 2, an initial estimate is made. Then, based on the slope of the likelihood function, and the amount of discrepancy between the observed and expected values, a revised, second estimate is produced. The second estimate produces a re-

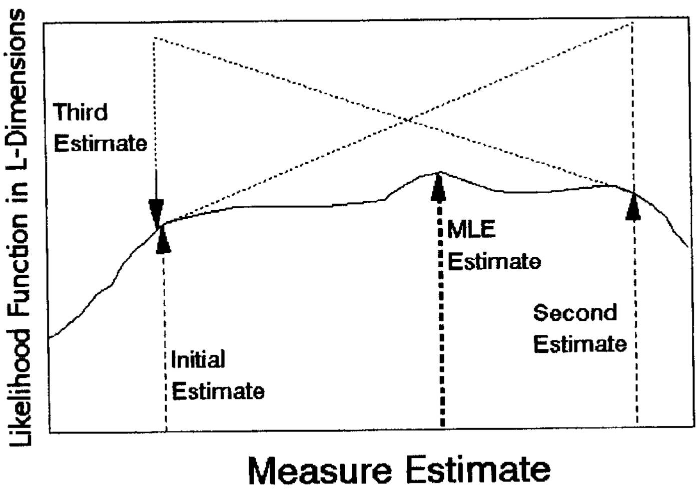

Figure 2. Effect of misestimation on observation fit.

vised slope and discrepancy, leading to a third estimate. In principle, these estimates rapidly approach the best estimate, i.e., the value of the parameter most likely to have generated the observed data.

In Figure 2, however, the Newton-Raphson approach is failing to converge on the best estimate. The slope of the mathematical function is such that the estimates are jumping from one side of the best estimate to the other. This oscillating behavior is usually an indication that the likelihood function is rather flat, so that a range of estimates are almost equivalent.

In Figure 2, it is seen that the third estimate is slightly further away from the best estimate than was the initial estimate. This type of divergent behavior has been observed with 3-PL estimation (Stocking, 1989). I have also observed this to occur with Rasch estimation of incomplete data sets, containing weakly connected subsets of data, or when there are categories of a rating scale that are rarely observed. Under these circumstances, other methods may be more robust. These methods include proportional curve-fitting (implemented in Winsteps) and Monte Carlo techniques.

Misestimation, Convergence and Fit

Paradoxically, as estimates of measures improve, their fit to the Rasch model can become worse. Consider a dichotomous test, such as the Knox Cube Test (Wright and Stone, 1979). Suppose that due to gross misestimation, every item was estimated to have the same difficulty, and every person was estimated to have an ability exactly targeted on that difficulty. Then the model expectation for each person on each item is .5, and the model binomial standard deviation of an observation about its expectation is 0.5. In fact, the observations are failure (0) or success (1). All the "observed-expected" residuals are 0.5, so every single standardized residual is 1.0. Chi-square statistics would report stochastically perfect model fit!

The moral of this story is that fit statistics can be misleading if the estimates are far from the latent parameter values. An estimate can always be thought of as "close", i.e., "different from the estimate for an adjacent raw score", if the expected raw score corresponding to that estimate is within 0.5 score points of its observed value. This places the estimate within 0.5 (SE^2) logits·of its most likely value, where SE is the standard error of the estimate. This level of precision in estimation is not usually required in complete data sets. For instance, for a typical dichotomous item, with p-value of .8, administered to 300 persons, the value of 0.5 (SE^2) = .01 logits is too small to have any noticeable impact on fit computations or almost any substantive decision, and much less than the difficulty estimate's S.E. of .14 logits.

For most estimation methods, it is the estimates corresponding to almost extreme scores that are the last to become close, i.e., to converge. Consequently, methods that announce convergence based on the small size of a change in some average or summary indicator may noticeably misestimate almost extreme measures, usually by making them too central. This also tends to make their fit too good.

Figure 3 plots the reported standardized residuals (on the y-axis) for dichotomous observations for different abilities relative to an item (the contours), and for different amounts of misestimation (on the x-axis). It is seen that that the standardized residuals corresponding to unexpected responses by persons far from an item are those that are most sensitive to misestimation. A one-logit misestimation associated with outlying unexpected responses can reduce their standardized residuals by 30%, effectively halving their squared residuals. These are the values that most influence conventional chi-square statistics, such as Wright's OUTFIT statistic, and, in a similar way, likelihood-ratio tests.

The Growing Family of Rasch Models

The "growing family of Rasch models" (Rost, 2000) presents both opportunities and challenges. Opportunities include widening the application of Rasch measurement methodology, and making better use of it in current areas of application. Challenges include more flexible measurement models and methods of estimation.

Rating Scale Estimation

Under Rasch model conditions, rating scale categories are conceptualized to be ordered indicators of performance on an item. The essential observation is the count of such categories up from the lowest category, whether or not that is the process used by the respondent on the way to being observed in that category. This implies that a category has a specific meaning, i.e., is "well-enough defined" (Masters and Wright, 1984). An obvious rating scale is the two category, "wrong", "right" scale associated with many dichotomous multiple-choice items.

This contrasts with the arbitrary categorization inherent in, say, the Graded Response model and other models which attempt to be "invariant under the grouping of adjacent response categories" (McCullagh, 1985, p.39). For these models, categories, though ordered, are not qualitatively different, but are merely defined for convenience. Such categorizations include arbitrary stratifications of percents, times, distances, weights and the like. These are often encountered in a large-scale surveys, such as a national census.

In practice, a rating scale definition is neither completely defined nor completely arbitrary. For scales of a few categories, the implications of each are usually fairly clear-cut, and, Rasch models, like Equation (1) are usually straight-forward to estimate. But as the number of categories increases, and their meanings become less clear-cut, the parameterization of individual categories can become insecure. Categories may have very low frequencies, or even not be observed.

There is a mathematical device to maintain unobserved intermediate categories in the rating scale structure (Wilson, 1991). If category x is not observed, then Equation (2) can be rewritten,

$$\log\left(\frac{\pi_{ni(x+1)}}{\pi_{ni(x-l)}}\right) = 2\beta_n - \left(2\delta_i + \tau_{x-1,x+1}\right) \qquad (20)$$

where $\tau_{x-1,x+1}$ parameterizes the point at which categories x-1 and x+1 are equally probable, and $\pi_{nix} = 0$. But this device may not be satisfactory from the standpoint of inference, because it predicts that category x will never be observed.

RUMM2010 produces estimates for a reparameterization of the rating scale structure in terms of its mean ("location"), dispersion

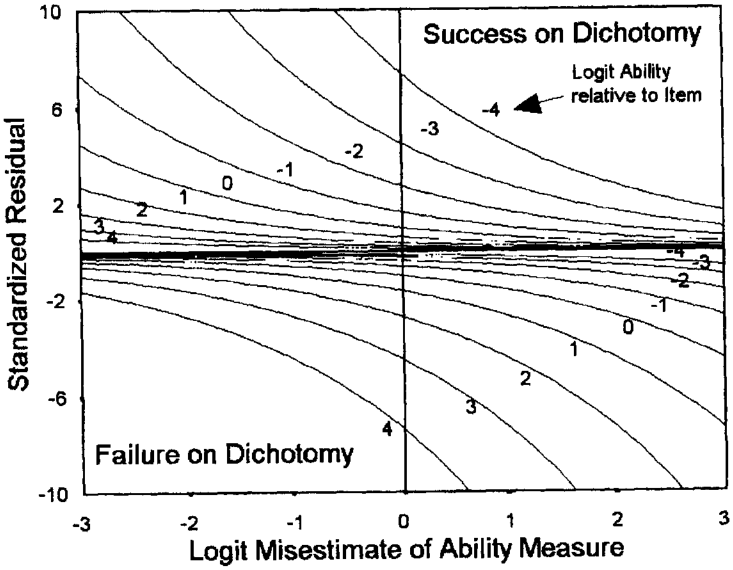

Figure 3. Effect of misestimation on observation fit.

("unit"), skewness and kurtosis. For long rating scales, this enables the inclusion of unobserved categories and regularizes the probabilistic relationship of all the categories. For instance, if the analyst believes that the categories define a regular progression, then this can be imposed on the estimates. Consequently, local, accidental variation in the use of categories is smoothed out. Simplification, however, comes at a cost. Just as with the assertion of distributions for MMLE, care must be taken to verify that the summary structure does not hide important features of the rating scale. For instance, a long bipolar scale might be conveniently summarized by mean and dispersion parameters, but these might obscure a malfunctioning central "don't know" category. Equation (2) can be expressed in terms of a scale structure mean, μ, and dispersion, ξ, as:

$$\log\left(\frac{\pi_{nix}}{\pi_{ni(x-1)}}\right) = \beta_n - (\mu_i + (2x - m)\xi) \quad (21)$$

In this formulation, negative ξ would indicate "disordered thresholds". Disordered thresholds do not imply disordered categories. Disordered categories represent a substantive disagreement between the ordering of the categories and the orientation of the latent variable. An example is a negatively-worded statement on an attitude survey. Unless responses to that item are reverse-scored, its categories will be disordered, in fact reverse-ordered, relative to the latent variable. Disordered categories are usually accomapnied by gross category-level and item-level misfit.

Disordered thresholds reflect the existence of non-modal categories, i.e., categories that do not have a higher probability of being observed than the other categories at any point along the latent variable. In principal, these present no additional estimation problems unless the non-modal category is unobserved, i.e., a sampling or incidental zero. In this case, the device described in (20) above may be employed.

Disordered thresholds indicate that a category corresponds to a narrow interval on the latent variable. This is good if the item is intended to be highly discriminating, so that a small distance along the latent variable corresponds to a large distance along the rating scale. This is bad if each category is intended to correspond to a broad step in development.

Empirically incorrect category and threshold ordering are among aspects of the functioning of rating scales discussed in Linacre (2002a). Omitted, however, from that discussion is the condition necessary for the existence of super-modal categories. Super-modal categories are those, that in some region of the latent variable, have a higher probability of being observed than all other categories combined. This contrasts with modal categories which may not go beyond being more probable than any other single category. The condition for supermodality is that the τ_k advance by (or that $\xi/2$ is) at least 1.1 + $m/10$ logits.

Figure 4 illustrates non-modal, modal and super-modal categories. In this Figure, the model probabilities of a person, of given ability relative to the item difficulty, being observed in each category are shown. Extreme categories, in this case 0 and 4, are always super-modal, because their probabilities asymptotically approach 1.0 at the extremes of the latent variable. Category 3 is here also super-modal as its peak probability exceeds .5. Category 2 is modal because its peak probability is higher than the probability of any other category at some point on the latent variable. Category 1 is non-modal. Either category 0 or category 2 is always more probable than category 1. It is seen that the points of equal probability between categories 0 and 1, τ_1, and between categories 1 and 2, τ_2, are in the reverse order of the categories along the latent variable. These are the "disordered thresholds".

Combinations of item structures and item discriminations

An area of estimation that is only starting to be addressed is the combining of different Rasch mqdels in one analysis. A quality of life instrument, for instance, may include true-false items, Likert scales, frequency scales, intensity scales, Poisson counts of events, Bernoulli trials and paired comparisons all in the same instrument. Rasch estimation methods vary greatly in their capacity to encompass different models in one analysis. JMLE is proving to be the most flexible. Facets (Linacre, 2002b) can co-calibrate all the item types just listed.

An obvious extension of Equation (2) to accommodate several different rating scales structures within the same instrument is:

$$\log\left(\frac{\pi_{nigx}}{\pi_{nig(x-1)}}\right) = \beta_n - (\delta_{ig} + \tau_{xg}) \qquad (22)$$

where *g* indicates a group of items which share the same rating scale response structure. If all items are assigned to the same one group, this is the Rating Scale (Andrich, 1978) model. If each item is assigned to its own group, this is the "Unrestricted" (Andrich, et al., 2000) or "Partial Credit" model (Masters, 1982).

It is typical, however, for each item grouping to exhibit its own sub-dimension, and also to have its own level of discrimination on the latent variable. Different sub-dimensions would be equivalent to incorporating temperature readings from alcohol thermometers, mercury thermometers and thermocouples in the same analysis. Construction of thermometers is now so regularized, that this is no longer thought to be of concern. Psychometrics is not yet so advanced, but, provided the impact of the sub-dimensions is seen to be small, there is little practical motivation for reporting two or more measures whose meaning, for decision-makers, is identical.

Differently discriminating sub-groups of items are more awkward to manage. These are equivalent to mixing Celsius and Fahrenheit temperatures in the same analysis without conversion. Of course, every item, just like every thermometer, has slightly different discrimination. Linacre (2000) indicates that discriminations in the range of 0.5-1.5 (0.9-2.5 on probit scale) can be usefully accommodated.

For wider ranges of discrimination, or for sub-groups comprising relatively high or low discriminating items, OPLM (Verhelst, et al., 1995) permits the imputing of discrimination parameters as though they are known constants. This differs from the 2-PL IRT model in which discriminations are estimated simultaneously with the item difficulties. Verhelst and Glass (1995) suggest statistics which may be useful in guiding the choice of the discrimination constants. A more direct option may be to read them off a graph, such as that provided in Linacre (2000). An alternative, if the subgroups of items are long

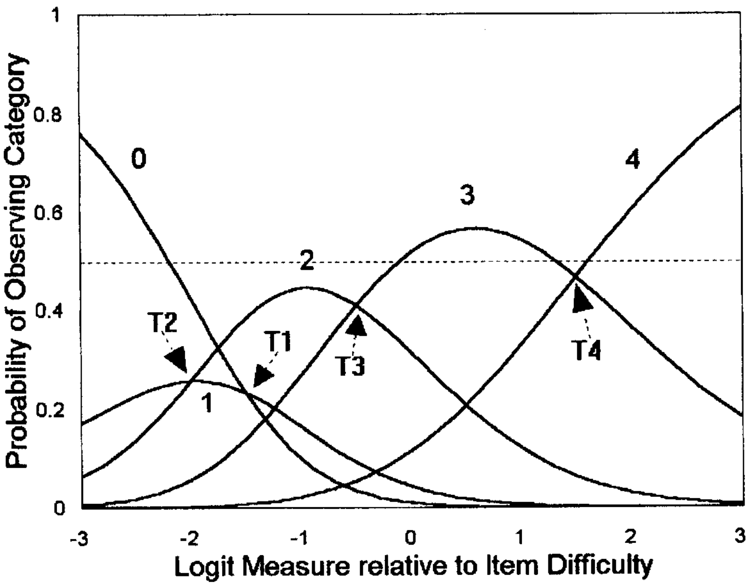

Figure 4. Super-modal, modal and non-modal categories.Table 1:

enough, is to analyze each subgroup separately, and then to combine the resulting estimates after conversion to a common linear scale, in exactly the same way that Celsius and Fahrenheit temperatures are routinely combined.

Conclusion

The remarkable variety and adaptability of Rasch measure estimation algorithms already supports the analysis of a vast range of ordered categorical data. Those few technical intricacies that actually have substantive implications can usually be overcome by reconceptualizing the analytical problem or applying an alternative estimation method. The challenge is no longer to estimate measures, it is to understand and communicate their meaning.

References

Andersen, E. B. (1973). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D., Lyne, A., Sheridan, B., and Luo, G. (2000) RUMM2010 Computer Program. Perth, Australia: Rumm Laboratory Pty. Ltd.

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

Camilli, G. (1994). Origin of the scaling constant d=1.7 in item response theory. *Journal of Educational and Behavioral Statistics, 19,* 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, and Molenaar, I. W. (1995) *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

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

Haberman, S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5,* 815-841.

Jansen, P. G., Van den Wollenberg, A. L., and Wierda, F. W. (1988). Correcting unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12,* 297-306.

Karabatsos, G. (2001). The Rasch model, additive conjoint measurement, and new models of probabilistic measurement theory. *Journal of Applied Measurement, 2,* 389-423.

Kelderman, H., and Steen, R. (1988). *LOGIMO Loglinear and Loglinear IRT Model Analysis.* Chicago: Scientific Software.

Lee, O.-K. (1991). Convergence: statistics or substance? *Rasch Measurement Transactions, 5,* 172.

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

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

Linacre, J. M. (2000). Item discrimination and infit mean-squares. *Rasch Measurement Transactions, 14,* 743.

Linacre, J. M. (2002a). Understanding Rasch measurement: Optimizing category effectiveness. *Journal of Applied Measurement, 3,* 85-106.

Linacre, J. M. (2002b). *Facets Rasch Measurement Computer Program.* Chicago: Facets.com

Linacre, J. M. (2002c). *WINSTEPS Rasch Measurement Computer Program.* Chicago: Winsteps.com

Masters, G. N. (1982) A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N., and Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529-544.

McCullagh, P. (1985). Statistical and scientific aspects of models for qualitative data. In P.

Nijkamp, et al., (Eds.), *Measuring the Unmeasurable* (pp. 39-49). Dordrecht, The Netherlands: Martinus Nijhoff.

Mislevy, R. J., and Bock, R. D. (1996). *BILOG computer program.* Chicago: Scientific Software International.

National Physical Laboratory. (1955). *Calibration of Temperature Measuring Instruments.* London: HMSO.

Nickerson, C.A. & McClelland, G.H. (1984). Scaling distortion in numerical conjoint measurement. *Applied Psychological Measurement, 8,* 182-198.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. H. Fischer & D. Laming (Eds.), *Contributions to Mathematical Psychology, Psychometrics and Methodology.* New York: Springer Verlag.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danish Institute for Educaitonal Research. (Expanded edition, 1980. Chicago: University of Chicago Press.)

Rasch, G. (1969). Models for description of the time-space distribution of traffic accidents. *Symposium on the Use of Statistical Methods in the Analysis of Road Accidents.* Copenhagen: Organization for Economic Cooperation and Development.

Rost, J. (2000). The growing family of Rasch models. Chapter 2 in A. Boomsma, M. A. J. van Duijn, T. A. B. Snijders (Eds.), *Essays on Item Response Theory.* New York: Springer Verlag.

Sheng, S., and Carrière, K. C. (2002) An improved CML estimation procedure for the Rasch model with item response data. *Statistics in Medicine, 21,* 407-416.

Stocking, M. L. (1989). *Empirical Estimation Errors in Item Response Theory as a Function of Test Properties. (Research Report RR-89-5).* Princeton, NJ: ETS.

Thissen, D. (1991). *MULTILOG IRT Computer Program.* Chicago: Scientific Software.

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

Verhelst, N. D., Glas, C. A. W., and Verstralen, H. H. F. M. (1995). *OPLM: One Parameter Logistic Model. Computer program and manual.* Arnhem, The Netherlands: CITO.

Wilson, M. (1991) Unobserved categories. *Rasch Measurement Transactions, 5,* 128.

Wright, B. D. (1988). The efficacy of unconditional maximum likelihood bias correction. *Applied Psychological Measurement, 12,* 315-318.

Wright, B. D., and Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1,* 281-294.

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., and Wilson, M. R. (1998). *ACER ConQuest: generalized 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.

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. *Applied Psychological Measurement, 19,* 369-375.