Volume 3, Number 2, 1999 ISSN 1090-655X

Journal of Outcome Measurement®

Dedicated to Health, Education, and Social Science

REHABILITATION
FOUNDATION
INC. EST. 1993
Research & Education

JOURNAL OF OUTCOME MEASUREMENT®

Volume 3, Number 2 1999

Articles

Indexing/Abstracting Services: JOM is currently indexed in the *Current Index to Journals in Education* (ERIC), *Index Medicus*, and MEDLINE. The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

JOURNAL OF OUTCOME MEASUREMENT®, *3*(2), 103–122

Investigating Rating Scale Category Utility

John M. Linacre
University of Chicago

Eight guidelines are suggested to aid the analyst in investigating whether rating scales categories are cooperating to produce observations on which valid measurement can be based. These guidelines are presented within the context of Rasch analysis. They address features of rating-scale-based data such as category frequency, ordering, rating-to-measure inferential coherence, and the quality of the scale from measurement and statistical perspectives. The manner in which the guidelines prompt recategorization or reconceptualization of the rating scale is indicated. Utilization of the guidelines is illustrated through their application to two published data sets.

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.

Introduction

A productive early step in the analysis of questionnaire and survey data is an investigation into the functioning of rating scale categories. Though polytomous observations can be used to implement multidimensional systems (Rasch and Stene, 1967; Fischer, 1995), observations on a rating scale are generally intended to capture degrees of just one attribute: "rating scales use descriptive terms relating to the factor in question" (Stanley and Hopkins, 1972, p. 290). This factor is also known as the "latent trait" or "variable". The rating scale categorizations presented to respondents are intended to elicit from those respondents unambiguous, ordinal indications of the locations of those respondents along such variables of interest. Sometimes, however, respondents fail to react to a rating scale in the manner the test constructor intended (Roberts, 1994).

Investigation of the choice and functioning of rating scale categories has a long history in social science. Rating scale categorizations should be well-defined, mutually exclusive, univocal and exhaustive (Guilford, 1965). An early finding by Rensis Likert (1932) is that differential category weighting schemes (beyond ordinal numbering) are unproductive. He proposes the well-known five category agreement scale. Nunnally (1967) favors eliminating the neutral category in bi-polar scales, such as Likert's, and presenting respondents an even number of categories. Nunnally (1967, p. 521), summarizing Guilford (1954), also reports that "in terms of psychometric theory, the advantage is always with using more rather than fewer steps." Nevertheless, he also states that "the only exception ... would occur in instances where a large number of steps confused subjects or irritated them." More recently, Stone and Wright (1994) demonstrate that, in a survey of perceived fear, combining five ordered categories into three in the data increases the test reliability for their sample. Zhu et al. (1997) report similar findings for a self-efficacy scale.

Since the analyst is always uncertain of the exact manner in which a particular rating scale will be used by a particular sample, investigation of the functioning of the rating scale is always merited. In cooperation with many other statistical and psycho-linguistic tools, Rasch analysis provides an effective framework within which to verify, and perhaps improve, the functioning of rating scale categorization.

Rasch Measurement Models for Rating Scales

A basic Rasch model for constructing measures from observations on an ordinal rating scale is (Andrich, 1978)

$$\log (P_{nik} / P_{ni(k-1)}) \equiv B_n - D_i - F_k \qquad (1)$$

where

P_{nik} is the probability that person *n*, on encountering item *i* would be observed in category *k*,
$P_{ni(k-1)}$ is the probability that the observation would be in category *k-1*,
B_n is the ability, (attitude etc.), of person *n*,
D_i is the difficulty of item *i*,
F_k is the impediment to being observed in category *k* relative to category *k-1*, i.e., the *k*th step calibration, where the categories are numbered 0,*m*.

This and similar models not only meet the necessary and sufficient conditions for the construction of linear measures from ordinal observations (Fischer, 1995), but also provide the basis for investigation of the operation of the rating scale itself. The Rasch parameters reflecting the structure of the rating scale, the step calibrations, are also known as thresholds (Andrich, 1978).

The prototypical Likert scale has five categories (Strongly Disagree, Disagree, Undecided, Agree, Strongly Agree). These are printed equally spaced and equally sized on the response form (see Figure 1). The intention is to convey to the respondent that these categories are of equal importance and require equal attention. They form a clear progression and they exhaust the underlying variable.

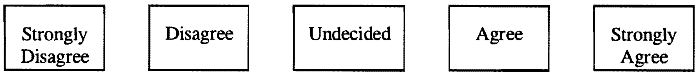


Figure 1. Prototypical Likert scale as presented to the respondent.

From a measurement perspective, the rating scale has a different appearance (Figure 2). The rating categories still form a progression and exhaust the underlying variable. The variable, however, is conceptually infinitely long, so that the two extreme categories are also infinitely wide. However strongly a particular respondent "agrees", we can always posit one who agrees yet more strongly, i.e., who exhibits more of the latent variable. The size of the intermediate categories depends on how they are perceived and used by the respondents. Changing the description of the middle category from "Undecided" to "Unsure" or "Don't Know" or

"Don't Care" will change its meaning psycho-linguistically, and so the amount of the underlying variable it represents, its size as depicted in Figure 2. In view of the general proclivity of respondents towards social conformity, agreeability or mere lethargy, the "agree" option is usually more attractive than the "disagree" one. Hence the "agree" category tends to represent a wider range of the underlying variable.

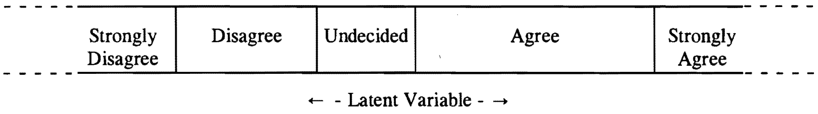


Figure 2. Prototypical Likert scale from a measurement perspective.

Empirically, the observations manifest a stochastic element. The analyst's expectation is that the probability of observing each category is greatest where that category is modeled to occur on the latent variable, but there is always some possibility of observing any category at any point on the continuum. Figure 3 shows probability curves for each category in accordance with Rasch model specifications (Wright and Masters, 1982, p. 81).

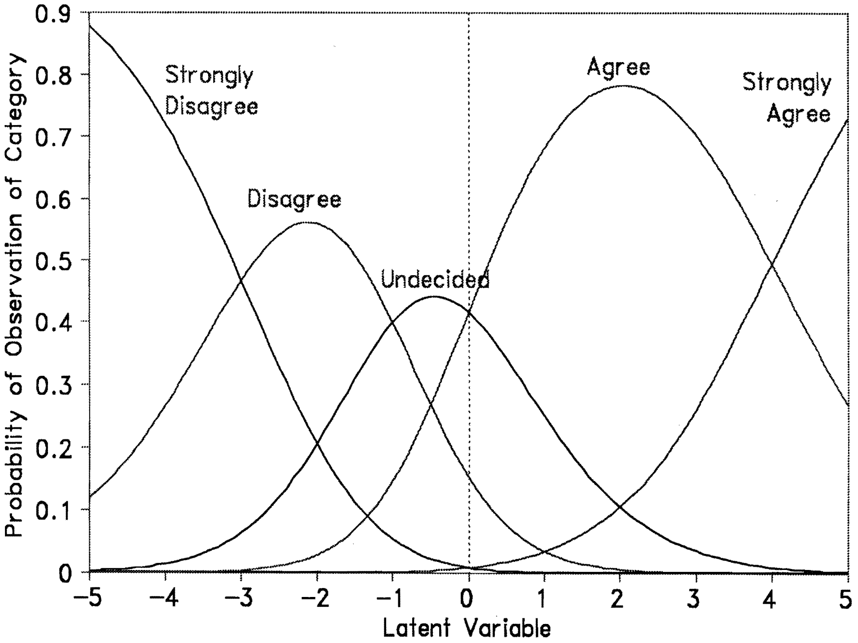


Figure 3. Category probability curves for 5 category Likert scale.

In practice, data do not conform exactly to Rasch model specifications, or those of any other ideal model. "For problem solving purposes, we do not require an exact, but only an approximate resemblance between theoretical results and experimental ones." (Laudan, 1977, p. 224). For analytical purposes, the challenge then becomes to ascertain that the rating scale observations conform reasonably closely to a specified model, such as that graphically depicted in Figure 3. When such conformity is lacking, the analyst requires notification as to the nature of the failure in the data and guidance as to how to remedy that failure in these or future data.

How the variable is divided into categories affects the reliability of a test. Mathematically it can be proven that, when the data fit the Rasch model, there is one best categorization to which all others are inferior (Jansen and Roskam, 1984). Since this best categorization may not be observed in the raw data, guidelines have been suggested for combining categories in order to improve overall measure quality (Wright and Linacre, 1992). Fit statistics, step calibrations and other indicators have also been suggested as diagnostic aids (Linacre, 1995; Andrich, 1996; Lopez, 1996).

The Heuristic Analysis of Rating Scale Observations

At this point we set aside the linguistic aspects of category definitions (Lopez, 1996), taking for granted that the categories implement a clearly defined, substantively relevant, conceptually exhaustive ordered sequence. We consider solely the numerical information that indicates to what extent the data produce coherent raw scores, i.e., raw scores that support the construction of Rasch measures. The description of the characteristics of an ideal rating scale, presented above, suggests an explicit procedure for verifying useful functioning and diagnosing malfunctioning.

Table 1

Analysis of Guilford's (1954) rating scale.

Category Label	Count	%	Average Measure	Expected Measure	OUTFIT MnSq	Step Calibration	Category Name
1	4	4%	-.85	-.73	.8	-	lowest
2	4	4%	-.11	-.57	2.6	-.63	
3	25	24%	-.36*	-.40	.9	-2.31*	
4	8	8%	-.43*	-.22	.5	.84	
5	31	30%	-.04	-.03	.8	-1.48*	middle
6	6	6%	-.46*	.16	4.1	1.71	
7	21	20%	.45	.34	.6	-1.01*	
8	3	3%	.74	.49	.5	2.35	
9	3	3%	.76	.61	.7	.53*	highest

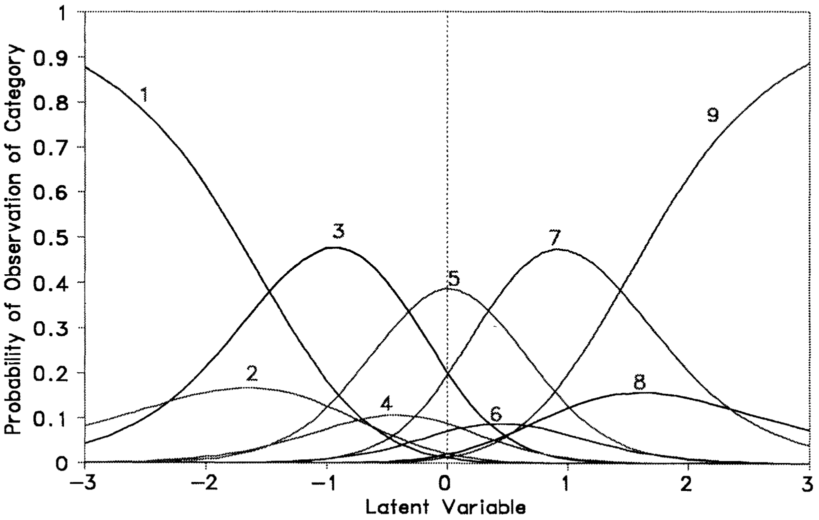


Figure 4. Rasch model category probability curves for Guilford's (1954) scale.

Table 2

Analysis of LFS rating scale data.

Category		Average	Expeoted	OUTFIT	Step	Coherenoe		Soore-to-Measure	
Label	Count	Measure	Measure	MnSq	Calibration	M->C	C->M	----Zone----	
0	378	-.87	-1.03	1.19	NONE	63%	42%	-∞	-1.18
1	620	.13	.33	.69	-.85	54%	71%	-1.18	1.18
2	852	2.23	2.15	1.46	.85	85%	78%	1.18	+∞

Consider the Ratings of Creativity (Guilford, 1954), also discussed from a different perspective in Linacre (1989). Table 1 contains results from an analysis by *Facets* (Linacre, 1989). The model category characteristic curves are shown in Figure 4. These will be contrasted with the "Liking for Science" (LFS) ratings reported in Wright and Masters (1982). Table 2 contains results from an analysis by *BIGSTEPS* (Wright and Linacre, 1991). The model category characteristic curves for the LFS data are shown in Figure 5.

Guideline #1: At least 10 observations of each category.

Each step calibration, F_k, is estimated from the log-ratio of the frequency of its adjacent categories. When category frequency is low, then the step calibration is imprecisely estimated and, more seriously, potentially unstable. The inclusion or exclusion of one observation can noticeably change the estimated scale structure.

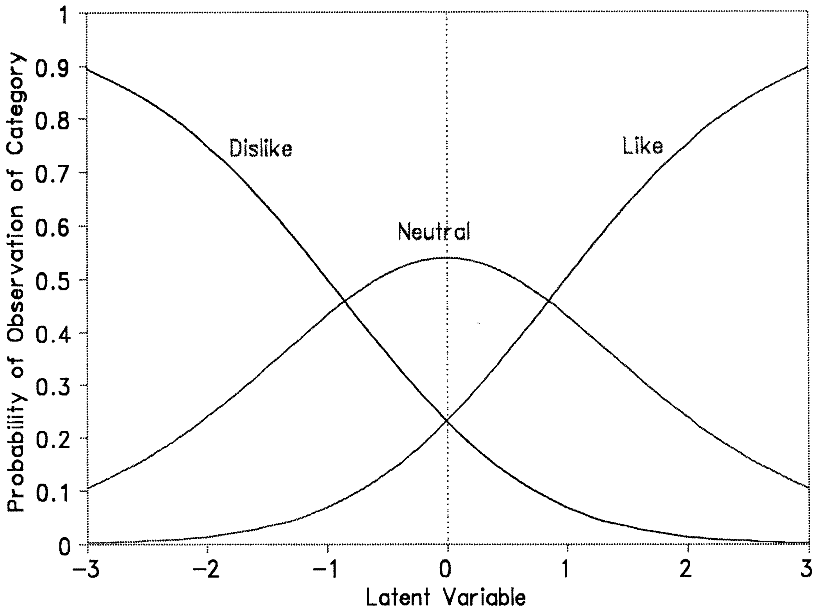


Figure 5. Model probability characteristic curves for LFS rating scale.

For instance, omitting one of ten observations changes the step calibration by more than .1 logits, (more than .2 logits for one of 5). If each item is defined to have its own rating scale, i.e., under partial credit conditions, this would also change the estimated item difficulty by .1/*m*, when there are *m*+1 categories and so *m* steps. For many data sets, this value would exceed the model standard error of the estimated item difficulty based on 100 observations. Consequently, the paradox can arise that a sample large enough to provide stable item difficulty estimates for less statistically informative dichotomous items (Linacre, 1994) may not be sufficiently large for more informative polytomous items.

Categories which are not observed in the current dataset require special attention. First, are these structural or incidental zeros? Structural zeros correspond to categories of the rating scale which will never be observed. They may be an artifact of the numbering of the observed categories, e.g., categories "2" and "4" cannot be observed when there are only three scale categories and these are numbered "1", "3" and "5". Or structural zeros occur for categories whose requirements are impossible to fulfil, e.g., in the 17th Century it was conventional to assign the top category to God-level performance. For these structural zeros, the categories are simply omitted, and the remaining categories renumbered sequentially to represent the only observable qualitative levels of performance.

Incidental zeroes are categories that have not been observed in this particular data set. Thus all categories of a 5 category scale cannot be seen in just three observations. There are several strategies that avoid modifying the data: i) treat those incidental zeroes as structural for this analysis, renumbering the categories without them; ii) impose a scale structure (by anchoring thresholds) that includes these categories; iii) use a mathematical device (Wilson, 1991) to keep intermediate zero categories in the analysis.

In the Guilford example (Table 1), category frequency counts as low as 3 are observed. When further relevant data cannot be easily obtained, one remedy is to combine adjacent categories to obtain a robust structure of high frequency categories. Another remedy is to omit observations in low frequency categories that may not be indicative of the main thrust of the latent variable. Such off-dimension categories may be labeled "don't know" or "not applicable". The frequency count column, by itself, suggests that the rarely observed categories, 1, 2, 4, 6, 8, 9, be combined with adjacent categories or their data be omitted. The remaining categories would be renumbered sequentially and then the data reanalyzed.

In the LFS example (Table 2), all category frequency counts are large, indicating that locally stable estimates of the rating scale structure can be produced.

Guideline #2: Regular observation distribution.

Irregularity in observation frequency across categories may signal aberrant category usage. A uniform distribution of observations across categories is optimal for step calibration. Other substantively meaningful distributions include unimodal distributions peaking in central or extreme categories, and bimodal distributions peaking in extreme categories. Problematic are distributions of "roller-coaster" form, and long tails of relatively infrequently used categories. On the other hand, when investigating highly skewed phenomena, e.g., criminal behavior or creative genius, the long tails of the observation distribution may capture the very information that is the goal of the investigation.

A consideration, when combining or omitting categories, is that the rating scale may have a substantive pivot-point, the point at which the substantive meaning of the ratings is dichotomized. For instance, when using a Likert scale to ask about socially-acceptable propositions, such as "Crime should be punished", the pivot point could be between "Strongly Agree", and "Agree". For negatively worded propositions, such as "Politicians are dishonest", the pivot could be between "Disagree" and "Neutral".

In Table 1, the frequency distribution is tri-modal with peaks at 3, 5, and 7, perhaps indicating that the judges are being asked to apply a 9 category scale to performances that they can only discriminate into three levels. Again, remedies include combining adjacent categories or omitting observations in categories, such as "Other", whose measurement implications are dubious. A regular frequency distribution in Table 1 could be obtained by combining categories 1, 2 and 3, totaling 33, also 4 and 5, totaling 39, and then 6, 7, 8, and 9, totaling 33.

In Table 2, the frequency distribution is unimodal and shows reassuringly smooth increases from approximately 380 to 620 (a jump of 240), and then from 620 to 850 (a jump of 230).

Guideline #3: Average measures advance monotonically with category.

Equation (1) specifies a Rasch measurement model. This is conceptualized to generate data in the following fashion:

$$B_n - D_i - \{F_k\} \Rightarrow X_{ni} \qquad (2)$$

where

X_{ni} is the rating observed when person *n* encountered item *i*,
$\{F_k\}$ is the set of step calibrations for all categories 0, *m*,
and other parameters have the meanings assigned in (1).

Within any one item or group of items modeled to have the same rating scale structure, the $\{F_k\}$ are constant across observations and may be ignored at this point. It is the combination of B_n and D_i (or their equivalent in any other Rasch model) that is crucial in producing, and then diagnosing, the empirical observation, X_{ni}. It is essential to our comprehension of the rating scale that, in general, higher measure combinations $(B_n - D_i)$ produce observations in higher categories and *vice-versa.* Accordingly a diagnostic indicator is the average of the measures, $(B_n - D_i)$, across all observations in each category.

These average measures are an empirical indicator of the context in which the category is used. In general, observations in higher categories must be produced by higher measures (or else we don't know what a "higher" measure implies). This means that the average measures by category, for each empirical set of observations, must advance monotonically up the rating scale. Otherwise the meaning of the rating scale is uncertain for that data set, and consequently any derived measures are of doubtful utility.

In Table 1, failures of average measures to demonstrate monotonicity are flagged by "*". In particular, the average measure corresponding to the 6 observations in category 6 is -.46, noticeably less than the -.04 for the 31 observations in category 5. Empirically, category 6 does not manifest higher performance levels than category 5. An immediate remedy is to combine non-advancing (or barely advancing) categories with those below them, and so obtain a clearly monotonic structure. The average measure column of Table 2, by itself, suggests that categories 2, 3, and 4 be combined, and also categories 5, 6, and 7. Categories 1, 8 and 9 are already monotonic.

In Table 2, the average measures increase monotonically with rating scale category from -.87 to .13 logits (a jump of 1.0), and then from .13 to 2.23 (a jump of 2.2). This advance is empirical confirmation of our intention that higher rating scale categories indicate more of the latent variable. The advances across categories, however, are uneven. This may be symptomatic of problems with the use of the rating scale or may merely reflect the item and sample distributions.

The "Expected Measure" columns in Tables 1 and 2 contain the values that the model predicts would appear in the "Average Measure" columns, were the data to fit the model. In Table 1, these values are diagnostically useful. For category 1, the observed and expected values, -.85 and -.73, are close. For category 2, however, the observed value of -.11 is .46 logits higher than the expected value of .57, and also higher than the expected value for category 4. Category 6 is yet more aberrant, with an observed average measure less than the expected average measure for category 3. The observations in categories 2 and 6 are so contradictory to the intended use of the rating scale, that, even on this slim evidence, it may be advisable to remove them from this data set.

In Table 2, the observed average measures appear reasonably close to their expected values.

Guideline #4: OUTFIT mean-squares less than 2.0.

The Rasch model is a stochastic model. It specifies that a reasonably uniform level of randomness must exist throughout the data. Areas within the data with too little randomness, i.e., where the data are too predictable, tend to expand the measurement system, making performances appear more different. Areas with excessive randomness tend to collapse the measurement system, making performances appear more similar. Of these two flaws, excessive randomness, "noise", is the more immediate threat to the measurement system.

For the Rasch model, mean-square fit statistics have been defined such that the model-specified uniform value of randomness is indicated by 1.0 (Wright and Panchapakesan, 1969). Simulation studies indicate that values above 1.5, i.e., with more than 50% unexplained randomness, are problematic (Smith, 1996). Values greater than 2.0 suggest that there is more unexplained noise than explained noise, so indicating there is more misinformation than information in the observations. For the outlier-sensitive OUTFIT mean-square, this misinformation may be confined to a few substantively explainable and easily remediable observations. Nevertheless large mean-squares do indicate that segments of the data may not support useful measurement.

For rating scales, a high mean-square associated with a particular category indicates that the category has been used in unexpected contexts. Unexpected use of an extreme category is more likely to produce a high mean-square than unexpected use of a central category. In fact, central categories often exhibit over-predictability, especially in situations where respondents are cautious or apathetic.

In Table 1, category 6 has an excessively high mean-square of 4.1. It has more than three times as much noise as explained stochasticity. From the standpoint of the Rasch model, these 6 observations were highly unpredictable. Inspection of the data, however, reveals that only one of the three raters used this category, and that it was used in an idiosyncratic manner. Exploratory solutions to the misfit problem could be to omit individual observations, combine categories or drop categories entirely. Category 2, with only 4 observations also has a problematic mean-square of 2.1. One solution, based on mean-square information alone, would be to omit all observations in categories 2 and 6 from the analysis.

In Table 2, central category 1 with mean-square .69 is showing some over-predictability. In the data, one respondent choose this category in responses to all 25 items, suggesting that eliminating that particular respondent's data would improve measurement without losing information. Extreme category 2 with mean-square 1.46 is somewhat noisy. This high value is cause by a mere 6 observations. Inspection of these ratings for data entry errors and other idiosyncracies is indicated.

Guideline #5: Step calibrations advance.

The previous guidelines have all considered aspects of the current sample's use of the rating scale. This guideline concerns the scale's inferential value. An essential conceptual feature of rating scale design is that increasing amounts of the underlying variable in a respondent correspond to a progression through the sequentially categories of the rating scale (Andrich, 1996). Thus as measures increase, or as individuals with incrementally higher measures are observed, each category of the scale in turn is designed to be most likely to be chosen. This intention corresponds to probability characteristic curves, like those in Figure 3, in which each category in turn is the most probable, i.e., modal. These probability curves look like a range of hills. The extreme categories always approach a probability of 1.0 asymptotically, because the model specifies that respondents with infinitely high (or low) measures must be observed in the highest (or lowest) categories, regardless as to how those categories are defined substantively or are used by the current sample.

The realization of this requirement for inferential interpretability of the rating scale is that the Rasch step calibrations, $\{F_k\}$, advance monotonically with the categories. Failure of these parameters to advance monotonically is referred to as "step disordering". Step disordering does not imply that the substantive definitions of the categories are disordered, only that their step calibrations are. Disordering reflects the low probability of observance of certain categories because of the manner in which those categories are used in the rating process. This degrades the interpretability of the resulting measures. Step disordering can indicate that a category represents too narrow a segment of the latent variable or a concept that is poorly defined in the minds of the respondents.

Disordering of step calibrations often occurs when the frequencies of category usage follow an irregular pattern. The most influential components in the estimation of the step calibrations are the log-ratio of the frequency of adjacent categories and the average measures of the respondents choosing each category. Thus,

$$F_k \approx \log (T_{k-1}/T_k) - B_k + B_{k-1} \qquad (3)$$

where

T_k is the observed frequency of category *k*,
T_{k-1} is the observed frequency of category *k-1*,
B_k is the average measure of respondents choosing category *k*,
and B_{k-1} is the average measure of those choosing category *k-1*.

It can be seen that step-disordering may result when a higher category is relatively rarely observed or a lower category is chosen by respondents with higher measures.

In Table 1, disordered step calibrations are indicated with "*". The step calibrations correspond to the intersections in the probability curve plot, Figure 4. The step calibration from category 2 to category 3, F_3, is -2.31 logits. In Figure 4, this is the point where the probability curves for categories 2 and 3 cross at the left side of the plot. It can be seen that category 2 is never modal, i.e, at no point on the variable is category 2 ever the most likely category to be observed. The peak of category 2's curve is submerged, and it does not appear as a distinct "hill". Figure 4 suggests that a distinct range of hills, and so strict ordering of the step calibrations, would occur if categories 2 and 3 were combined, and also 4, 5, and 6, and finally 7 and 8. Since the extreme categories, 1 and 9, are always modal, it is not clear from this plot whether it would be advantageous to combine one or both of them with a neighboring, more central category.

In Table 2, the step calibrations, -.85 and +.85 are ordered. The corresponding probability curves in Figure 5 exhibit the desired appearance of a range of hills.

Guideline #6: Ratings imply measures, and measures imply ratings.

In clinical settings, action is often based on one observation. Consequently it is vital that, in general, a single observation imply an equivalent underlying measure. Similarly, from an underlying measure is inferred what behavior can be expected and so, in general, what rating would be observed on a single item. The expected item score ogive, the model item characteristic curve (ICC), depicts the relationship between measures and average expected ratings.

Figure 6 shows the expected score ogive for the 5 category Likert scale depicted in Figure 3. The y-axis shows the average expected rating. Since only discrete ratings can be observed, this axis has been partitioned at the intermediate .5 average rating points. To the practitioner, an expected average rating near to 4.0, (e.g., 3.75), implies that a rating of "4" will be observed. The expected score ogive facilitates the mapping of these score ranges on the y-axis into measure zones on the x-axis, the latent variable. The implication is that measures in, say, the "4" zone on the x-axis, will be manifested by average ratings between 3.5 and 4.5, and so be observed as ratings of "4". Equally, to be interpretable, observed

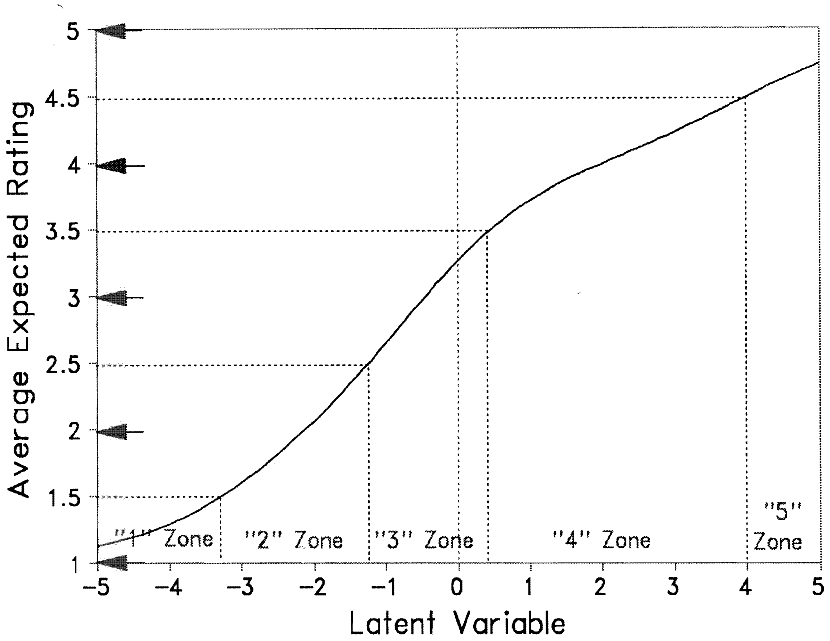


Figure 6. Expected score ogive for 5 category Likert scale showing rating-measure zones.

ratings of "4" on the y-axis imply respondent measures within the "4" zone of the latent variable.

In Table 2, the "Coherence" columns report on the empirical relationship between ratings and measures for the LFS data. The computation of Coherence is outlined in Table 3. M->C (Measure implies Category %) reports what percentage of the ratings, expected to be observed in a category (according to the measures), are actually observed to be in that category.

The locations of the measure "zone" boundaries for each category are shown in Table 2 by the Score-to-Measure Zone columns. Consider the M->C of category 0. 63% of the ratings that the measures would place in category 0 were observed to be there. The inference of measures-to-ratings is generally successful. The C->M (Category implies Measure %) for category 0 is more troublesome. Only 42% of the occurrences of category 0 were placed by the measures in category 0. The inference of ratings-to-measures is generally less successful. Nevertheless, experience with other data sets (not reported here) indicates that 40% is an empirically useful level of coherence.

Table 3.

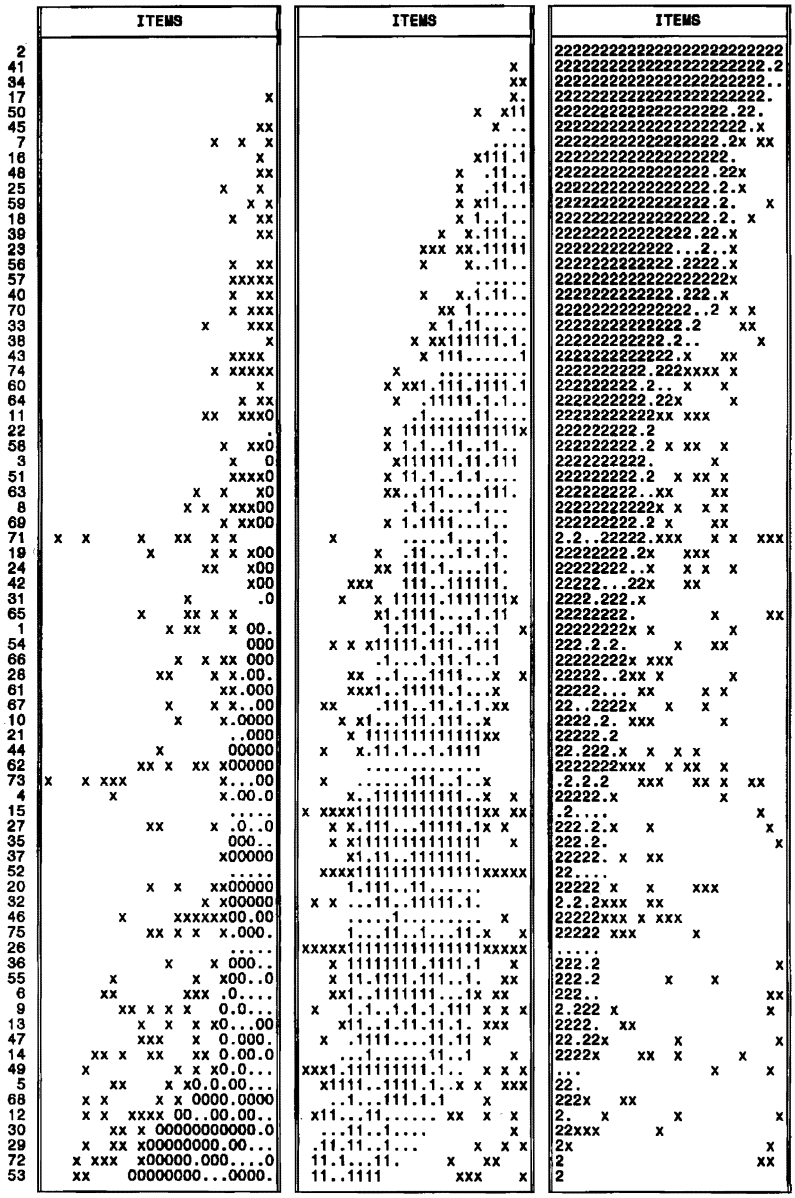
Coherence of Observations.

	Observed Rating in Category	Observed Rating **outside** Category
Observed Measure in Zone	ICIZ (Rating in Figure 7)	OCIZ ("." in Figure 7)
Observed Measure **outside** Zone	ICOZ ("x" in Figure 7)	-
M->C = In Category & Zone / All in Zone = ICIZ / (ICIZ + OCIZ) * 100%		
C->M = In Category & Zone / All in Category = ICIZ / (ICIZ + ICOZ) * 100%		

Figure 7 shows the Guttman scalogram for the LFS data, partitioned by category, for categories 0, 1, and 2, left to right. In each section ratings observed where their measures predict are reported by their rating value, "0", "1", or "2". Ratings observed outside their expected measure zone are marked by "x". Ratings expected in the specified category, but not observed there, are marked by ".". In each partition, the percentage of ratings reported by their category numbers to such ratings and "."s is given by M->C. The percentage of ratings reported by their category numbers to such ratings and "x"s is given by C->M. In the left-hand panel, for category 0, the there are about twice as many "0"s as "."s, so C->M coherence of 63% is good. On the other hand, there are more "x"s than "0"s, so M->C coherence of 42% is fragile. The inference from measures to ratings for category 0 is strong, but from ratings to measures is less so. This suggests that local inference for these data would be more secure were categories 0 and 1 to be combined.

Guideline #7: Step difficulties advance by at least 1.4 logits.

It is helpful to communicate location on a rating scale in terms of categories below the location, i.e., passed, and categories above the location, i.e., not yet reached. This conceptualizes the rating scale as a set of dichotomous items. Under Rasch model conditions, a test of *m* dichotomous items is mathematically equivalent to a rating scale of *m*+1 categories (Huynh, 1994). But a rating scale of *m*+1 categories is only equivalent to test of *m* dichotomous items under specific conditions (Huynh, 1996).



```
    |          ITEMS          | |          ITEMS          | |          ITEMS          |
    |-------------------------| |-------------------------| |-------------------------|
  2 |                         | |                         | |2222222222222222222222222|
 41 |                         | |                       x | |22222222222222222222222.2|
 34 |                         | |                       xx| |22222222222222222222222..|
 17 |                        x| |                       x.| |22222222222222222222222. |
 50 |                         | |                   x  x11| |2222222222222222222.22.  |
 45 |                       xx| |                     x ..| |222222222222222222222.x  |
  7 |                  x  x  x| |                     ....| |222222222222222222.2x xx |
 16 |                       x | |                   x111.1| |2222222222222222222.     |
 48 |                       xx| |                 x  .11..| |22222222222222222.22x    |
 25 |                   x   x | |                 x  .11.1| |22222222222222222.2.x    |
 59 |                      x x| |                 x x11...| |22222222222222222.2.   x |
 18 |                    x  xx| |                 x 1..1..| |22222222222222222.2. x   |
 39 |                       xx| |               x  x.111..| |222222222222222.22.x     |
 23 |                         | |             xxx xx.11111| |2222222222222...2..x     |
 56 |                    x  xx| |             x    x..11..| |2222222222222.2222.x     |
 57 |                    xxxxx| |                   ......| |2222222222222222222x     |
 40 |                    x  xx| |             x   x.1.11..| |2222222222222.222.x      |
 70 |                    x xxx| |               xx 1......| |222222222222222..2 x x   |
 33 |                 x    xxx| |              x 1.11.....| |22222222222222.2    xx   |
 38 |                        x| |            x xx111111.1.| |222222222222.2..      x  |
 43 |                    xxxx | |             x 111......1| |2222222222222.x   xx     |
 74 |                  x xxxxx| |          x    ..........| |2222222222.222xxxx x     |
 60 |                       x | |         x xx1.111.1111.1| |222222222.2.. x   x      |
 64 |                     x xx| |          x  .11111.1.1..| |2222222222.22x     x     |
 11 |                 xx  xxx0| |            .1.....11....| |22222222222xx xxx        |
 22 |                        .| |         x 1111111111111x| |222222222.2              |
 58 |                   x  xx0| |         x 1.1..11..11.. | |222222222.2 x xx  x      |
  3 |                    x   0| |          x111111.11.111 | |2222222222.      x       |
 51 |                    xxxx0| |         x 11.1..1.1.... | |222222222.2  x xx x      |
 63 |                x  x   x0| |         xx..111....111. | |222222222..xx    xx      |
  8 |               x x  xxx00| |           .1.1....1...  | |22222222222x x  x x      |
 69 |                   x xx00| |         x 1.1111...1..  | |222222222.2 x    xx      |
 71 | x  x     x   xx  x x    | |   x       .....1....1.  | |2.2..22222.xxx   x x  xxx|
 19 |           x      x x x00| |        x  .11...1.1.1.  | |2222222.2x    xxx        |
 24 |                 xx   x00| |        xx 111.1....11.  | |2222222..x    x x  x     |
 42 |                      x00| |     xxx   111..111111.  | |22222...22x   xx         |
 31 |               x       .0| |    x   x 11111.1111111x | |2222.222.x               |
 65 |          x    xx x x    | |        x1.1111....1.11  | |22222222.        x     xx|
  1 |             x xx   x 00.| |         1.11.1..11..1  x| |2222222x x         x     |
 54 |                      000| |   x x x11111.111..111   | |222.2.2.     x   xx      |
 66 |              x  x xx 000| |        .1...1.11.1..1   | |2222222x xxx             |
 28 |            xx    x x.00.| |     xx ..1...1111...x  x| |22222..2xx x       x     |
 61 |                   xx.000| |     xxx1..11111.1...x   | |22222... xx     x x      |
 67 |             x    x x..00| |  xx    .111..11.1.1.xx  | |22..2222x   x   x        |
 10 |              x    x.0000| |    x x1...111.111..x    | |2222.2. xxx       x      |
 21 |                    ..000| |     x 1111111111111xx   | |22222.2                  |
 44 |            x       00000| |  x   x.11.1..1.1111     | |22.222.x  x  x x         |
 62 |          xx x  xx x00000| |       .............     | |2222222xxx  x xx  x      |
 73 |x   x xxx          x...00| |  x   ......111..1..x    | |.2.2.2   xxx   xx x  xx  |
  4 |       x           x.00.0| |     x..1111111111..x  x | |22222.x           x      |
 15 |                    .....| |x xxxx11111111111111xx xx| |.2....                x  |
 27 |           xx     x .0..0| |   x x.111...11111.1x x  | |222.2.x   x            x |
 35 |                    000..| |   x x11111111111111   x | |222.2.                  x|
 37 |                   x00000| |     x1.11..1111111.     | |22222. x  xx             |
 52 |                    .....| |  xxxx11111111111111xxxxx| |22....                   |
 20 |           x  x   xx00000| |     1.111..11......     | |22222 x   x    xxx       |
 32 |                 x x00000| | x x ...11..11111.1.     | |2.2.2xxx  xx             |
 46 |        x     xxxxxx00.00| |     .....1.........  x  | |22222xxx x xxx           |
 75 |           xx x x  x.000.| |     1...11..1...11.x   x| |22222 xxx      x         |
 26 |                    .....| |xxxxx111111111111111xxxxx| |.....                    |
 36 |             x    x 000..| |   x 11111111.1111.1   x | |222.2                   x|
 55 |       x        x  x00..0| |   x 11.1111.111..1.  xx | |222.2       x    x       |
  6 |      xx       xxx .0....| |   xx1..1111111...1x xx  | |222..                  xx|
  9 |        xx x x x   0.0...| | x   1.1..1.1.1.111 x x x| |2.222 x                x |
 13 |          x  x  x x0...00| |    x11..1.11.11.1. xxx  | |2222.  xx                |
 47 |          xxx   x  0.000.| |  x  .1111....11.11 x    | |22.22x       x          x|
 14 |     xx x  xx   xx 0.00.0| |    ...1......11..1    x | |2222x    xx  x      x    |
 49 |    x         x x x0.0...| |xxx1.11111111.1..   x x x| |...              x     x |
  5 |       xx    x x0.0.00...| |  x1111..1111.1..x x  xxx| |22.                      |
 68 |    x x     x x 0000.0000| |   ..1...111.1.1    x    | |222x   xx                |
 12 |    x x  xxxx 00..00.00..| | x11...11...... xx  x  x | |2.   x       x          x|
 30 |       xx x 00000000000.0| |  ...11..1....         x | |22xxx      x             |
 29 |    x  xx x00000000.00...| | .11.11..1...      x  x x| |2x                     x |
 72 |   x xxx  x00000.000....0| | 11.1...11.     x   xx   | |2                     xx |
 53 |   xx    00000000...0000.| | 11..1111        xxx    x| |2                        |
```

Figure 7. Guttman scalograms of LFS data, flagging out-of-zone observations wiht “x”.

For practical purposes, when all step difficulty advances are larger than 1.4 logits, then a rating scale of m+1 categories can be decomposed, in theory, into a series of independent dichotomous items. Even though such dichotomies may not be empirically meaningful, their possibility implies that the rating scale is equivalent to a sub-test of *m* dichotomies. For developmental scales, this supports the interpretation that a rating of *k* implies successful leaping of *k* hurdles. Nevertheless, this degree of rating scale refinement is usually not required in order for valid and inferentially useful measures to be constructed from rating scale observations.

The necessary degree of advance in step difficulties lessens as the number of categories increases. For a three category scale, the advance must be at least 1.4 logits between step calibrations in order for the scale to be equivalent to two dichotomies. For a five category rating scale, advances of at least 1.0 logits between step calibrations are needed in order for that scale to be equivalent to four dichotomies.

In Table 2, the step calibrations advance from -.85 to +.85 logits, a distance of 1.7. This is sufficiently large to consider the LFS scale statistically equivalent to a 2-item sub-test with its items about 1.2 logits apart. When the two step calibrations are -.7 and +.7, then the advance is 1.4 logits (the smallest to meet this guideline), and the equivalent sub-test comprises two items of equal difficulty. When the advance is less than 1.4 logits, redefining the categories to have wider substantive meaning or combining categories may be indicated.

Guideline #8: Step difficulties advance by less than 5.0 logits

The purpose of adding categories is to probe a wider range of the variable, or a narrow range more thoroughly. When a category represents a very wide range of performance, so that its category boundaries are far apart, then a "dead zone" develops in the middle of the category in which measurement loses its precision. This is evidenced statistically by a dip in the information function. In practice, this can result from Guttman-style (forced consensus) rating procedures or response sets.

In Figure 8, the information functions for three category (two step) items are shown. When the step calibrations are less than 3 logits apart, then the information has one peak, mid-way between the step calibrations. As the step calibrations become farther apart, the information function sags in the center, indicating that the scale is providing less information about the respondents apparently targeted best by the scale. Now the scale is better at probing respondents at lower and higher decision points than at the center. When the distance between step calibrations is more than 5 logits, the information provided at the item's center is less than half that provided by a simple dichotomy. When ratings collected under circumstances which encourage rater consensus are subjected to Rasch analysis, wide distances between step calibrations may be observed. Distances of 30 logits have been seen. Such results suggest that the raters using such scales are not locally-independent experts, but rather rating machines. A reconceptualization of the function of the raters or the use of the rating scale in the measurement process may be needed.

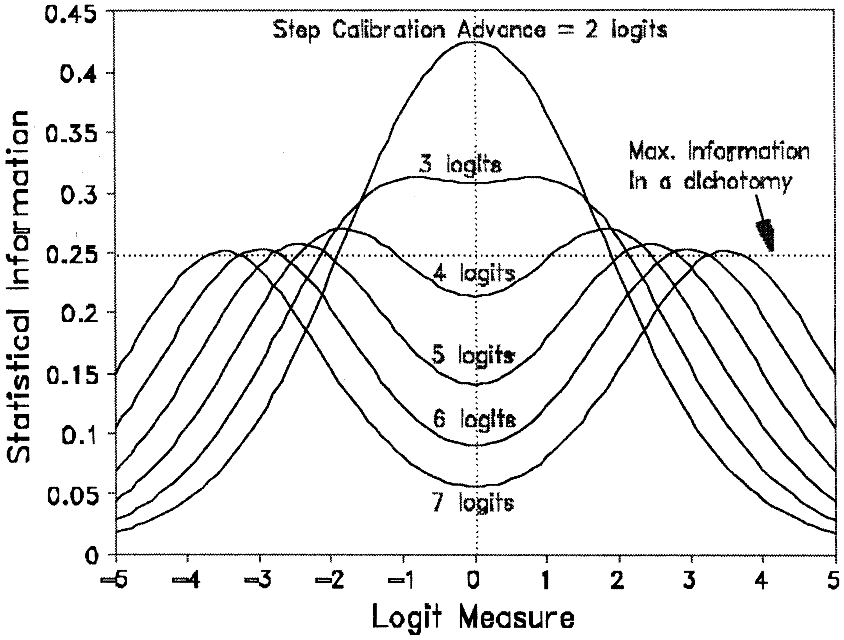


Figure 8. Information functions for a three-category rating scale.

In clinical applications, discovery of a very wide intermediate category suggests that it may be productive to redefine the category as two narrower categories. This redefinition will necessarily move all category thresholds, but the clinical impact of redefinition of one category on other clearly defined categories is likely to be minor, and indeed may be advantageous.

Conclusion

Unless the rating scales which form the basis of data collection are functioning effectively, any conclusions based on those data will be insecure. Rasch analysis provides a technique for obtaining insight into how the data cooperate to construct measures. The purpose of these guidelines is to assist the analyst in verifying and improving the functioning of rating scale categories in data that are already extant. Not all guidelines are relevant to any particular data analysis. The guidelines may even suggest contradictory remedies. Nevertheless they provide a useful starting-point for evaluating the functioning of rating scales.

References

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

Andrich, D. A. (1996). Measurement criteria for choosing among models for graded responses. In A. von Eye and C. C. Clogg (Eds.) *Analysis of categorical variables in developmental research.* Orlando FL: Academic Press. Chapter 1, 3-35.

Fischer, G. H. (1995). The derivation of polytomous Rasch models. Chapter 16 in G. H. Fischer and I. W. Molenaar (Eds.) *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer Verlag.

Guilford, J. P. (1954). *Psychometric Methods. 2nd Edn.* New York: McGraw-Hill.

Guilford, J.P. (1965). *Fundamental Statistics in Psychology and Education,* 4th Edn. New York: McGraw-Hill.

Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika, 59*, 111-119.

Huynh, H. (1996). Decomposition of a Rasch partial credit item into independent binary and indecomposable trinary items. *Psychometrika, 61*(1) 31-39.

Jansen, P.G.W. and Roskam, E.E. (1984). The polytomous Rasch model and dichotomization of graded responses. p. 413-430. in E. Degreef and J. van Buggenhaut (Eds), *Trends in Mathematical Psychology.* Amsterdam: North-Holland.

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

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology.* 140:1-55.

Linacre, J.M. (1989). *Facets computer program for many-facet Rasch measurement.* Chicago: MESA Press.

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

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4) 328.

Linacre, J.M. (1995). Categorical misfit statistics. *Rasch Measurement Transactions, 9*(3) 450-1.

Lopez, W. (1996). Communication validity and rating scales. *Rasch Measurement Transactions, 10*(1) 482.

Nunnally, J. C. (1967). *Psychometric Theory.* New York: McGraw Hill.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Institute for Educational Research. Reprinted, 1992, Chicago: MESA Press.

Rasch, G. and Stene, J. (1967). *Some remarks concerning inference about items with more than two categories.* (Unpublished paper).

Roberts, J. (1994). Rating scale functioning. *Rasch Measurement Transactions, 8*(3) 386.

Smith, R.M. (1996). Polytomous mean-square statistics. *Rasch Measurement Transactions, 10*(3) p. 516-517.

Stanley, J. C. and Hopkins, K. D. (1972). *Educational and Psychological Measurement and Evaluation.* Englewood Cliffs, N.J.: Prentice-Hall Inc.

Stone M. H. and Wright B.D. (1994). Maximizing rating scale information. *Rasch Measurement Transactions, 8*(3) 386.

Wilson, M. (1991). Unobserved categories. *Rasch Measurement Transactions 5*(1) 128.

Wright, B.D. and Linacre, J.M. (1992). Combining and splitting categories. *Rasch Measurement Transactions, 6*(3) 233-235.

Wright, B.D. and Linacre, J.M. (1991). *BIGSTEPS computer program for Rasch measurement.* Chicago: MESA Press.

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

Wright, B. D. and Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

Zhu, W., Updyke, W.F. and Lewandowski C. (1997). Post-Hoc Rasch analysis of optimal categorization of an ordered response scale. *Journal of Outcome Measurement, 1*(4) 286-304.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(2), 123–133

Using IRT Variable Maps to Enrich Understanding of Rehabilitation Data

Wendy Coster
Boston University

Larry Ludlow
Boston College

Marisa Mancini
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil

One of the benefits of item response theory (IRT) applications in scale development is the greater transparency of resulting scores. This feature allows translation of a total score on a particular scale into a profile of probable item responses with relative ease, for example by using the variable map that often is part of IRT analysis output. Although there have been a few examples in the literature using variable maps to enrich clinical interpretation of individual functional assessment scores, this feature of IRT output has received very limited application in rehabilitation research. The present paper illustrates the application of variable maps to support more in-depth interpretation of functional assessment scores in research and clinical contexts. Two examples are presented from an outcome prediction study conducted during the standardization of a new functional assessment for elementary school students with disabilities, the *School Function Assessment.* Two different applications are described: creating a dichotomous outcome variable using scores from a continuous scale, and interpreting the meaning of a classification cut-off score identified through Classification and Regression Tree (CART) analysis.

Requests for reprints should be sent to Wendy Coster, Ph.D., OTR/L, Department of Occupational Therapy, Sargent College of Health and Rehabilitation Sciences, Boston University, 635 Commonwealth Avenue, Boston, MA 02215.

Support for this project was provided in part by grant #133G30055 from the U.S. Department of Education, National Institute on Disability and Rehabilitation Research to Boston University.

One of the benefits of item response theory (IRT) applications in scale development is the greater transparency of resulting scores (e.g., Fisher, 1993). That is, assuming a well-fitting model, a total score on a particular scale can be translated into a profile of probable item responses with relative ease. Graphic versions of these profiles, often referred to as "variable maps", are readily obtained in the output from programs such as BIGSTEPS (Linacre and Wright, 1993). In some instances this graphic output has been incorporated into research reports as a useful way to display information about item distribution along the continuum represented by the scale (e.g., Ludlow, Haley, and Gans, 1992; Wright, Linacre, and Heinemann, 1993). To date, however, this feature of IRT has received very limited attention in the research literature compared to the growing body of information on applications of IRT in scale development.

The purpose of the present paper is to encourage greater use of variable maps from IRT analyses to support theoretically and clinically meaningful use of functional assessment scores in both research and clinical contexts. Application of this feature can help bridge the gap between quantitative summaries of rehabilitation status and questions regarding the profile of strengths and limitations that such scores may represent. The paper expands on previous reports in the literature (e.g., Haley, Ludlow and Coster, 1993) that have focused on interpretation of results from an individual clinical assessment by focusing on use of variable maps to interpret data from a group of individuals.

The examples used for this illustration are drawn from research conducted during the standardization of a new functional assessment for elementary school students with disabilities, the *School Function Assessment* (SFA) (Coster, Deeney, Haltiwanger, and Haley, 1998). After an overview of the instrument and its features, the paper will describe two different applications of SFA variable maps during an outcome prediction study. Other rehabilitation research and clinical contexts where use of variable maps could enrich interpretation of results will also be presented.

General Background

The focus of the study from which these examples are derived was to identify a set of predictors that accurately classified students with disabilities into two groups: those with high and low levels of participation in the regular elementary school program. The analysis method chosen to address this question was the non-parametric Classification and Regression Tree (CART) or recursive partitioning approach (Breiman, Friedman, Olshen, and Stone, 1993). This method was chosen over multivariate regression or discriminant analysis methods because it allows greater examination of outcomes at the level of individuals.

Instrument

The *School Function Assessment* examines three different aspects of elementary school functioning: level of participation, need for supports, and functional activity performance across six major activity settings, including classroom, playground, transportation, transitions, mealtime, and bathroom. It is a criterion-referenced, judgement-based instrument whose primary purpose is to assist the student's special education team to identify important functional strengths and limitations in order to plan effective educational programs, and to measure student progress after the implementation of intervention or support services.

The present examples involve two sections of the SFA: Part I (Participation) and Part III (Activity Performance). The Part I Participation scale examines the student's level of active participation in the important tasks and activities of the six major school settings listed above. Ratings for each setting are completed using a 6 point scale where each rating represents a different profile of participation: 1= Extremely limited; 2 = Participation in a few activities; 3 = Participation with constant supervision; 4 = Participation with occasional assistance; 5 = Modified full participation; 6 = Full Participation. Ratings are summed to yield a total Participation raw score. Raw scores are converted to Rasch measures (estimates) and then transformed to scaled scores on a 0-100 continuum. The scaled scores are interpreted in a criterion-referenced (as compared to norm-referenced) manner.

Part III, the Activity Performance section, consists of 18 independent scales, each of which examines performance of related activities within a specific task domain (e.g., Travel, Using Materials, Functional Communication, Positive Interaction). There are between 10 and 20 items in each scale. Each item is scored on a 4-point rating scale based on the student's typical performance of the particular activity: 1 = Does not/cannot perform; 2 = Partial performance (student does some meaningful portion of activity); 3 = Inconsistent performance (student initiates and completes activity, but not consistently); and, 4 = Consistent performance (student initiates and completes activity to level expected of typical same grade peers). Item ratings are summed to yield a total raw score for the scale, which are then converted to Rasch measures (estimates) and transformed to scaled scores on a 0-100 continuum. Like the Part I scores, these scaled scores are also interpreted in a criterion-referenced manner.

All scales of the SFA were developed using the Rasch partial credit model (BIGSTEPS; Linacre and Wright, 1993). Final score conversion tables were derived directly from the item measure information. Scores were transformed onto a 0 to 100 continuum for greater ease of use (Ludlow and Haley, 1995). The data presented in this paper involved the Standardization edition of the SFA. The final, published version of the SFA is identical to the Standardization version except for two items that were dropped from Part III scales because of serious goodness-of-fit problems. Psychometric information on the SFA is detailed in the *User's Manual* (Coster et al, 1998). Studies have provided favorable evidence of internal consistency and coherence, as well as stability of scores across assessment occasions (test-retest *r*'s > .90).

Participants

The sample from which the current data were obtained consisted of 341 elementary school students with disabilities with a mean age of 9.0 years. Approximately 65% were boys and 35% were girls. Data were collected from 120 public school sites across the United States, which included a representative mix of urban, suburban, and rural sites as well as racial/ethnic groups. Forty-six percent of the students were identified as having a primary physical impairment (e.g., cerebral palsy, spina bifida) and 54% were identified as having a primary cognitive/behavioral impairment (e.g., autism, mental retardation, ADHD).

Students were identified by school personnel, following general guidelines established by the project coordinator. The major concern during sample selection was to maximize diversity in the sample, in terms of school location and clinical diagnosis, in order to assess whether the scales were relevant and appropriate for all geographic regions and for students with different types of functional limitations. Hence, only a small number of students from each individual school were included. Since diversity of participants was the most essential requirement, and normative standards were not being established, random selection was not deemed feasible.

Procedure

The data collection and sample selection were conducted by volunteer school professionals. Because the authors' priority was to create an instrument that could be applied readily in typical school situations, it was designed to require no special training in administration or scoring. Participants were asked to rely on the description and instructions in the test booklet to understand both the items and the rating scales. Because the SFA is so comprehensive, it is unlikely that any one person would have all the information required to complete all scales. Typically, two or more persons who worked with the student, often a teacher and therapist, were involved as respondents. No additional instructions were given about how this collaborative effort should be conducted.

Application Example 1: Setting a criterion to create a dichotomous variable

Rehabilitation outcome studies like the one described here often involve dichotomous variables, e.g., examination of factors associated with good versus poor treatment outcome. Outcome group criteria can be established in a variety of ways. Sometimes there is a specific definition of what constitutes "good" versus "poor" outcome for a particular group. For example, in outcome studies of stroke rehabilitation, good outcome may be meaningfully defined as "discharge to the community" and poor outcome as "discharge to nursing home". In other circumstances, however, the outcome of interest is measured on a continuous scale, and the researcher must then decide how to create the dichotomous split. This situation arises almost any time that a functional scale is used to examine patient performance because most of these scales are continuous. A variety of methods can be used to split the groups in this situation, for example doing a median split, or splitting subjects into those above or below the mean. The drawback to such methods is that the selected dividing point may be statistically meaningful, but may or may not have any real world meaning. That is, members of the two groups may not necessarily differ in ways that are congruent with our clinical understanding of good and poor outcomes. For scales developed using IRT methodology, variable maps offer an alternative approach.

In the present analysis, the SFA Part I Participation total score was chosen as the outcome variable since this score reflected the students' overall degree of success in achieving active participation across the six different school environments. This variable needed to be dichotomized in order to conduct a classification analysis. To identify a meaningful split point, the variable map for the Participation scale was examined in conjunction with the definitions of the rating categories for that scale.

The rating definitions suggested a logical split between students whose ratings were between 1 and 3 (i.e., those who needed intensive levels of physical assistance or supervision in order to perform important school tasks), and those with ratings between 4 and 6 (i.e., those who were able to do many or all school tasks without assistance).

One option in this situation would have been to use a frequency count to classify the participants, for example, by setting the criterion that all students who achieved at least four ratings of "4" or better would be put in the "good outcome" group. This approach, however, would assign equal weight to all the school settings, ignoring information from IRT analyses indicating that the settings present different degrees of difficulty in their demands. Such an approach, in fact, defeats the major purpose of using IRT to construct summary scores that reflect actual item difficulty.

A sounder alternative is to use the variable map from the Participation scale, which is reproduced in Figure 1. By looking at the map, one could identify the specific summary score (transformed score) that best represented the performance profile for what was considered "high participation". Because the variable maps reflect the information on relative item and rating difficulty, one could decide the settings in which it was most important for participants to have achieved a minimum rating of "4" in order to be included in the "high participation" group. Two possibilities are illustrated in Figure 1. In the first, a lower criterion is set (dotted line). This criterion reflects a judgement that when the student's total score indicates there are at least two settings in which "4" is the expected rating, he or she is assigned to the "high participation" group. A second choice, (solid line), is to set a higher criterion. Here, the final cut score is set at the point where the student would be expected to have ratings of "4" in all settings.

An important point here is that this alternative strategy for cut-score determination rests on the researcher's *clinical understanding* of which profile best represents the outcome of interest for the particular study. Thus, rather than rely on more arbitrary groupings based on traditional statistical evidence such as frequency counts or the midpoint of a distribution, the researcher can set a criterion that incorporates information regarding the difficulty of the items and the desired level of performance defined by the positive and less positive outcomes selected.

Application Example 2: Interpreting the meaning of results

The focus of the study used for these examples was to identify predictors of good and poor outcomes, with an ultimate aim of understanding path-

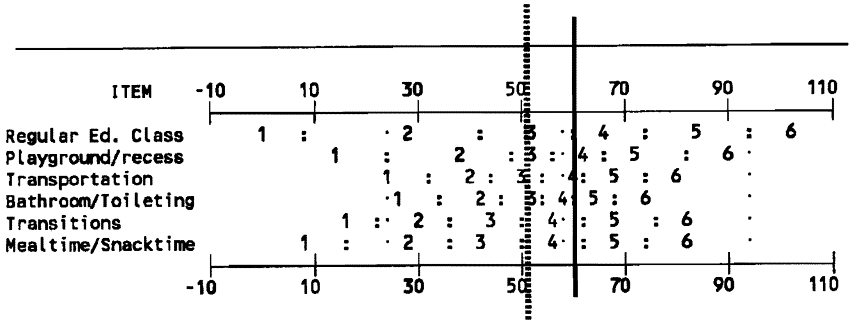


Figure 1. Participation Variable Map (6 school setting items). Adapted from the *School Function Assessment.* Copyright© 1998 by Therapy Skill Builders, a division of The Psychological Corporation. Reproduced by permission. All rights reserved.

Note: The two lines indicate potential cut-off points to dichotomize the Participation variable. For the first option, a lower scaled score is selected (dotted line) at the point where a student is expected to have a rating of "4" (Participation with occasional assistance) in at least two settings. Once a student achieves this score, he or she will be assigned to the "high participation" group. For the second option (solid line), a higher criterion is set by choosing a cut-off score at the point where the student would have an expected rating of at least "4" in all settings.

ways to successful participation and the variables that help identify those students most likely to benefit from rehabilitation services. Because the ultimate goal was to generate outcome predictions for individuals, a Classification and Regression Tree (CART) analysis approach (Breiman, Friedman, Olshen, and Stone, 1993) was chosen. A full discussion of CART is beyond the scope of the present paper, howěver the essential features will be described as needed to understand this application example. A complete discussion of the present study and its results is found elsewhere (Mancini,1997).

CART is a non-parametric multivariate procedure that can be used to classify individuals into specified outcome categories. The analysis proceeds through a series of binary recursive partitions or splits to select from a larger set of predictor variables those that, considered in sequence, provide the most accurate classification of the individuals in the sample. This method is particularly helpful to identify interactions among predictor variables in situations where such interactions are likely but there is limited previous research to guide the analysis. (See Falconer, Naughton, Dunlop, Roth, Strasser, and Sinacore, 1994, for an application of CART methodology in the analysis of stroke rehabilitation outcomes).

Another valuable feature of CART is that, for each predictor variable that is selected, all potential divisions or values of the variable are tested to identify the one that best separates individuals into the groups of interest. This cut-point can be useful for clinicians who want to use the CART decision tree to identify persons likely to have better or worse outcomes in a similar clinical context. However, for researchers interested in understanding the pathways to different outcomes, it would be helpful if there were some means to ascertain *why* a particular cut-point might have been selected. That is, what distinguishes the performance of persons above and below that score level? This is the situation where, if scores are IRT-based measures, application of the relevant variable maps may prove very useful.

In the present example, a CART analysis was conducted to examine predictors of school participation, using the SFA Participation variable,

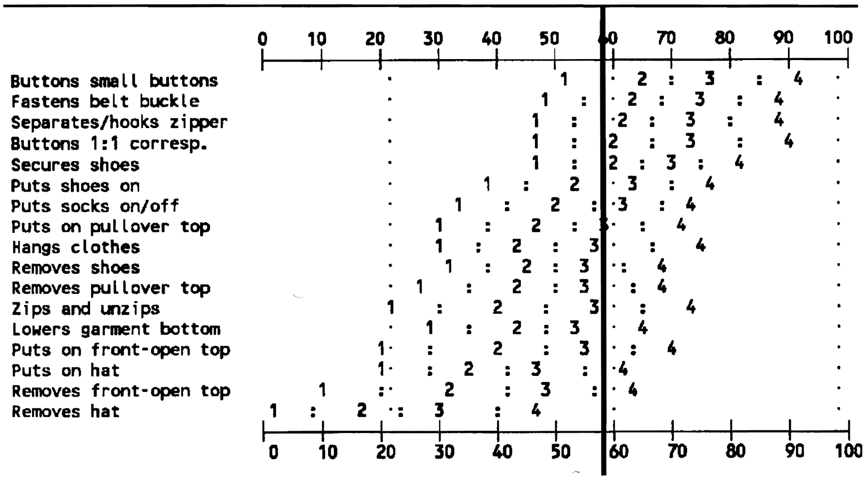


Figure 2. Clothing Management Variable Map (17 functional activity items). Adapted from the *School Function Assessment.* Copyright© 1998 by Therapy Skill Builders, a division of The Psychological Corporation. Reproduced by permission. All rights reserved.

Note: The solid line indicates the cut-point of 59 selected by the CART analysis. The expected functional activity performance profile of students with scores below 59 is described by the ratings to the left of this line; that of students above the cut-point is described by the ratings to the right of the line. Results indicate that students may be classified into the "high participation" group even though they may still be expected to have difficulty (i.e., as indicated by ratings of 1, "Does not perform" or 2, "Partial performance") on fine motor activities such as buttoning buttons.

dichotomized as described above, as the outcome measure and SFA Part III scaled scores as predictors. The specific research question was: which of the functional tasks examined in Part III are most informative in predicting high or low participation? The first variable selected by the program was the Clothing Management score, with a cut point set at a score of 59. Although this result seemed counter-intuitive, given the outcome focus on school participation, a review of the content of this scale suggested that it was probably selected because it measures functional performance in activities that require a variety of both gross motor and fine motor skills. Thus, the variable may be serving as an indirect measure of severity of physical disability (for more detailed discussion see Mancini, 1997). However, application of the variable map for this scale provided more insight into the result.

The variable map for this scale is reproduced in Figure 2, and the selected cut point is identified with the solid vertical line. The line identifies the expected pattern of item performance associated with scores above (to the right) and below (to the left) of 59. Review of the items indicates that students with scores below 59 would be expected to have difficulty with the basic gross motor and postural control aspects of dressing. The profile suggests that these students are generally not able to consistently initiate and/or complete (ratings <3) lower body dressing activities such as raising and lowering pants or putting on and taking off shoes and socks, nor can they do any significant portions of manipulative tasks (ratings of 1). In contrast, students with scores above 59 manage the gross motor aspects of dressing relatively well, although many would still be expected to have difficulty with fine motor aspects such as doing fasteners (e.g., expected ratings of 2).

This more in-depth interpretation of results supported by the variable map is valuable on several counts. First, examining the items on either side of the cut-point helped confirm the initial interpretation that the scale was selected because it provided an indirect measure of severity of mobility limitations. This result makes clinical sense since severe mobility restrictions can pose significant challenges to active participation in school across a wide variety of contexts. On the other hand, the results also suggested that limitations in the performance of manipulative activities did *not*, on their own, significantly predict school limitations. The latter result is somewhat surprising, given the number and variety of fine motor activities (e.g., writing and other tool use, eating activities) typically expected during the school day. This result invites further research into the question of what the minimum threshold of functional performance for school-related fine motor activities may be, and the degree to which limitations in this area may be accommodated successfully through various adaptations. This more precise definition of questions for future research would not have been possible without the additional information provided through examination of IRT variable maps.

Discussion

This paper has presented two illustrations of research applications of variable maps: to guide definition of outcome groups and to obtain a more in-depth understanding of results. Both types of applications are relevant to a variety of other clinical and research situations where IRT-based scales are used. For example, clinical facilities and researchers analyzing patient outcomes may face similar questions of how best to describe positive and negative outcome groups. If they are using IRT-based measures for which variable maps are available, they can follow a similar rational procedure for deciding which cut-off point is most appropriate, given their questions.

The more common applications are those in which variable maps can enhance understanding of particular scores related to research or clinical outcomes. For example, rather than simply reporting the mean functional performance score for a group of patients discharged from an intervention program, one could use the variable map to describe the expected functional performance profile associated with that score and the proportion of patients who achieved that level or better. Similarly, descriptive analyses of change over time in a particular patient group would be much more meaningful if the scores at admission and discharge were interpreted, using variable maps, in terms of the expected functional profiles represented by each score. For example, consider an intervention program that enabled a substantial portion of students to achieve scores above 59 on the Clothing Management scale used in the previous example. In addition to reporting the average amount of change in the scaled score after intervention, one could also describe the increased level of independent function that this change represents, i.e., that these students are now able to manage basic dressing tasks more consistently on their own, implying less need for assistance from the teacher. In this context, interpreting results using variable map information on the expected functional performance profiles associated with pre and post scores can help to resolve debates over whether particular "statistical" differences in scores represent "clinically meaningful" change.

References

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1993). *Classification and regression trees.* NY: Chapman and Hall.

Coster, W.J., Deeney, T.A., Haltiwanger, J.T., and Haley, S.M. (1998). *School Function Assessment.* San Antonio, TX: The Psychological Corporation/ Therapy Skill Builders.

Falconer, J.A., Naughton, B.J., Dunlop, D.D., Roth, E.J., Strasser, D.C., and Sinacore, J.M. (1994). Predicting stroke inpatient rehabilitation outcome using a classification tree approach. *Archives of Physical Medicine and Rehabilitation*, *75*, 619-625.

Fisher, W.P. (1993). Measurement-related problems in functional assessment. *American Journal of Occupational Therapy*, *47*, 331-338.

Haley, S.M., Ludlow, L.H., and Coster, W.J. (1993). Clinical interpretation of summary scores using Rasch Rating Scale methodology. In C. Granger and G. Gresham (eds.), *Physical Medicine and Rehabilitation Clinics of North America*: Volume 4, No. 3: New developments in functional assessment. (pp. 529-540). Philadelphia: Saunders.

Linacre, J.M., and Wright, B.D. (1993). *A user's guide to BIGSTEPS: Rasch-model computer program.* Chicago: MESA Press.

Ludlow, L.H., Haley,S.M. and Gans, B.M. (1992). A hierarchical model of functional performance in rehabilitation medicine. *Evaluation and The Health Professions*, *15*, 59-74.

Ludlow, L.H. and Haley, S.M. (1995). Rasch model logits: Interpretation, use, and transformation. *Educational and Psychological Measurement*, *55*, 967-975.

Mancini, M. C. (1997). Predicting elementary school participation in children with disabilities. Unpublished doctoral dissertation, Boston University.

Wright, B.D., Linacre, J.M., and Heinemann, A.W. (1993). Measuring functional status in rehabilitation. In C. Granger and G. Gresham (eds.), *Physical Medicine and Rehabilitation Clinics of North America* Volume 4, No. 3: New developments in functional assessment:. (pp. 475-491). Philadelphia: Saunders.

JOURNAL OF OUTCOME MEASUREMENT®, *3*(2), 134–161

Measuring Pretest-Posttest Change with a Rasch Rating Scale Model

Edward W. Wolfe
University of Florida

Chris W.T. Chiu
Michigan State University

When measures are taken on the same individual over time, it is difficult to determine whether observed differences are the result of changes in the person or changes in other facets of the measurement situation (e.g., interpretation of items or use of rating scale). This paper describes a method for disentangling changes in persons from changes in the interpretation of Likert-type questionnaire items and the use of rating scales (Wright, 1996a). The procedure relies on anchoring strategies to create a common frame of reference for interpreting measures that are taken at different times and provides a detailed illustration of how to implement these procedures using FACETS.

Requests for reprints should be sent to Edward W. Wolfe, 1403 Norman Hall, University of Florida, Gainesville, FL 32611, e-mail: ewolfe@ufl.edu. This research was supported by ACT, Inc. and a post-doctoral fellowship at Educational Testing Service. The authors thank Carol Myford, Mark Reckase, and Hon Yuen for their input on a previous version of this manuscript. A previous draft of this paper was presented in March of 1997 at the *International Objective Measurement Workshop 9* in Chicago, IL.

Measuring Pretest-Posttest Change with a Rasch Rating Scale Model

Measuring change over time presents particularly difficult challenges for program evaluators. A number of potential confounds may distort the measurement of change, making it unclear whether the observed changes in the outcome variable are due to the intervention or some other effect such as regression toward the mean (Lord, 1967), maturation of participants, or idiosyncrasies of participants who drop out of the program (Cook and Campbell, 1979). When rating scales or assessment instruments are used to measure changes in an outcome variable, additional confounds may be introduced into the evaluation process. For example, participants may improve their performance on an assessment instrument that is used as both a pre-test and post-test because of familiarity with the test items (Cook and Campbell, 1979). Alternatively, when changes are measured with Likert-type questionnaires, participants may interpret the items or the rating scale options differently on the two occasions (Wright, 1996a).

This article describes and illustrates an equating procedure proposed by Wright (1996a) that can be applied to rating scale data to compensate for the latter of these potential confounds to measuring change over time. That is, we describe a method for reducing the effect that changes in participants' interpretations of questionnaire items and rating scale options may have on the measurement of change on the underlying construct. We outline the procedures for making this correction, illustrate how these procedures are carried out, and demonstrate how the employment of these procedures can lead to the discovery of changes that would not be apparent otherwise. By implementing this equating procedure, evaluators can eliminate at least one potential threat to the valid interpretation of changes in attitudes or opinions as measured by Likert-type questionnaires.

Theoretical Framework

In many program evaluation settings, evaluators are interested in measuring changes in the behaviors or attitudes of non-random samples of participants who are drawn from a population of interest. Changes in the measures of the outcome variable are typically attributed to participation in the program in question. Of course, numerous threats to the validity of this inference exist, and each of these threats highlights a potential confound that must be taken into account when designing an evaluation, collecting and analyzing data, and interpreting the results. These threats to the validity of interpretations that are drawn from a program evaluation may relate to *statistical validity* (the accuracy of the statistical inferences drawn about the relationship between the program and the outcome variable), *construct validity* (the accuracy of the inferred relationship between the measurement procedures and the latent construct they are intended to represent), *external validity* (the accuracy of the inferred relationship between the participants and the population that they are intended to represent), or *internal validity* (the accuracy of the theory-based inferences drawn about the relationship between the program and the outcome variable). Methods for avoiding or reducing each of these threats to drawing valid inferences are outlined by Cook and Campbell (1979).

The problem addressed by this article represents one of several potential threats to internal validity. That is, we are concerned with whether observed changes in the outcome variable are truly caused by participation in the program or whether observed changes can be attributed to other variables that are byproducts of the evaluation setting. In a program evaluation, threats to internal validity may arise when changes in participants can be attributed to maturation, changes in participants' familiarity with the measurement instrument, mortality of participants, the procedures used to assign participants to treatments, statistical regression toward the mean, or changes in the measurement instrument rather than the treatment itself. The threat to internal validity that we discuss arises when Likert-type questionnaire items are used to measure attitudinal changes. More specifically, we are concerned with the degree to which changes in the way participants interpret questionnaire items and use rating scales confounds the measurement of changes in attitudes or opinions.

Prior research in this area has shown that participants' interpretations of items or rating scales may change over time and that this is a common concern for those who use questionnaires to measure outcome variables. For example, Zhu (1996) investigated how children's psychomotoric self-efficacy changes over time. In this study, children completed a questionnaire designed to measure the strength of their confidence about their abilities to perform a variety of physical exercises. The results of this study indicated that some of the activities were perceived as being less difficult to perform, relative to the remaining activities, over repeated administrations of the questionnaire. Such differential functioning of items over time threatens the validity of interpretations that might be drawn from the results of Zhu's study.

In order to evaluate changes in persons over time, participants' interpretations of the items and rating scales that are used to measure this change must be stable across multiple administrations of the questionnaire. Only if interpretations of items and rating scales demonstrate such stability can differences between measures of the persons be validly interpreted (Wilson, 1992; Wright, 1996b). To further exacerbate the problem, summated composite scores are not comparable across time when items are added, removed, or reworded; items are skipped by some subjects; or response options change from pre-test to post-test—all problems that are common with multiple questionnaire administrations (Roderick and Stone, 1996). In order to mitigate some of these problems, scaling methods are often used to place measures from different administrations of a questionnaire onto a common scale.

Rasch Rating Scale Model

The Rasch Rating Scale Model (RSM) is an additive linear model that describes the probability that a specific person (n) will respond to a specific Likert-type item (i) with a specific rating scale category (x) (Andrich, 1978). The mathematical model for this probability (Equation 1) contains three parameters: the person's *ability* (β_n), the item's *difficulty* (δ_i), and the difficulty of each scale step (i.e., the threshold between two adjacent scale levels, x and x-1) (τ_j). Calibration of questionnaire data to this model results in a separate parameter estimate and a standard error for that estimate for each person, item, and scale step in the measurement context.

$$P(X_{ni} = x) = \frac{\exp \sum_{j=0}^{k} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} \left[\beta_n - \left(\delta_i + \tau_j\right)\right]}, \; x = 0,1,\ldots,m \qquad (1)$$

where, $P(X_{ni} = x)$ is the probability that a person n responds with rating scale category x to item i, which has $m+1$ response options.

The fit of these estimates to the RSM (i.e., model-data fit) can be evaluated using a standardized mean square fit statistic as shown in Equation 2 (Wright and Masters, 1982). When the data fit the RSM, t_e has a mean near zero and a standard deviation near one. Estimates with $|t_e|>2.00$ exhibit poor fit to the model and should be further examined to determine whether there are problems with the scores associated with that particular person, item, or rating scale step.

$$t_e = \left(\sqrt[3]{v_e} - 1\right)\left(\frac{3}{q_e}\right) + \left(\frac{q_e}{3}\right) \qquad (2)$$

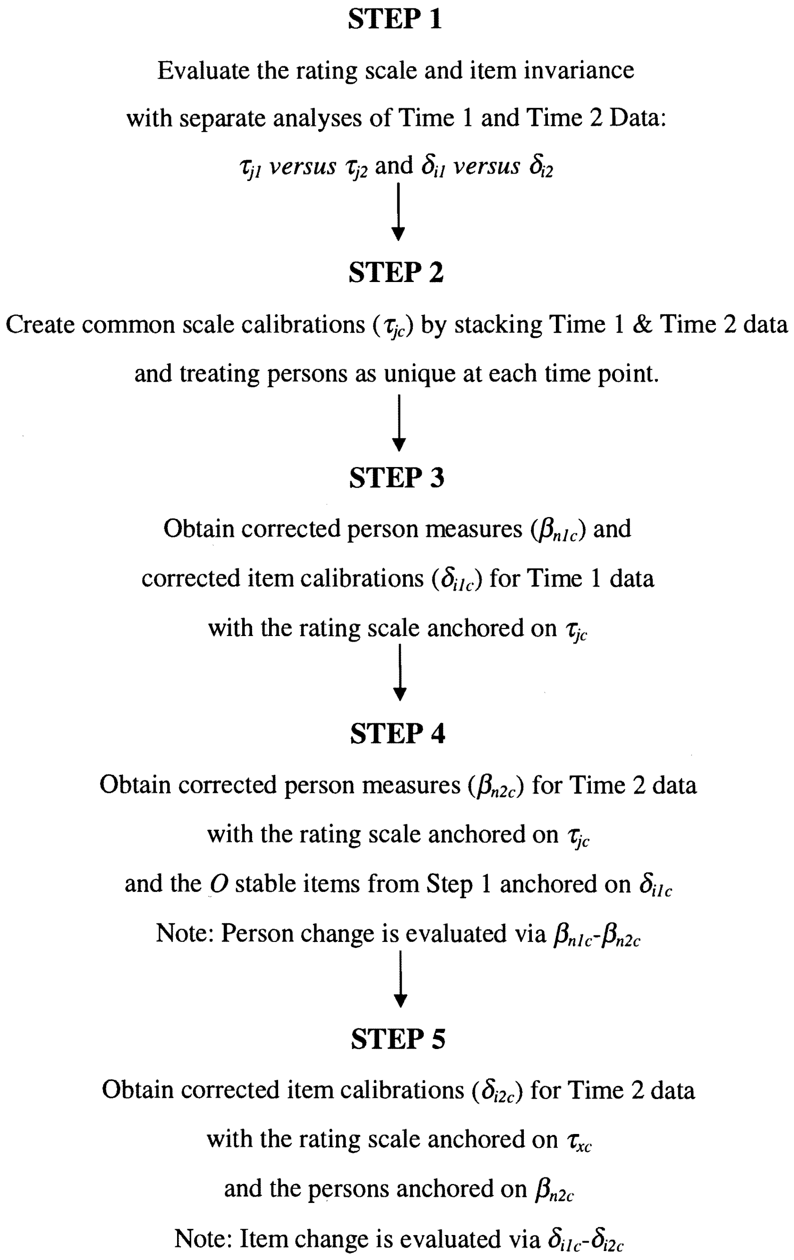
where v_e is the weighted mean of the squared residuals (weighted by their variances) of the observed data from their expected values and q_e is the model standard deviation of the weighted mean square.

An important feature of the RSM is that it allows one to evaluate the extent to which item calibrations are stable across samples of persons or the extent to which person measures are stable across samples of items (i.e., to determine the *invariance* of parameter estimates). This feature is useful when comparing two groups of persons who respond to the same set of items or equating two tests (each composed of different items) that are taken separately by a single group of people. In the present context, invariance evaluation is useful because it allows one to determine the extent to which item calibrations and person measures are stable across two measurement occasions. The stability of two parameter estimates ($\hat{\theta}_1$ and $\hat{\theta}_2$) that are obtained on different occasions is evaluated by examining the standardized difference (Equation 3) between the two estimates (Wright and Masters, 1982). The standardized differences for a population or item pool that conform to the RSM have an expected value of 0.00 and an expected standard deviation of 1.00. Large departures in observed data from these expected values indicate estimates that are less stable over time than would be expected.

$$z = \frac{\hat{\theta}_1 - \hat{\theta}_2}{\sqrt{\left[SE\left(\hat{\theta}_1\right)\right]^2 + \left[SE\left(\hat{\theta}_2\right)\right]^2}} \qquad (3)$$

Examining Change Over Time with the Rating Scale Model

Measuring change over time requires a stable frame of reference, and differential functioning of items and rating scales disrupts the establishment of such a frame of reference. In order to measure changes in the performance of persons across time, other changes in the measurement framework must be eliminated or controlled. There are several methods for accomplishing this (Wright, 1996b). For example, facets other than the persons may be assumed to be constant by forcing the elements of each facet to



STEP 1

Evaluate the rating scale and item invariance
with separate analyses of Time 1 and Time 2 Data:
τ_{j1} *versus* τ_{j2} and δ_{i1} *versus* δ_{i2}

STEP 2

Create common scale calibrations (τ_{jc}) by stacking Time 1 & Time 2 data
and treating persons as unique at each time point.

STEP 3

Obtain corrected person measures (β_{n1c}) and
corrected item calibrations (δ_{i1c}) for Time 1 data
with the rating scale anchored on τ_{jc}

STEP 4

Obtain corrected person measures (β_{n2c}) for Time 2 data
with the rating scale anchored on τ_{jc}
and the *O* stable items from Step 1 anchored on δ_{i1c}
Note: Person change is evaluated via β_{n1c}-β_{n2c}

STEP 5

Obtain corrected item calibrations (δ_{i2c}) for Time 2 data
with the rating scale anchored on τ_{xc}
and the persons anchored on β_{n2c}
Note: Item change is evaluated via δ_{i1c}-δ_{i2c}

Figure 1. Steps for creating a frame of reference using Rasch measurement.

remain fixed. Alternatively, facets that exhibit noticeable change from one occasion to another may be assumed to be truly different indicators of the construct in question and may, therefore, be treated as being completely different elements at each time. Finally, a compromise can be achieved between different administrations of an instrument by creating an "average" frame of reference and allowing facets to vary about that average.

The method we describe was originally proposed by Wright (1996a), and this method creates a common frame of reference by assuming that some elements of the measurement situation remain constant and by allowing others to vary over time. Many researchers desire to identify whether differences demonstrated by specific items or persons are large enough to be of importance, and the method presented in this article allows for such a distinction. Once a common frame of reference has been created, differences between person measures or between item calibrations at each measurement occasion can be evaluated by examining the standardized differences of the parameter estimates produced for each occasion. The method is described here as a five step procedure as portrayed in Figure 1.

Step 1: Evaluate Rating Scale and Item Invariance

The first step in using the RSM to measure change over time is to determine whether interpretations of the scale steps and the items are stable across the two measurement occasions. If the item and step calibrations do demonstrate stability over time (i.e., they are invariant), then differences between person measures at the two occasions are valid indicators of changes in persons over time (i.e., they are free from potential confounding due to changes in interpretations of items or uses of rating scales). However, if the scale step and item calibrations are not invariant over time, then the researcher must disentangle the changes in the scale steps, items, and persons to determine which elements of the measurement context are indeed changing (*Steps* 2 through *5*).

To determine whether the scale step or item calibrations are invariant over time, one must generate two data sets—one containing the responses of each person (n) to each item (i) at Time 1 and the other containing the responses of each person to each item at Time 2. The layout of these data sets is shown in Figure 2. Item and step calibrations, as well as person measures, are obtained for each data set separately so that there is a pair of estimates, one for Time 1 and one for Time 2, for each scale step (τ_{j1} & τ_{j2}), each item (δ_{i1} & δ_{i2}), and each person (β_{n1} & β_{n2})

$$Time1:\begin{bmatrix} 1 & 1-I & x_{111} & x_{121}\cdots x_{1I1} \\ 2 & 1-I & x_{211} & x_{221}\cdots x_{2I1} \\ \vdots & \vdots & \vdots & \vdots \;\; \vdots \;\; \vdots \\ N & 1-I & x_{N11} & x_{N21}\cdots x_{NI1} \end{bmatrix}$$

$$Time2:\begin{bmatrix} 1 & 1-I & x_{112} & x_{122}\cdots x_{1I2} \\ 2 & 1-I & x_{212} & x_{222}\cdots x_{2I2} \\ \vdots & \vdots & \vdots & \vdots \;\; \vdots \;\; \vdots \\ N & 1-I & x_{N12} & x_{N22}\cdots x_{NI2} \end{bmatrix}$$

Figure 2. FACETS Data Layout for Step 1: The first column shows the examinee, the second column shows the item range, N is the number of examinees, and I is the number of items.

(where β_{n1} refers to the measure for person *n* at Time 1 and β_{n2} refers to the for person *n* measure at Time 2).

To evaluate item and step calibration invariance over time, one compares the pair of calibrations for each element of these two facets. That is, one compares τ_{j1} and τ_{j2} for each step level and compares δ_{i1} to δ_{i2} for each item. This comparison can be made using the standardized difference of the two estimates (Equation 3). Items or scale steps that exhibit large differences between their Time 1 and Time 2 calibrations (e.g., $|z|> 2.00$) are not invariant over time (i.e., they are unstable). Such differences between the way the scale steps were used or the items were interpreted at each occasion may confound any inferences that are drawn based on observed differences in the person measures for Time 1 and Time 2, and the researcher must make corrections (*Steps 2* through *5*). For now, note that there are *O* invariant items. If there are no large differences between step and item calibrations from the two occasions, then it is safe to interpret the differences between the person measures from the two occasions as indicators of change in persons over time. Again, this can be done by examining the standardized differences (Equation 3) between the two measures for each person. (β_{n1} & β_{n2}).

Step 2: Create Common Scale Calibrations

If the analyses in *Step 1* reveal that any of the step or item calibrations are not stable across time, then there is a need to constrain this variability before interpreting observed changes in person measures. In order to measure change, one must assume that at least one facet of the measurement context remained stable. The current method relies on the assumption that the rating scale remained stable by using the average value for the scale steps to create the measurement framework. Thus, this method assumes that a common underlying, equal-interval scale adequately portrays the data and that departures from that underlying scale are due only to random fluctuations.

Therefore, the second step in measuring change over time is to create common step calibrations so that person measures and item calibrations from Time 1 and Time 2 can be compared on a common underlying rating scale. To accomplish this, we allow persons to float from Time 1 to Time 2, and items are assumed to be invariant from Time 1 to Time 2. That is, persons are treated as being different objects of measurement on each of the two occasions. This means that the two data sets from *Step 1* must be reconfigured by assigning two unique identifiers to each person—one for Time 1 responses (*n*.1) and one for Time 2 responses (*n*.2) and appending them (i.e., stacking them to create a single data set). The format of the reconfigured data is shown in Figure 3.

$$Stacked:\begin{bmatrix} 1.1 & 1-I & x_{1.111} & x_{1.121} & \cdots & x_{1.1I1} \\ 2.1 & 1-I & x_{2.111} & x_{2.121} & \cdots & x_{2.1I1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ N.1 & 1-I & x_{N.111} & x_{N.121} & \cdots & x_{N.1I1} \\ 1.2 & 1-I & x_{1.212} & x_{1.222} & \cdots & x_{1.2I2} \\ 2.2 & 1-I & x_{2.212} & x_{2.222} & \cdots & x_{2.2I2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ N.2 & 1-I & x_{N.212} & x_{N.222} & \cdots & x_{N.2I2} \end{bmatrix}$$

Figure 3. FACETS Data Layout for Step 2: The first column shows the examinee, the second column shows the items, N is the number of examinees, and I is the number of items.

This stacked data set is analyzed to obtain step calibrations that are consistent with person performance and item functioning across both occasions. The values of these common scale estimates (τ_{jc}) are used in *Steps* 3 through *5* as anchors for the scale steps. Analysis of the stacked data set also produces a single set of item calibrations and two separate measures for each person—one portraying the person at Time 1 and another portraying the person (as a different person) at Time 2. These item calibrations and person measures are ignored. Note that the calibrations from this step for the items that were identified as showing instability over time in Step 1 should show greater misfit to the RSM (Wright, 1996b).

Step 3: Correct the Time 1 Estimates

Once a common rating scale has been created for the two occasions, that scale is used as a frame of reference for the Time 1 and Time 2 data sets. In *Step* 3 of the procedure, the Time 1 data are re-analyzed using the step calibrations from *Step 2* (i.e., τ_{jc}) as anchors for the rating scale. This results in two sets of estimates: a) corrected person measures for all persons (β_{n1c}), and b) corrected item calibrations for all items (δ_{i1c}). These estimates are referenced to the common scale that was created in the *Step 2* analyses, and they are used as the basis for measuring change in *Steps 4* and *5*.

Step 4: Correct the Time 2 Person Measures

In *Steps 2* and *3*, a frame of reference was created for interpreting changes in person measures at Time 2 by creating a rating scale that is common to both occasions and determining the corrected Time 1 person measures and item calibrations. In *Step 4*, the Time 2 data are re-analyzed by anchoring the steps on the common-scale values obtained in *Step 2* (i.e., τ_{jc}) and anchoring the *O* invariant items from *Step 1* on the corrected item calibrations from *Step 3* (δ_{o1c}). The *I-O* items that were found to be unstable from one occasion to the next in *Step 1*, however, are not anchored (i.e., they are allowed to float).

The *Step 4* analyses produce corrected person measures at Time 2 (β_{n2c}) that are referenced to a rating scale that is common to both Time 1 and Time 2 and a set of items that are invariant across time. Any differences between these corrected person measures and the corrected measures obtained in *Step 3* (β_{n1c}) indicate changes in persons, rather than interpretations of items or uses of the rating scale, over time. For each person, the corrected Time 1 measure (β_{n1c}) and the corrected Time 2 measure (β_{n2c}) can be compared using the standardized difference as shown in Equation 3. Persons that exhibit large variability (e.g., $|z| > 2.00$) have changed over time. The analysis also produces calibrations for the *I-O* unstable items (i.e., the items that were allowed to float—$\delta_{(I-O)2}$). These calibrations are ignored.

Step 5: Correct the Time 2 Item Calibrations

The final step in the procedure is to determine the extent to which item functioning changed over time while controlling for changes in person measures. In *Step 5*, the Time 2 data are re-calibrated by anchoring the scale steps on the joint calibrations obtained in *Step 2* (τ_{jc}) and anchoring the person measures on the corrected Time 2 estimates from *Step 4* (β_{n2c}). All items are allowed to float. This analysis results in item calibrations (for all items) at Time 2 (δ_{i2c}) that are corrected for changes in both the interpretation of the rating scale and the performance of people. To determine how much item functioning changed across occasions, the corrected Time 1 item calibrations (δ_{i1c}) are compared to the corrected Time 2 item calibrations (δ_{i2c}). The comparison can be made by computing the standardized differences between these two estimates (Equation 3). This comparison is free from potential confounds due to changes in the use of the rating scale or the performance of persons across time. It is important to note that calibrations for items that were found to be unstable over time in the *Step 1* analyses have been treated as different items in the estimation of person measures regardless of how much their corrected calibrations differ.

Example

The remainder of this article illustrates how this procedure can be applied to the measurement of change in questionnaire data. We demonstrate this technique on data that are typical of many program evaluations (i.e., *single group, pretest, intervention, posttest*). Our analyses emphasize how using the procedure results in different interpretations of how persons and items change over time.

Participants

The data for our illustration come from mathematics, science, and language arts teachers from 14 public and private secondary schools in different regions of the United States. These teachers participated in a nine-month program designed to help them develop portfolio assessments. Approximately 12 teachers from each school participated in the program (n = 168). At the beginning of the school year (in September), teachers responded to a questionnaire designed to assess the strength with which teachers perceive potential barriers to the implementation of a portfolio assessment program to be problematic (Wolfe and Miller, 1997). After participating in the program for an academic year (in June), teachers completed the questionnaire a second time. A comparison of a teacher's responses from September (Time 1) with the responses provided in June (Time 2) was interpreted as a measure of change in the teacher's perception of barriers to the implementation of portfolio assessments. Complete data for Time 1 and Time 2 were available for 117 of the 168 teachers who participated in the program (a 30% attrition rate).

Instrument

The questionnaire asked teachers how problematic they perceived 30 potential barriers to the implementation of a portfolio assessment system to be. The barriers referenced issues such as the amount of *time* required to use portfolios, resistance from *people* to the idea of using portfolios, the difficulty of assigning *scores* to portfolio entries, changes in *instruction* that are required when portfolios are used, and the availability of *resources* for using portfolio assessment. Each barrier was formatted as the stem for a four-point Likert-type item. Teachers responded to each barrier by indicating whether the barrier is a(n) *unlikely*, *minor*, *difficult*, or *serious* problem. For each of the 30 barriers, teachers indicated the option that best describes the difficulty of that specific barrier. *Unlikely* problems were defined as those that would likely have no impact on the teacher's use of portfolios. *Minor* problems were those that may cause the teacher to use portfolios differently than they would be used in an ideal situation. *Difficult* problems were defined as problems that may cause the teacher to reconsider using portfolios in his or her classroom. *Serious* problems were those that would cause the teacher not to use portfolios at all.

Analyses and Results

These data were analyzed with a Rasch RSM. For substantive meaning, all facets were scaled so that higher logit values were associated with more difficult portfolio implementation. That is, higher values of teacher measures were associated with the perception of portfolio implementation as being more difficult, and higher values of barrier and rating scale step calibrations were associated with barriers that are more difficult to overcome. In each of the following sections, we detail the steps of the anchoring method described by Wright (1996a). Prior to illustrating the five steps, however, the fit of the data to the RSM is evaluated.

Evaluating Fit

For *Step 1*, the data were placed in two data sets—one containing the teachers' responses from September (Time 1) and the other containing teachers' responses from June (Time 2) (see Figure 2). Each data set contained three variables: a) teacher (person) identifier, b) barrier (item) identifier, and c) the teacher's response (rating) to that barrier. In our description, we use FACETS (Linacre, 1989) to obtain parameter estimates for these data sets. It should be noted, however, that these analyses can be performed using any item response software that allows for the analysis of rating scale data and the anchoring of measurement facets. The two data sets, one from each of the two occasions, were calibrated on separate FACETS analyses. An example FACETS command file for performing *Step 1* on the September data is shown in Appendix A. A similar command file is written for the June data. These analyses result in two sets of barrier calibrations and teacher measures—one for the September data and one for the June data.

To evaluate the fit of the data to the model, the standardized mean square fit statistics (Equation 2) for the parameter estimates of each teacher and barrier were examined at each occasion. Teachers and barriers with fit statistics greater than two were flagged as potential problems. However, no teachers or barriers were eliminated from the analyses based on misfit because inspection of their response patterns revealed no conspicuous anomalies. Sixteen of the teachers (14%) showed poor fit to the model in the September data and twelve (10%) showed poor fit in June. Three of the barriers in our questionnaire (10%) had large fit statistics for the September data and only one (3%) showed poor fit in June.

Step 1: Evaluate Rating Scale and Barrier Invariance

As described in the previous section, in *Step 1* the September and June responses were analyzed separately so that each teacher, barrier, and rating scale step received a pair of parameter estimates—one for September and one for June. The pair of estimates for each teacher, barrier, and rating scale step are referred to here as β_{n1} and β_{n2}, δ_{i1} and δ_{i2}, and τ_{j1} and τ_{j2}, respectively. In subsequent sections, these estimates will also be referred to as *uncorrected* estimates. To determine whether differences between β_{n1} and β_{n2} are valid indicators of change in teacher measures over time, we computed the standardized differences (Equation 3) between each pair of step calibrations (τ_{j1} and τ_{j2}) and each pair of barrier calibrations (δ_{i1} and δ_{i2}). The parameter estimates for September and June, their standard errors, and the associated standardized differences are shown in Table 1 and Figure 4.

The analyses from *Step 1* reveal that there are large differences in the way that the rating scale steps were used at September and June as indicated by the large standardized difference for two of the three scale

Table 1

Rating Scale Step Calibrations from Step 1 for September and June

Scale Step	**τ_{j1} Logit**	**τ_{j1} Error**	**τ_{j2} Logit**	**τ_{j2} Error**	**z**
0 to 1	-1.82	0.04	-1.24	0.04	-10.25
1 to 2	0.05	0.05	0.12	0.05	-0.99
2 to 3	1.77	0.10	1.13	0.08	5.00
Mean	**0.00**	**0.06**	**0.00**	**0.06**	**-2.08**
(SD)	**(1.80)**	**(0.03)**	**(1.19)**	**(0.02)**	**(7.68)**

Note: τ_{j1} represents the rating scale step calibrations obtained in *Step 1* for September, and τ_{j2} represents the rating scale step calibrations obtained in *Step 1* for June. $|z|>2.00$ is considered large enough to indicate unstable uses of rating scale steps across occasions.

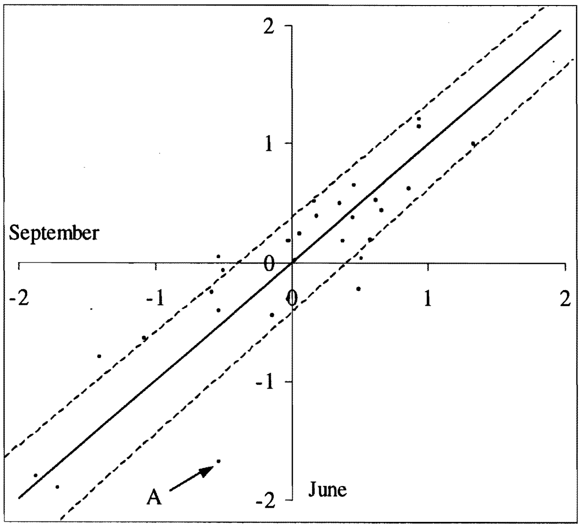


Figure 4. Scatter plot of uncorrected September and June barrier calibrations.

step calibrations (Table 1). Specifically, the difference between September and June scale steps for the 0 to 1 and the 2 to 3 transitions are so much larger than their standard errors that their invariance across time is suspect. Furthermore, several of the barriers showed unexpectedly large changes in their calibrations over the two administrations of the questionnaire. In fact, 7 of the 30 barriers (23%) have absolute standardized differences greater than 2.00 (i.e., $O = 23$) as evidenced by the points falling outside of the 95% confidence bands shown in Figure 4. This is a large percentage of barriers when compared to the expectation derived from the standard normal distribution (about five percent). The largest change in barrier calibrations, indicated as point A on the scatterplot, was 4.49 standard errors (from –0.58 to –1.68 logits). In addition, the standard deviation of the standardized differences (1.80) is considerably larger than the expected value of 1.00. These statistics suggest that differences in the functioning of barriers and rating scale steps over time may cloud any interpretations that we make of differences in teacher measures, so our example proceeds with *Steps 2* through *5* of Wright's (1996a) correction procedure.

Step 2: Correct the Scale Calibrations

In *Step 2* of the procedure, a common rating scale is created so that teacher attitudes and barrier severity estimates from September and June can be estimated in a common frame of reference. To this end, we stack the two data sets from *Step 1*, reassigning teacher identifiers to each teacher for the June responses (as shown in Figure 3). In our example, we simply added 1000 to the original identifier (as shown in the example FACETS command file presented in Appendix B). Because this step of the analysis portrays each teacher as being a different person in June than in September and allows barriers to remain stable across administration of the questionnaire, the output of this command file results in a pair of measures for each teacher and a single calibration for each barrier. All of these values are ignored. The rating scale step calibrations (τ_{jc}) from this analysis, however, are of interest and will be utilized as anchor values for the remaining steps of the procedure. Table 2 compares the scale step calibrations from *Step 1* of the procedure to those obtained from *Step 2*. As one would expect, the values from *Step 2* (i.e., the step calibrations for the scale that are common to September and June) are between the two values obtained in *Step 1* (i.e., the step calibrations for the separate September and June scales). In addition, because each calibration is based on a larger number

Table 2

Rating Scale Step calibrations from Step 1 and Step 2

Scale Step	τ_{j1} Logit	τ_{j2} Logit	τ_{jc} Logit	τ_{jc} SE
0 to 1	-1.82	-1.24	-1.46	0.03
1 to 2	0.05	0.12	0.10	0.03
2 to 3	1.77	1.13	1.36	0.06
Mean	**0.00**	**0.00**	**0.00**	**0.04**
(SD)	**(1.80)**	**(1.19)**	**(1.41)**	**(0.02)**

Note: τ_{j1} represents the rating scale step calibrations obtained in *Step 1* for September, τ_{j2} represents the scale step calibrations obtained in *Step 1* for June, and τ_{jc} represents the scale step calibrations obtained in *Step 2* for the combined September and June data set (i.e., the common scale).

of observations, the standard errors of these calibrations are smaller than those for the *Step 1* calibrations.

Step 3: Corrected September Estimates

In *Step 3*, corrected estimates are obtained for teachers (β_{n1c}) and barriers (δ_{i1c}) in September by anchoring rating scale steps on the values obtained in *Step 2* (τ_{jc}). Appendix C shows an example FACETS command file for this analysis. Note that the command file is the same as the command file used in *Step 1* with the exception that rating scale steps are now anchored on their τ_{jc} values. The data file is the same one used for the September analysis in *Step 1*. The *Step 3* analyses result in two sets of values for the September data. The corrected teacher measures (β_{n1c}) and the corrected barrier calibrations (δ_{i1c}) are used as the basis for measuring changes in teachers and barriers in *Steps 4* and *5*. These estimates are also referred to as the *corrected* September estimates.

Step 4: Correct the June Teacher Measures

In *Step 4*, the common rating scale step calibrations from *Step 2* (τ_{jc}) and the corrected barrier calibrations obtained in *Step* 3 (δ_{j1c}) for the *O* (23) items that were found to be invariant across time in the *Step 1* analyses (referred to here as δ_{o1c}) are used as anchors so that corrected teacher measures (β_{n2c}) can be estimated for the June data. As shown in Appendix D, the seven barriers that were found to be unstable across time in *Step 1* are not anchored (i.e., they are allowed to float). Note that new calibrations ($\delta_{(I-O)2}$) are obtained for these barriers in *Step 4*, but these values are ignored. Otherwise, the procedures for analyzing the June data are the same as they were in *Step 1*. The resulting teacher measures (β_{n2c}) have been corrected for changes in perceptions of barriers and uses of the rating scale over time through this anchoring process. As a result, a comparison of the corrected June teacher measures (β_{n2c}) with the corrected September teacher measures (β_{n1c}) reveals how people have changed over time without confounding from changes in barrier or rating scale functioning. This comparison can be made by examining the standardized difference (Equation 3) for each teacher's pair of corrected measures.

Comparison of Uncorrected and Corrected Teacher Measures

Figure 5 displays the uncorrected and corrected measures for the 117 teachers in this study. This scatter plot shows that, overall, the standardized differences that were based on the corrected teacher measures were greater than those based on the uncorrected measures. This is evident from the fact that the majority of the points fall above the identity line in Figure 5. That is, teachers appear to have a more positive view as a result of participation in the program when the corrected measures are considered. For example, the teacher represented by point A had a

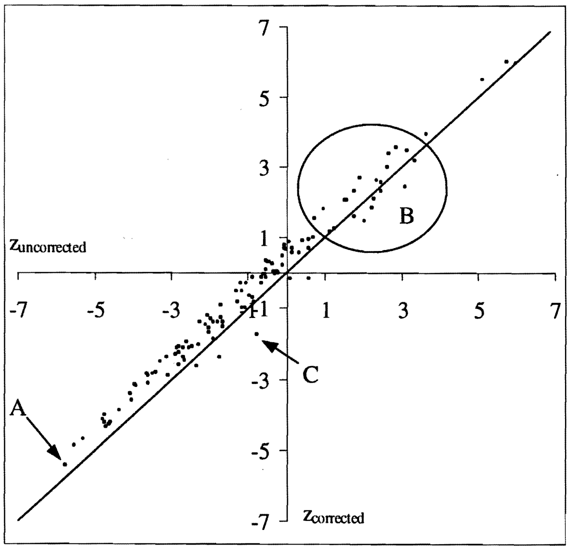


Figure 5. Scatter plot of uncorrected and corrected teacher standardized differences.

$z_{uncorrected}$ = -5.78 and a $z_{corrected}$ = -5.40. Although there were only minor differences in most of the teacher standardized differences that were based on the uncorrected and the corrected measures (r = .99), the correction method made noticeable differences for some teachers. This is particularly true for teachers who had positive standardized differences (i.e., showed a decreased sensitivity to barriers between the pretest and posttest) as evidenced by the fact that the points are more dispersed in the upper right quadrant of Figure 5. As shown, the teachers included in subset B are more dispersed than are other teachers in the sample. As for changes in individual teachers' standardized differences, the largest discrepancy between teacher standardized differences between the uncorrected and corrected teacher measures was 0.97 (i.e., $z_{uncorrected}$ = -0.77 and $z_{corrected}$ = -1.74), indicated in the scatterplot as point C.

Table 3 summarizes how the application of the method for measuring change influenced the distributions of teacher measures. This table shows that the utilization of the correction method resulted in slightly fewer misfitting teachers ($\%_{uncorrected}$ = 12%, $\%_{corrected}$ = 11%). In addition, the correction method also resulted in smaller differences between pretest and posttest measures as evidenced by the mean standardized difference (Mean $z_{uncorrected}$ = -0.94, Mean $z_{corrected}$ = -0.53). As a result, we would draw somewhat different conclusions about the amount of change that teachers exhibited, depending on whether we interpret the uncorrected or the corrected

Table 3

Uncorrected and Corrected Teacher Measure Summary Statistics

Statistic	Uncorrected Measures	Corrected Measures
Number with Fit > 2	29 (12%)	26 (11%)
Mean (z)	-0.94	-0.53
Number with Significant z	57 (49%)	54 (46%)
SD (z)	2.43	2.33

Note: Number with fit > 2 represents the number of teachers with large fit statistics summed across both occasions. Therefore the percent shown is the total number of misfitting teachers divided by 234 (117 × 2). Mean (z) is the average standardized difference across the 117 teachers, and *SD* (z) is the standard deviation of the standardized differences. Number with significant z represents the number of teachers with absolute standardized differences > 2.

measures. This observation is supported by the fact that a smaller percent of the teachers' corrected absolute standardized differences were greater than two ($\%_{\text{uncorrected}}$ = 49%, $\%_{\text{corrected}}$ = 46%). Not only would our interpretation of the amount of change exhibited by the group of teachers change, but so would our interpretation of which teachers changed. Comparison of the standardized differences revealed that different conclusions would be drawn about 13 (11%) of the teachers based on the uncorrected and corrected standardized differences. Of course, the majority of these teachers had smaller standardized differences when their measures were corrected.

Step 5: Correct the June Barrier Calibrations

In *Step 5*, the common rating scale step calibrations from *Step 2* (τ_{jc}) and the corrected person measures for June obtained in *Step 4* (β_{n2c}) are used as anchors so that corrected barrier calibrations can be estimated for the June data. As shown in the example command file in Appendix E, this anchoring is the only difference between the analyses of the June data for *Steps 1* and *5*. The resulting barrier calibrations (δ_{i2c}) have been corrected for changes in teachers and uses of the rating scale over time. The corrected June barrier calibrations can be compared to the corrected calibrations for September (δ_{i1c}) obtained in *Step 3* to identify how the perception of barriers changed over time. As in the previous analyses, this comparison is made by examining the standardized difference (Equation 3) for each barrier's pair of corrected calibrations.

Comparison of Uncorrected and Corrected Barrier Calibrations

Figure 6 displays the uncorrected and corrected calibrations for the 30 barriers in this study. This scatter plot shows that there were larger differences between the standardized differences that were based on the uncorrected and the corrected calibrations than for the teacher measures as evidenced by the wider dispersion of points around the identity line. However, the correlation between corrected and uncorrected standardized differences for barrier calibrations is still strong (r = .93), albeit somewhat inflated by the outlier represented by point A on the scatterplot. This resulted in changes in individual barriers' standardized differences as large as 1.42 (i.e., $z_{\text{uncorrected}} = -1.52$ and $z_{\text{corrected}} = -2.94$, point B), depending on whether the uncorrected and corrected calibrations are considered.

Table 4 summarizes how the application of the method for measuring change influenced the distributions of barrier calibrations. This table

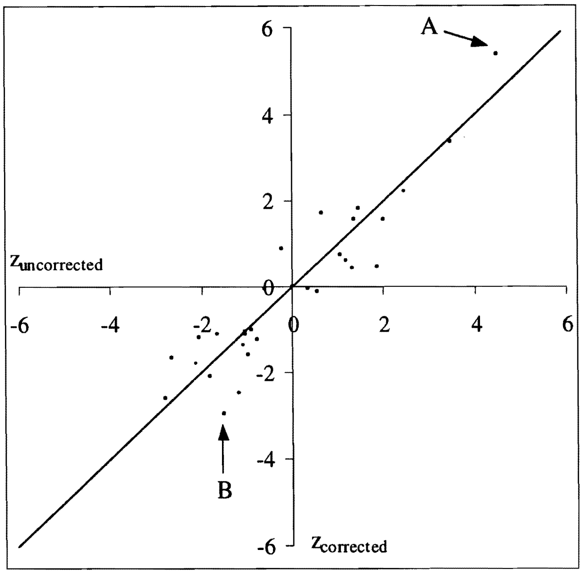


Figure 6. Scatter plot of uncorrected and corrected barrier standardized differences.

Table 4

Uncorrected and Corrected Barrier Calibration Summary Statistics

Statistic	Uncorrected Calibrations	Corrected Calibrations
Number with Fit > 2	4 (7%)	2 (3%)
Mean (*z*)	-0.01	-0.09
Number with Significant *z*	7 (23%)	7 (23%)
SD (*z*)	1.80	1.88

Note: Number with fit > 2 represents the number of barriers with large fit statistics summed across both occasions. Therefore the percent shown is the total number of misfitting barriers divided by 60 (30 x 2). Mean (*z*) is the average standardized difference across the 30 barriers, and *SD* (*z*) is the standard deviation of the standardized differences. Number with significant *z* represents the number of barriers with absolute standardized differences > 2.

shows that the utilization of the correction method resulted in fewer misfitting barriers ($\%_{\text{uncorrected}} = 7\%$, $\%_{\text{corrected}} = 3\%$). In addition, the correction method also resulted in larger differences between pretest and posttest measures as evidenced by the mean standardized difference (Mean

$z_{\text{uncorrected}}$ = -0.01, Mean $z_{\text{corrected}}$ = -0.09). Again, this result would lead us to draw somewhat different conclusions about the nature of change in this study, depending on whether we interpret the uncorrected or the corrected barrier calibrations—with the corrected measures we would say that teachers changed less and barriers changed more than they did when uncorrected values are considered. And, as was true for the teacher measures, we would draw different conclusions about several of the individual barriers based on the uncorrected and corrected standardized differences. Decisions concerning a total of six of the barriers (20%) would be different (half moving from significant to non-significant and half moving from non-significant to significant).

Conclusions

We have illustrated a procedure for removing potentially-confounding sources of variability from the measures of changes in persons over time. Application of this procedure to the data in our example revealed four things about our perception of change in these data that were not apparent when this correction procedure was not applied. First, use of this procedure reduced the perceived overall change exhibited by this group of teachers. That is, depending on whether teacher measures were corrected or not, a different magnitude of change in teacher measures from pretest to posttest was observed. Second, the procedure resulted in changes in the perception of which teachers changed over time. In the example given here, we would draw different conclusions about 11% of the teachers as a result of applying the procedure. Third, the application of the procedure resulted in better fit of the teacher measures to the RSM. By removing the confounding of changes in perceptions of barriers from the teacher measures and fluctuations in the use of the rating scale, we were able to produce better measures of our teachers. Finally, we were able to detect changes in the perceptions of items over time that were not apparent without the correction procedure. In fact, we would draw different conclusions about 20% of the items after applying the correction procedure to our data.

Overall, this procedure seems useful for disentangling changes in the item functioning and rating scale use from changes in person performance when Likert-type questionnaires are used to measure the impact that a program has in participants. As a result, the procedure could prove useful for program evaluators who are interested in measuring changes in attitudes and opinions. We suggest four directions for further exploration of this procedure. One direction would be to determine how well the method can be adapted to multi-faceted measurement contexts or to measurement models based on different response structures (e.g., partial credit models). Often program evaluations are ongoing and involve measuring changes across several consecutive years, so a second direction for future work might involve extending this correction procedure to settings with more than two measurement occasions. Third, it would be interesting to compare this method to other methods for disentangling sources of change in questionnaire data. For example, Chang and Chan (1995) identified four Rasch-based methods for measuring change that are distinct from the one presented in this article. Comparison of these methods would reveal consistencies and differences in the ways that change is depicted. Of course, because there are multiple methods, such a comparison would probably lead to conflicting results. Therefore, we suggest a fourth direction for future research concerning the variety of models for measuring change over time. We believe that any work that is based on these models is incomplete without evidence that the change that is being detected is indeed true change. As a result, we believe that it will be necessary to use simulations to determine the adequacy with which of each of the variety of methods recovers true change over time.

References

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

Chang, W.C., and Chan, C. (1995). Rasch analysis for outcome measures: Some methodological considerations. *Archives of Physical Medicine and Rehabilitation, 76*, 934-939.

Cook, T.C., and Campbell, D.T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin.

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

Lord, F.M. (1967). Elementary models for measuring change. In C.W. Harris (Ed.), *Problems in Measuring Change* (pp. 21-38). Madison, WI: University of Wisconsin Press.

Roderick, M., and Stone, S. (1996, April). Is it changing opinions or changing kids? Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Wilson, M. (1992). Measuring changes in the quality of school life. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (pp. 77-96). Norwood, NJ: Ablex Publishing.

Wolfe, E.W., and Miller, T.R. (1997). Barriers to the implementation of portfolio assessment in secondary education. *Applied Measurement in Education, 10*, 235-251.

Wright, B.D. (1996a). Time 1 to time 2 Comparison. *Rasch Measurement Transactions, 10*, 478-479.

Wright, B.D. (1996b). Comparisons require stability. *Rasch Measurement Transactions, 10*, 506.

Wright, B.D., and Masters, G.N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Zhu, W. (1996, April). Many-faceted rasch analysis of children's change in self-efficacy. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Appendix A. FACETS Command File for *Step 1*

```
Title = STEP 1, TIME 1: EVALUATE SCALE AND BARRIER
INVARIANCE
Output = STEP1T1.OUT,STEP1T1.ANC
Scorefile = STEP1T1.S
Facets = 2
Positive = 1,2
Arrange = m,f
Models =
?,?,LIKERT
*
Rating Scale=LIKERT,R3
0=UNLIKELY
1=MINOR
2=DIFFICULT
3=SERIOUS
*
Labels =
   1,TEACHER
      1 = T1; teacher 1 time 1
      2 = T2; teacher 2 time 1
      .
      .
      .
      117 = T117; teacher 117 at time 1
*
   2,BARRIER
      1 = B1; barrier 1
      2 = B2; barrier 2
      .
      .
      .
      30 = B30; barrier 30
Data = STEP1T1.DAT; Time 1 data configured as shown
in Figure 1
```

Appendix B. FACETS Command File for *Step 2*

```
Title = STEP 2: CREATE COMMON SCALE CALIBRATIONS
Output = STEP2.OUT,STEP2.ANC
Scorefile = STEP2.S
Facets = 2
Positive = 1,2
Arrange = m,f
Models =
?,?,LIKERT
*
Rating Scale=LIKERT,R3
0=UNLIKELY
1=MINOR
2=DIFFICULT
3=SERIOUS
*
Labels =
   1,TEACHER
       1 = T1; teacher 1 time 1
      2 = T2; teacher 2 time 1
      .
      .
      .
      117 = T117; teacher 117 at time 1
       1001 = T1; teacher 1 time 2
      1002 = T2; teacher 2 time 2
      .
      .
      .
      1117 = T117; teacher 117 at time 2
*
   2,BARRIER
       1 = B1; barrier 1
      2 = B2; barrier 2
      .
      .
      .
      30 = B30; barrier 30
data = STEP2.DAT; Time 1 & Time 2 data stacked as in
Figure 2
```

Appendix C. FACETS Command File for *Step 3*

```
Title = STEP 3: CORRECT TIME 1 ESTIMATES
Output = STEP3.OUT,STEP3.ANC
Scorefile = STEP3.S
Facets = 2
Positive = 1,2
Arrange = m,f
Models =
?,?,LIKERT
*
Rating Scale=LIKERT,R3
0=UNLIKELY,0,A; ALWAYS ANCHOR ON 0
1=MINOR,-1.46,A; ANCHOR ON VALUE FROM STEP 2
2=DIFFICULT,0.10,A
3=SERIOUS,1.36,A
*
Labels =
   1,TEACHER
       1 = T1; teacher 1 time 1
      2 = T2; teacher 2 time 1
      .
      .
      .
      117 = T117; teacher 117 at time 1
*
   2,BARRIER
       1 = B1; barrier 1
      2 = B2; barrier 2
      .
      .
      .
      30 = B30; barrier 30
data = STEP1T1.DAT; Time 1 data as shown in Figure 1
```

Appendix D. FACETS Command File for *Step 4*

```
Title = STEP 4: CORRECT TIME 2 TEACHER MEASURES
Output = STEP4.OUT,STEP4.ANC
Scorefile = STEP4.S
Facets = 2
Positive = 1,2
Arrange = m,f
Models =
?,?,LIKERT
*
Rating Scale=LIKERT,R3
0=UNLIKELY,0,A; ALWAYS ANCHOR ON 0
1=MINOR,-1.46,A; ANCHOR ON VALUE FROM STEP 2
2=DIFFICULT,0.10,A
3=SERIOUS,1.36,A
*
Labels =
   1,TEACHER
       1001 = T1; teacher 1 time 2
      1002 = T2; teacher 2 time 2
      .
      .
      .
      1117 = T117; teacher 117 at time 2
*
  2,BARRIER,A
   1 = B1,-.46; barrier 1 (invariant) anchored on
Step 3 value
   2 = B2; barrier 2 (unstable) with no anchor value
   .
   .
   .
   30 = B30,-.51; barrier 30 (invariant) anchored on
Step 3 value
data = STEP1T2.DAT; Time 2 data as shown in Figure 1
```

Appendix E. FACETS Command File for *Step 5*

```
Title = STEP 5: CORRECT TIME 2 ITEM CALBRATIONS
Output = STEP5.OUT,STEP5.ANC
Scorefile = STEP5.S
Facets = 2
Positive = 1,2
Arrange = m,f
Models =
?,?,LIKERT
*
Rating Scale=LIKERT,R3
0=UNLIKELY,0,A; ALWAYS ANCHOR ON 0
1=MINOR,-1.46,A; ANCHOR ON VALUE FROM STEP 2
2=DIFFICULT,0.10,A
3=SERIOUS,1.36,A
*
Labels =
  1,TEACHER,A
    1001 = T1,.34; teacher 1 time 2 anchored on Step
4 value
    1002 = T2,.01; teacher 2 time 2 anchored on Step
4 value
    .
    .
    .
    1117 = T117,-.61; teacher 117 time 2 anchored on
Step 4 value
*
   2,BARRIER
       1 = B1; barrier 1
      2 = B2; barrier 2
      .
      .
      .
      30 = B30; barrier 30
data = STEP1T2.DAT; Time 2 data as shown in Figure 1
```

JOURNAL OF OUTCOME MEASUREMENT®, *3*(2), 162–175

Grades of Severity and the Validation of an Atopic Dermatitis Assessment Measure (ADAM)

Denise P. Charman
Victoria University of Technology

George A. Varigos
Royal Children's Hospital, Victoria, Australia

There has generally been a dearth of good clinical descriptions of grades of disease severity. The aim of this study was to produce reliable and valid descriptions of grades of severity of Atopic Dermatitis (AD). The ADAM (AD Assessment Measure) measure was used to assess AD severity in 171 male and female paediatric patients (mean age=54 months) at the Royal Children's Hospital in Melbourne, Australia. The assessments were subject to Partial Credit analyses to produce clinically relevant "word pictures" of grades of severity of AD. Patterns of AD were shown to vary according to age, sex and severity. These descriptions will be useful for clinical training and research. Moreover, the approach to validation adopted here has important implications for the future of measurement in medicine.

Requests for reprints should be sent to Denise Charman, Victoria University of Technology, P. O. Box 14428, MCMC, Melbourne 8001, Victoria, Australia, e-mail: denisecharman@vut.edu.au

Introduction

In the clinical and research literature, there is a dearth of good clinical descriptions of grades of disease severity which are generally implicit and acquired from an apprenticeship style of clinical training. Under these circumstances it is not surprising that agreement studies have produced less than optimal results (Charman, Varigos, Horne, and Oberklaid, 1999; Sackett, et al., 1991). The extent of agreement, in general, has been moderate only. Moreover, agreement appears to vary such that it has been better with "mild" grades of severity than with "moderate" and "severe" grades (Charman, et al., 1999; Hall, et al., 1987).

Such findings have been noted in studies of agreement on grades of severity of Atopic Dermatitis (AD) (for example, European Task Force on Atopic Dermatitis, 1993). These findings suggest that the bases for clinical judgements of grades of AD severity could be more clearly defined. If definitions are clear, agreement should improve and response biases diminished (Streiner and Norman, 1989; Spiteri, et al., 1988). Moreover, definitions would also help to ensure that a scale or measure is "valid".

It should be noted that there is a web-site (http://www.adserver.sante.univ-nantes.fr/) which does provide photographs for selected morphological features at severity grade levels, 1, 2 and 3. This site is an adaptation from a CD-ROM developed for the European Task Force on AD (European Task Force on Atopic Dermatitis, 1997). However, the clinical photographs and the morphological features were derived from consensus: They are not necessarily statistically reliable and valid. This point is emphasised by the reported 54% agreement on grade 1 for one of the morphological features (edema/populations) and yet, this feature was retained (presumably) because of clinical commitment. The overall SCORAD AD severity score is calculated from a weighted linear combination of all of the (a priori) selected intensity features, including edema/populations, extent and subjective symptoms of pruritus and insomnia. There does not appear to be any direct relationship between the severity score and the definitions of AD severity.

It is the aim of this paper to report on a study to establish the validity of the ADAM measure (Charman, et al., 1999) and in so doing develop clear definitions of grades of severity to be derived from mathematically modelling severity of AD. The ADAM measure requires clinicians to record ratings which are clinical judgements of grades of severity. Each sign and symptom is rated as present or absent and, if present, present to a degree of severity, that is, none, mild, moderate or severe.

Clinical rating behaviours can be analyzed with mathematical (or measurement) models available to disciplines such as education, psychology and psychiatry. Essentially, modelling is the search for "latent trait" where latent trait is a statistical concept to "explain the consistency of people's responses to the items in the scale" (Streiner and Norman, 1989, p. 51). It is assumed that the latent trait underscores the pattern of ratings on the items of an measure and lies on a continuum from "less" to "more" (Wright and Masters, 1982). Applied to medical ratings, latent trait can be conceptualised as severity of disease.

In this study a variant of the Rasch measurement model, the Partial Credit model, was adopted for use as it can analyse items on ordered categories, "none", "mild", "moderate" and "severe". Moreover, the items can be evaluated independently of the sample upon which it is calibrated. This independence, or separation, is mathematically expressed as an additive system of the person's qualities (case estimates) and the item level or "difficulty" (item estimates). This additivity is achieved with a log transformation of the raw score to produce a logit score. The effect is that items and cases are measured on an interval scale with a common unit (Wright and Masters, 1982; Wright and Linacre, 1989). The items are distributed along the scale, which may, if the item fit is good, be conceptualised as the latent trait scale. Cases, too, are distributed along the same latent trait continuum from "less" of the trait ("mild") to "more" of the trait ("severe"). Cases (and items) that do not "fit" for whatever reasons (for example, misdiagnosis) may be identified and omitted from further analysis.

Method

Participants: Children (N = 171) with active AD who were consecutive patients to dermatology clinics at the Royal Children's Hospital over a twelve month period were recruited. The children were new and old patients to the clinic (mean age = 54 months, ranging from 4 to 193 months). There were 98 (57.3%) males (mean age = 47 months) and 68 (39.8%) females (mean age = 62 months). Five (2.9%) children did not have their sex recorded and 19 (11.1%) children did not have their age recorded. *Doctors*: There were seven available staff members comprised of three dermatologists and two dermatology trainees, and two medical trainees on six month rotations under the supervision of the Head of the Unit. *The ADAM measure*: The ADAM measure (Charman, et al., 1999) comprises items scored on either a four point rating scale (0 "none", 1 "mild", 2 "moderate" or 3 "severe") or a two point scale ("present" or "absent"). AD morphology items were for erythema, scale/dryness, lichenification and excoriations. Sites were face, arms, hands, legs, feet, trunk, head and neck, and flexures. *Procedure:* Participants were assessed during scheduled appointments in the dermatology clinics of the Royal Children's Hospital in Melbourne. Their treating doctor was instructed to rate the AD "as it is now" using the ADAM measure. The doctor was also advised that, where AD was absent, a blank (rather than a written zero) could be left for that item on the measure. *Analysis:* The Partial Credit analysis was computed with a software package called QUEST for MS-DOS, Student Version (1.3) (Adams and Khoo, 1993) using threshold as a measure of item difficulty. The items of the ADAM measure which were analysed consisted of the 28 site and morphology items. The remaining items did not refer to clinical manifestations of AD and were therefore excluded from the analysis.

Results

The item estimates (thresholds) were plotted as a "map" which revealed a bimodal distribution with almost all items coded as "severe" (3) located together at one end of the scale. These severe item estimates were based upon low frequencies of endorsement by assessing dermatologists. Therefore, "severe" was re-coded as "moderate/severe" and the data re-analysed. The new item estimates are provided in Table 1 and an item estimate map is provided as Figure 1. The Item Separation Reliability was R =.76, with SD Adjusted=.58. The Infit and Outfit Mean Squares were *M*=1.00 and .99 (*SD*=.12 and .23 respectively) and *t* values were .03. and .01 (*SD*=1.2 and 1.3 respectively). Thus, the summary statistics were within the accepted range (that is, *M*=1, *t*=0). The items of the ADAM measure fitted the model. The Case Separation Reliability was R = .86, with SD Adjusted=.86, with fit mean squares consistent with those expected if the model holds, that is, *M*=1.01 and .99 (*SD*=.39 and .44 respectively). The *t* values were also consistent with the model, -.06 and .00 (*SD*= 1.44 and 1.07), respectively.

The recoded data satisfied the requirement for uni-dimensionality with one item, erythema on the face, close to misfitting. Item estimates were distributed along a latent trait scale to be referred to as "AD Severity".

Table1.

Item Estimates (Thresholds) (N = 171 L = 30)

The ADAM Measure:		Thresholds		INFT	OUTFT	INFT	OUTFT
Item Name	1	2	3	MNSQ	MNSQ	t	t
1. Pruritus	-4.88	-1.56	.98	.77	.77	-2.3	-1.8
2. Face/Scale	-2.25	-.41	2.65	1.25	1.30	2.5	2.2
3. Face/Lichenification	-.41	.58	2.61	1.17	1.40	1.2	1.6
4. Face/Erythema	-1.81	-.85	.88	1.42	1.51	3.9	3.3
5. Face/Excoriations	-.34	.50	1.94	1.24	1.30	1.6	1.2
6. Arms/Scale	-2.56	-.29	2.62	1.08	1.13	.8	-1.3
7. Arms/Lichenification	-1.41	-.013	.12	1.01	.99	.1	-1.2
8. Arms/Erythema	-2.09	-.12	2.10	.88	.87	-1.2	-1.6
9. Arms/Excoriations	-.73	.073	.01	1.00	1.00	.0	1.5
10. Hands/Scale	-1.34	.30	2.22	1.00	0.99	.0	-.2
11. Hands/Lichenification	-.78	.11	2.98	.87	.79	-1.2	.1
12. Hands/Erythema	-.97	.29	NA	.90	.84	-1.0	.0
13. Hands/Excoriations	.09	.83	2.29	.89	.59	-.6	-1.1
14. Legs/Scale	-2.66	-.39	1.55	1.10	1.11	1.0	.7
15. Legs/Lichenification	-1.63	-.31	1.05	1.10	1.17	.9	.9
16. Legs/Erythema	-2.22	-.22	2.57	.89	.89	-1.1	-1.1
17. Legs/Excoriations	-1.03	-.06	1.70	.92	.87	-.7	-1.1
18. Feet/Scale	-.91	.62	.95	.91	.88	-.7	-1.0
19. Feet/Lichenification	-.47	.091	.82	.88	.69	-.9	-1.4
20. Feet/Erythema	-.66	.31	2.83	.85	0.77	-1.3	.9
21. Feet/Excoriations	.09	.83	2.29	.83	.55	-.9	1.2
22. Trunk/Scale	-2.3	4.04	2.03	1.03	1.03	.3	-.9
23. Trunk/Lichenification	-.78	.11	1.56	.98	1.10	-.1	-.7
24. Trunk/Erythema	-1.56	.27	2.26	1.04	1.06	.4	-.7
25. Trunk/Excoriations	-.47	.54	1.94	1.08	1.34	.6	1.5
26. Scalp	-.57			1.03	0.97	.5	-.2
27. NapkinArea	-.03			1.04	1.00	.4	.1
28. Head & NeckFlexures	-1.39			.99	1.01	-.2	.1
29. Limb Flexures	-2.15			1.09	1.31	1.11	2.1
30. Global Severity Rating	-5.28	-1.49	1.21	.70	.70	-3.2	-2.6
M	.00			1.00	1.00	.0	.0
SD	.86			.15	.24	1.4	1.3

NA Code 3 not used.

However, it is yet to be established that this is an accurate description for the trait.

Item placement

To define the scale for "AD Severity", item placements along the item estimate map were examined to determine whether placements were consistent with the a priori order as provided by the dermatologists. Firstly, the moderate ("2") codes were at the more "severe" end of the scale and were all consistently higher than the mild ("1") codes. Trunk and face

```
 2.0                        X

                            X
                                  13.2    21.2

                            X     18.2
                                  3.2     5.2     25.2

 1.0
                         XXXX      10.2   12.2    20.2    24.2
                            X      23.2
                            X      9.2    11.2    13.1    19.2
                                   21.1
                      XXXXXXX      7.2    17.2    27
                       XXXXXX      8.2
                          XXX      6.2    15.2    16.2
                            X      2.2     5.1    14.2
                       XXXXXX      3.1    19.1    25.1
 0.0                   XXXXXX     26
                    XXXXXXXXX     20.1
                           XX      9.1    11.1    23.1
                         XXXX      4.2    18.1
                   XXXXXXXXXX     12.1    17.1
                        XXXXX
                XXXXXXXXXXXXX
                       XXXXXX     10.1
                      XXXXXXX      7.1    28
-1.0              XXXXXXXXXXX     24.1
                        XXXXX     15.1
               XXXXXXXXXXXXXX
                     XXXXXXXX      4.1
                      XXXXXXX
                       XXXXXX      8.1
                                   2.1    16.1    29
                      XXXXXXX     22.1
                         XXXX
                                   6.1
-2.0                      XXX     14.1

                        XXXXX

                           XX
                           XX
-3.0
                                     X  represents one case
                          XXX     13.2  represents item 13 (see
                                        Table 1) at code level 2
```

Figure 1. Item estimates (thresholds) Map, 28 items (N=171)

items (22 to 25 and 2 to 5 respectively) were evenly placed along the scale, suggesting that these two sites were common in all manifestations of AD and that the type of morphological feature on these sites dictated placement on the severity scale. Face lichenification and excoriation were at the severe end of the scale. Face erythema and scale/dryness were at the less severe end of the scale. Head and neck flexures and limb flexures (items 28 and 29) were on the "milder" end of the scale. Arms and legs were placed at the moderate and mild parts of the scale.

Excoriations at all sites (items 5, 9, 13, 17, 21 and 25) were placed at the more severe end of the scale with hands and feet coded as moderate/ severe (items 13 and 21) at the extreme end of the scale. Mild excoriation on hands and feet had item estimates above the sample mean. Erythema (items 4, 8, 12, 16, 20 and 24) was present along the scale, but grade of erythema did not clearly indicate grade of severity. Therefore erythema could be a distraction despite the belief that it is an essential feature of AD but is consistent with the clinical observations that the signs of AD are similar to those for inflammation. Scale/dryness (items 2, 6, 10, 14, 18, and 22) was present across the continuum. Therefore, scale/dryness was an essential feature of "AD Severity" in all grades. This scale pattern is consistent with the clinical practice of prescribing moisturizers as an ameliorating or even a preventive treatment to avoid the anticipated development of lichenification and excoriations.

The number and distribution of the items might indicate that some of the items could be redundant for the purposes of measurement. However, apart form Face/Erythema, the fit statistics for the items are satisfactory. Moreover, clinically, cases may have equivalent severity scores but have quite different AD distributions and AD morphologies.

This ordering of difficulty as depicted by thresholds (see Table 1) was somewhat surprising in that lichenification was on the "less severe" end of the scale compared with excoriations. It was expected that excoriations were a precursor to lichenification, that is excoriations would have had lower thresholds. Either this proposition was incorrect or the calibration sample was a "chronic" sample with their skin already lichenified when the ratings took place. The sample was indeed "chronic", as the mean age was 4½ years, with mean onset of AD within the first 12 months and having waited a mean of 6 months for an appointment with the dermatologists.

Examination of the pattern of step 2 thresholds showed approximately equivalent thresholds for scale, erythema, lichenification and excoriations. Thus, severe manifestations of AD included all four morphological features.

Case Placement

The case fit estimates were M= -0.73, SD = .93 (i.e., the mean was within one SD of zero), and the range of case fit estimates varied from -3.29 to +1.99. The mode consisted of two equally occurring estimates, -1.13 and -.62. Both of these values fell in the "mild" end of the latent trait scale. Forty-two (24.4%) cases had positive case estimates, that is estimates > 0. Three (1.7%) cases were "too" conforming to the model (Wright and Masters, 1982), with raw scores of only 2, out of a possible maximum of 52 (estimates=-3.29), three standard deviations below the mean. When the original protocols for these three cases were examined, one case had mild scale on the face and arms, one case had mild scale on the arms only, and one had mild scale and mild lichenification on the face. These three cases were retained as representative of "trivial" AD which were occasionally seen in clinic and who may have presented for review appointments after successful treatment.

Concurrent validity

Concurrent validity was explored by determining if placement of items along the scale could define grades of AD Severity. These defined "grades" were compared with the grades of severity as provided by ratings on Global Rating of Severity, item 30 on the ADAM measure. According to the dermatologists' ratings on Global Rating of Severity, approximately 7% of the sample had severe AD, 48% had moderate AD, 43% had mild AD and 2% had negligible AD.

To define grades of severity based on the item placements on the latent trait scale, it was necessary to divide the scale into four sections, representing trivial, mild, moderate and severe. As the same scale defined case placements, this case scale was divided so that each section represented the appropriate number of cases which were equivalent to the above percentages. Occasionally, where there were "tied" cases, the numbers in the section exceeded the percentage. The final details of the grades are provided in Table 2.

Table 2.

Logit Estimates for Grades of AD Severity

Description	Total Score Range	Case and Item Logit Range	Measurement Error	N
Trivial	0 - 2	-3.29	.74	3 (2%)
Mild	3 - 14	-2.84 to -.91	.32 to .61	75 (43.6%)
Moderate	15 - 31	-.81 to .48	.27 to .31	77 (45.3%)
Severe	32 - 46	.56 to 1.99	.28 to .41	16 (9.3%)
Total				171 (100%)

In order to test whether there was an association between the grade of severity determined by the logit range and the grade of severity rated directly by dermatologists, a frequency table was constructed and is provided as Table 3.

Table 3.
Frequency of Global Ratings of Severity for Each Logit Range

Logit Range	Trivial	Mild	Moderate	Severe	Total
Trivial	1	2	0	0	3 (1.8%)
Mild	2	53	19	0	74 (43.3%)
Moderate	0	19	52	7	78 (45.6%)
Severe	0	0	11	5	16 (9.4%)
Total	3	74	82	12	171
%	1.8%	43.3%	48.0%	7.0%	100.0%

Because there were 8 cells with expected cell size less than 5, the table was collapsed by combining the cells for trivial global rating and trivial logit scores with the mild cells. This produced a 3 by 3 table, $\chi^2(4)=65.4$, $p<.01$. Thus grades based upon the Partial Credit estimates (logit ranges) were not independent of the Global Ratings of Severity (code levels) by the dermatologists and agreement was "marginal", kappa=.40, SE=.06, $p<.05$. Inspection of the table revealed some differences in the two gradings. Dermatologists more frequently rated AD as "mild" compared to the logit estimates, with 53 agreements and 42 (2,2, 19 and 19) disagreements. Dermatologists less frequently rated AD as "severe" with only 5 agreements and 18 (7 and 11) disagreements.

As stated above, because cases and items have a common scale, the cut-off points to define frequency of cases in each section also defined groups of items. These items were summarized as "word pictures" and are provided in Table 4. The last column in Table 4 consists of the dermatologists' clinical descriptions developed for the Global Rating item on the ADAM measure.

It can be seen that word descriptions based upon the model provided more comprehensive and consistent definitions of grades of severity which matched the a priori clinical expectations. On the other hand, clinical description using "admittable" is somewhat vague and could be influenced by a range of other criteria, such as whether a hospital bed was available, what psychosocial issues were salient for the patient and so forth. In short, the Partial Credit model of the latent trait, "AD Severity" does appear to have content, construct and concurrent validity.

Table 4

Descriptions of Grades of AD Severity

Grade	Description based on item analysis	Description by dermatologists
Trivial	Mild scale on arms or legs	
Mild	Mild scale and erythema on face, arms and/or legs, and possibly in the flexures of the limbs. Mild erythema on the trunk. Mild lichenification on arms and/or legs.	A bit scaly, and a bit red.
Moderate	Moderate/severe scale and erythema on hands and feet. Moderate/severe lichenification of arms and/or legs and mild lichenification on other sites of the body. Mild excoriations on face, arms legs and/or trunk. AD is present on the scalp and head and neck flexures.	Lichenification present. A few exoriations. Finger marks. No vesiculation.
Severe	Moderate/severe scale and lichenification on other sites, including hands. Moderate/severe erythema on hands, feet and trunk. Moderate/severe excoriations on arms, legs, feet, face, trunk and just appeared on hands and feet. There is AD on the napkin area.	Nodules exoriated. Lichenification widespread. Admittable or almost admittable. Vesiculation.

Visible AD

Since hospital utilization and adherence may be related to whether AD is on visible sites (Mechanic, 1978), certain questions needed to be asked. These questions relate to how visibility of AD was accommodated within this latent trait model and if children with visible AD had higher dermatologists' severity ratings than children with no visible AD. To examine these questions, visible areas were defined as face and hands. The word pictures illustrate that, as the level of AD severity increased, the features of AD intensified on the face with moderate/severe face items on the severe end of the scale. Mild face scale was at the "trivial" end of the scale. It was only in the moderate and severe grades that morphological features appeared and intensified on the hands. Thus, visibility of AD was associated with more severe "AD Severity".

Sex variations

An analysis was undertaken to investigate whether there was a significant difference in AD severity between males and females. Case estimates were $M = -.67$, $SD=.95$ and $M = -.80$, $SD=.93$, for males (n=98) and females (n=68) respectively. The difference between the means was not statistically significant, $F(1,164)=.76$, $p=.39$, with 5 missing cases.

Since, these null results may be an outcome of the way in which the model was estimated using grouped data, the Partial Credit model was re-estimated for males and females separately. The item estimates for each sex were plotted against each other and a regression line fitted to the data. One outlier, at 3 standard deviations, was identified. This outlier was moderate/severe excoriations on the face (Face/Excoriations). The male threshold for this item was .90. For females the threshold was 2.64 which is at the extreme end of the *female* "AD Severity" scale. Excoriations on the face were only seen in very severe manifestations of AD in females, and in males with less severe AD. Excoriations on the hands and feet were beyond two standard deviations and scored only in severe *male* AD. Thus Partial Credit analyses found that, when AD is mild to moderate, males were more likely to have excoriated faces and females to have excoriated hands and feet. Only when AD is very severe AD do females have excoriated faces and males have excoriated hands and feet.

Age variations

Age variations have been reported by Rajka (1989). The infantile phase for AD is up to 2 years of age during which AD is found on the face and scalp. The childhood phase is between 3 and 11 years and AD is on the limbs. The adolescent phase is 12 years onwards and AD is mostly on the trunk and the hands (in more severe manifestations). To investigate whether these reported age patterns or variations of AD were found in this sample a Partial Credit analysis was performed.

The numbers of patients in each of the age groups were 61, 44 and 14. Consequently, the groups of patients older than 2 years were combined. The Partial Credit model was re-estimated separately for infants (equal to or less than 2 years) and children (more than 2 years). The item estimates for infants and children were plotted against each other and a regression line fitted to the data. No outlying items were detected at three standard deviations. At two standard deviations the outlying items were erythema and severe excoriations on the face and mild scale on the legs. Infants with *mild* AD did have AD (erythema) on the face. Infants with *severe* AD had AD on the face, hands and feet. Rajka (1989) stated that older children had AD on limbs. This too was confirmed. AD appeared on the limbs, but, only in children with *mild* AD. Older children with *severe* AD had AD on the face. Rajka (1989) also stated that AD on the flexures was not always present in infancy, and these sites only get involved when the child is older than two years. Flexure item thresholds for infants and children were -.50 and -1.19 (for infants) and -.93 and -1.86 (for children). Once again, descriptions by Rajka (1989) have been confirmed. Flexures may not be present in infancy. However, some infants did have AD on the flexures, but these cases had more severe AD than the older children who had AD in the flexures.

Discussion

Partial Credit analyses of the site and morphology items of the ADAM measure showed that items and cases were distributed along a continuous scale and that placements conformed with clinical expectations. The thresholds suggested that the sample had chronic AD rather than recent onset AD. Also, the model produced coherent "word pictures" of AD severity which compared favourably with clinically derived descriptions. Thus, the continuous scale was conceived as "AD Severity" and found to have content and construct validity.

When the data was estimated separately for males and females, differences emerged. Essentially, males had more visible AD at less severe AD. This may account, in part, for the apparent better prognosis for males. Perhaps, parents respond to the visibility of the AD in their male children by obtaining earlier treatment. When the data was estimated separately for infants and children, the clinical descriptions by Rajka (1989) were confirmed only for *mild* AD. The descriptions of *moderate/severe* AD had variations on these basic descriptions.

The ADAM measure was designed to provide ratings of site and morphology together rather than separately as in previously published scoring systems. An earlier agreement study revealed, that like similar measures in clinical work, operational definitions of grades of AD severity were not well-enough defined (Charman, et al., 1999). In this current study, the application of Partial Credit modelling revealed "word pictures" and sex and age variations were identified. These word pictures constitute operational definitions of grades of severity.

The less reliable items and the ill-fitting items could be omitted from the ADAM measure. However, in doing this important clinical information may be missed. At this stage, instead, it is recommended that doctors have special training sessions in the identification and grading of lichenification and erythema in particular. Face items, especially erythema on the face may be distractor items. There are many reasons why a patient might have a red face at the doctors.

Use of clinical ratings of signs to characterise AD, or any other disease for that matter, is fraught with problems, not the least of which is the reliability of the ratings. The procedure adopted here is recommended as a basis for generating operational definitions of severity wherever grades are used as a basis for treatment. Unusual results such as the finding for erythema can inform and direct clinical training and practice and research.

References

Adams, R.J. and Khoo, S.T. (1993). *Quest: The interactive test analysis system.* Melbourne: Australian Council for Educational Research.

Charman, D.P., Varigos, G.A., Horne, D.J. and Oberklaid, F. (1999). ADAM: The development of a practical and reliable severity measure for Atopic Dermatitis. *Journal of Outcome Measurement, 3*(1), 21-34.

Charman, D.P. (1997). *Patterns in Atopic Dermatitis: Developing models to predict hospital utilization patterns.* University of Melbourne, Australia. Ph.D. thesis.

Diepgen, T. L., Sauerbrei, W. and Fartasch, M. (1996). Development and validation of diagnostic scores for atopic dermatitis incorporating criteria of dta quality and practical usefulness. *Journal of Clinical Epiedmiology, 49*(9), 1031-1038.

European Task Force on Atopic Dermatitis. (1993). Severity scoring of Atopic Dermatitis: The SCORAD index. *Dermatology, 186*, 23-31.

Hall, D.E., Lynn, J.M., Altieri, J. and Segers, V.D. (1987). Inter-intrajudge reliability of the stuttering severity instrument. *Journal of Fluency Disorders, 12*(13): 167-173.

Heiman, G.W. (1992). *Basic statistics for the behavioural sciences.* Boston: Houghton Mifflin Company.

Mechanic, D. (1978). *Medical Sociology* (2nd Ed.) New York: Free Press.

Rajka, G. (1989). *Essential aspects of Atopic Dermatitis.* Berlin: Springes-Verlag.

Sackett, D.L., Haynes, R.B., Guyatt, G.H. and Tugwell, P. (1991). *Clinical epidemiology: A basic science for clinical medicine.* (2nd ed). Boston: Little, Brown and Company.

Spiteri, M.A., Cook, D.G. and Clarke, S.W. (1988). Reliability of eliciting physical signs in examination of the chest. *The Lancet*, *6*, 873-875.

Streiner, D.L. and Norman, G.R. (1989). *Health measurement scales: A practical guide to their development and use.* New York: Oxford University Press.

Wright, B.D. and Linacre, J.M. (1989). Observations are always ordinal: Measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, *70*(12): 857-60.

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

JOURNAL OF OUTCOME MEASUREMENT®, *3*(2), 176–199

Parameter Recovery for the Rating Scale Model Using PARSCALE

Guy A. French

Barbara G. Dodd

The University of Texas at Austin

The purpose of the present study was to investigate item and trait parameter recovery for Andrich's rating scale model using the PARSCALE computer program. The four factors upon which the simulated data matrices varied were (a) the distribution of the scale values for the items (skewed or uniform), (b) the number of category response options (4 or 5), (c) the distribution of known trait levels (normal or skewed), and (d) the sample size (60, 125, 250, 500, or 1,000). Each condition was replicated 10 times resulting in 400 data matrices. Accurate item and trait parameter estimates were obtained for all sample sizes examined. As expected, sample size seemed to have little influence on the recovery of trait parameters but did influence item parameter recovery. The distribution of known trait levels did not seriously impact the item parameter recovery. It was concluded that Andrich's rating scale model allows for the use of considerably smaller calibration samples than are typically recommended for other polytomous IRT models.

Requests for reprints should be sent to Guy A. French, Crew System Interface Division; 2255 H Street; Wright-Patterson, OH 45433-7022.

When a model is employed to measure a latent construct, it is difficult to determine how accurate the model is by investigating the actual construct in the real world. The very nature of a latent construct means that it does not lend itself readily to the confirmation of a model's parameters using some computer program. Therefore, simulated data, where the latent construct is perfectly known, are frequently used. Accuracy is typically assessed in terms of the discrepancies between the known model parameters and those recovered by the calibration computer program. The present simulation study was performed to assess the accuracy of the PARSCALE (Muraki and Bock, 1993) computer program to estimate the parameters of Andrich's (1978a, 1978b) rating scale model, which is appropriate for attitude measurement. Several other item response theory models and computer programs have been assessed in similar types of studies.

Most research in the area of parameter recovery has focused on the dichotomous item response models. These models have been useful for the construction, administration, and scoring of cognitive ability or achievement tests (Hambleton, Jones, and Rogers, 1993). Far fewer studies have examined the accuracy of the polytomous item response models. One such study, by Reise and Yu (1990), looked at the recovery of item and ability parameters using the graded response model (Samejima, 1969) and the calibration program MULTILOG (Thissen, 1986). Three factors were manipulated: true ability distribution, true item discrimination distribution, and calibration sample size. A test length of 25 items with five response categories per item was used for all conditions. Results indicated that sample sizes of 500 or greater consistently provided true-to-estimated parameter correlation coefficients of 0.85 or larger. Root mean square error differences were concluded to be comparable with those reported in dichotomous model parameter recovery studies (Hulin, Lissak, and Drasgow, 1982; Yen, 1987). A sample of 500 for this study is equivalent to a 4:1 ratio of sample size to the number of item parameters.

Reise and Yu (1990) also found that the calibration sample size had little effect on the recovery of ability parameters, but influenced the recovery of item parameters. In addition, they concluded that the true ability distribution and true item discrimination magnitude influenced the recovery of ability and item parameters. Uniformly distributed true ability parameters and large true item discrimination values resulted in the best parameter recovery.

Another parameter recovery study by Walker-Bartnick (1990) examined the accuracy of item parameter estimation for the partial credit model (Masters, 1982) and the calibration program MSTEPS (Wright, Congdon, and Schultz, 1988). Three factors were manipulated: true theta distribution, ratio of sample size to number of parameters to be estimated, and number of response categories per item. The test length was held constant at 80 items. The results led Walker-Bartnick to conclude that the true theta distribution and the number of response categories per item did not affect the stability of the recovered parameters. Walker-Bartnick also concluded that a ratio of sample size to number of parameters of 2:1 was the minimum sufficient for stable parameter recovery. It should be noted that the 2:1 ratio for this study required fairly large sample sizes of 640 and 800 for the five- and six-response categories conditions, respectively.

A study by De Ayala (in press) examined parameter recovery using the nominal response model (Bock, 1972) and MULTILOG (Thissen, 1991) computer program. Three factors were studied: sample size ratio, true theta distribution, and item information level. Results indicated that as the ability distribution departs from a uniform distribution, the accuracy of the slope parameter recovery declines. De Ayala also concluded that a 5:1 ratio of examinees to parameters will produce reasonably accurate item parameter estimates.

Choi, Cook, and Dodd (1997) investigated the parameter recovery for the partial credit model using the MULTILOG (Thissen, 1991) calibration program. The factors studied were the number of item parameters, step values per item, and sample size. Choi et al. recommended that the sample size to number of item parameters ratio guideline for accurate item parameter estimation needs to be adjusted upward if the number of response categories in an item is large. That is, as the number of response categories per item increases, larger calibration samples are necessary.

Findings for the particular polytomous IRT models that have been studied may not generalize to other polytomous IRT models because the number of item parameters and their definitions can vary from model to model. As a consequence, the primary objective of the present study was to investigate the effect of sample size, the distribution of trait levels, the number of response categories, and the distribution of item parameters (scale values) on the recovery of known item and person parameters for the rating scale model (Andrich, 1978a, 1978b) using the PARSCALE computer program (Muraki and Bock, 1993). The goal of the present study was to determine if the general guidelines that have been recommended for the partial credit model, the nominal response model, or the graded response model generalize to Andrich's rating scale model.

Andrich's Rating Scale Model

The primary purpose for the development of the rating scale model was to give researchers a way to analyze attitudinal response data using item response theory. Andrich (1978a, 1978b) extended the Rasch model for dichotomously scored items to the case where there are more than two response options available for each item.

In the rating scale model, a single scale value for each item is estimated on the same metric as the trait parameter. Simultaneously, a set of response thresholds is estimated for the entire scale. The number of thresholds estimated for the scale is one less than the number of response categories. Andrich defined the probability of responding in a given category, x ($x = 0, 1, \ldots, m_i$), on item i as

$$P_{xi}(\Theta) = \frac{\exp\left[\sum_{j=0}^{xi}(\Theta - (b_i + t_j))\right]}{\sum_{k=0}^{mi}\exp\left[\sum_{j=0}^{k}(\Theta - (b_i + t_j))\right]},$$

where P_x is the probability of responding in a particular response category, x, on item i, Θ is the trait level for a given individual, b_i is the scale value of item i, and t_j is the threshold value for the response category in question. As can be noted by its absence from the formula, item discrimination is assumed to be constant across items. Figure 1 depicts the operating characteristic function for a rating scale item with four response options. The scale value for this item is 0 and the threshold values for the scale are -0.8, 0, and 0.8.

Method

Overview of the Design

Monte Carlo data were generated to study the ability of the PARSCALE program to recover item and trait parameters according to Andrich's rating scale model under several conditions. All conditions used a common test length of 30 items. There were 40 simulated data matrices corresponding to a fully crossed design of the following variables: (a) the distribution of the true scale values–skewed and uniform; (b) the number of response categories per item–4 and 5; (c) the distribution of known trait levels–normal and skewed; and (d) the sample size–60, 125, 250, 500, and 1000. Each cell in the fully crossed design was replicated 10

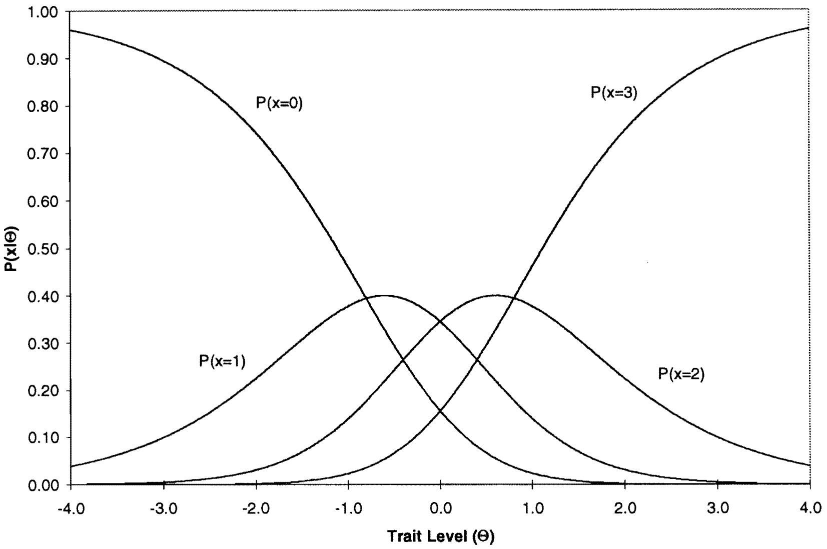


Figure 1. Operating characteristic function for a four response category rating scale item.

times using different seed values to generate new subjects; thus, there were a total of 400 (1040) matrices examined in this study.

Item Pools

Four item pools were constructed by completely crossing the two levels of the distribution of the scale values with the two levels of the number of response categories per item. Each item pool consisted of 30 items. The uniform and skewed distributions of scale value parameters are presented in Table 1. The uniform distribution was created by spacing the scale values evenly across the range of -2.0 to 2.0. The skewed distribution of scale values was created to have a mean of -0.8, standard deviation of 0.75, and a skewness index of 0.6.

Table 2 contains the threshold parameters, which were selected to be representative of the threshold values that have been obtained for real data sets (Dodd, 1990).

Simulated Data Generation

Known thetas for the simulees in the normal distribution of trait level condition were generated by randomly selecting z scores from a unit normal distribution. For a particular simulee, the probability of responding in each response category was calculated and compared to a random

Table 1

True Scale Values for the Various Distribution Conditions

	Distribution	
Item Number	**Uniform**	**Skewed**
1	-2.000	-1.985
2	-1.862	-1.802
3	-1.724	-1.769
4	-1.586	-1.652
5	-1.448	-1.541
6	-1.310	-1.531
7	-1.172	-1.487
8	-1.034	-1.415
9	-0.897	-1.408
10	-0.759	-1.313
11	-0.621	-1.266
12	-0.483	-1.215
13	-0.345	-1.039
14	-0.207	-1.007
15	-0.069	-0.998
16	0.069	-0.939
17	0.207	-0.873
18	0.345	-0.816
19	0.483	-0.713
20	0.621	-0.687
21	0.759	-0.670
22	0.897	-0.364
23	1.034	-0.305
24	1.172	-0.226
25	1.310	-0.102
26	1.448	-0.075
27	1.586	0.050
28	1.724	0.389
29	1.862	0.744
30	2.000	0.829

Table 2

True Scale Threshold Values for the Various Response Category Conditions

Number of Categories	t_1	t_2	t_3	t_4
4	-1.20	0.00	1.20	
5	-1.50	-0.75	0.75	1.50

number drawn from a uniform distribution ranging from 0 to 1. The random number was successively compared to the cumulative probability values from the lowest response category to the highest response category. For each response category where the random number was greater than the cumulative probability value, the simulee received 1 point for the response category. When the random number was equal to or less than the cumulative probability, the simulee was assigned 0 points for the response category. The simulee's score for the item was simply the sum of the response category scores for the item.

While it may seem more intuitive to simply calculate cumulative probabilities for each response category and find the response category within which the random number response falls to obtain the item score, the summation of response category scores provides the same answer. For each item, a simulee could receive a score ranging from zero points for failing to exceed the cumulative probability value associated with the first response category to four points for exceeding the cumulative probability value associated with the next to last response category of the five-option scale. The same process was applied to generate the responses to the four-option scale using the same known thetas, with the maximum possible item score being three. The process was conducted for each simulee to create the response strings that were used as the input to the PARSCALE program.

The procedures used to generate the known thetas for the skewed distribution of trait level condition were identical to the procedures used for the normal distribution of trait level condition except that a skewed distribution was used instead of the unit normal distribution. More specifically, the known trait levels were drawn from a skewed distribution with a mean of -0.8, standard deviation of 0.75, and a skewness index of 0.6. These parameters were selected so that the skewed trait level distribution would match the skewed distribution condition for the scale values.

Parameter Estimation

Item and person parameters were estimated according to Andrich's rating scale model with the PARSCALE (Muraki and Bock, 1993) computer program. PARSCALE employs the marginal maximum likelihood estimation procedure described by Bock and Aitkin (1981) to estimate item parameters. The person parameters can be estimated with either the maximum likelihood or expected aposterior estimation procedures. For this study the maximum likelihood procedure was chosen to estimate the trait levels of the simulees. To deal with the indeterminacy of the trait and item parameter scale, PARSCALE sets the ability scale to be normal with a mean of zero and a standard deviation of one.

Analysis

As criteria for judging the recovery of the parameters, three common indices were selected. The root mean square error (RMSE) index, the Bias index, and the Pearson correlation coefficient were calculated between the known parameters and the recovered parameters and averaged across replications.

RMSE was calculated for the theta estimates using the formula

$$\mathrm{RMSE}(\Theta_j) = \sqrt{\frac{\sum_{j=1}^{n}\left(\hat{\Theta}_j - \Theta_j\right)^2}{n}},$$

where Θ_j represents true trait, $\hat{\Theta}_j$ represents estimated ability, and n represents the number of simulees in the condition. For the item parameters (scale values and thresholds) the Θ_js were replaced, respectively, with b_is and t_{ik}s for the true values, and $\hat{b}_i$s and $\hat{t}_{ik}$s for the estimated values.

For the Bias calculations the following formula was used:

$$\mathrm{Bias}(\Theta_j) = \frac{\left(\hat{\Theta}_j - \Theta_j\right)}{n}$$

where the symbols represent the same quantities as in the RMSE formula. In keeping with the results reported by Reise and Yu (1990), also computed were the averages across the main factors that were manipulated in the current study for each distribution of trait level condition, respectively.

Results

Recovery indices are reported for the theta levels, the scale values, and the threshold parameters. As indicated by Reise and Yu (1990), there are no standards by which to measure the power of the parameter recovery efforts within the polytomous item response models. Reise and Yu used the recovery indices from a two-parameter logistic study (Hulin et al., 1982) and one that used marginal-maximum likelihood (Yen, 1987). For this study, the values obtained by Choi et al. (1997) were used for comparison.

Trait level (Θ)

Normal datasets. Table 3 contains the RMSE, correlation, and Bias indices averaged across replications for the trait level parameters in the normal distribution of trait level condition. The RMSE results are depicted graphically in Figure 2, while the correlation coefficients that were obtained are presented in Figure 3. Table 4 contains the RMSE, correlation coefficients, and Bias indices averaged for each main effect of the manipulated variables.

As pointed out in Reise and Yu (1990) and Choi et al. (1997), sample size doesn't appear to impact the recovery of theta parameters. Holding the other factors constant, there is little difference between the RMSEs at any levels of sample size, nor is there any apparent trend in these minor differences. Differences between the skewed and uniform distribution of the scale value conditions also appear to be negligible (the average differ-

Table 3

RMSE, Bias, and Correlation Indices for Trait Parameter Recovery Averaged Across Replications for the Normal Distribution of Known Trait Level Condition

Distribution of Items	Number of Categories	Number of Simulees	RMSE	Correlation	Bias
Skewed	**4**	**60**	0.2585	0.9663	-1.83E-06
		125	0.2552	0.9673	8.00E-07
		250	0.2687	0.9639	-1.48E-06
		500	0.2599	0.9662	9.40E-07
		1000	0.2686	0.9639	1.67E-06
	5	**60**	0.2204	0.9755	5.00E-07
		125	0.2215	0.9754	-1.12E-06
		250	0.2289	0.9738	-4.00E-08
		500	0.2341	0.9726	-4.00E-07
		1000	0.2318	0.9731	-5.50E-07
Uniform	**4**	**60**	0.2431	0.9702	2.17E-06
		125	0.2536	0.9678	8.00E-08
		250	0.2556	0.9672	-1.60E-06
		500	0.2552	0.9674	1.06E-06
		1000	0.2569	0.9670	-7.00E-07
	5	**60**	0.2228	0.9750	-1.00E-06
		125	0.2161	0.9766	-1.04E-06
		250	0.2283	0.9739	2.16E-06
		500	0.2274	0.9741	1.40E-06
		1000	0.2274	0.9741	-1.15E-06

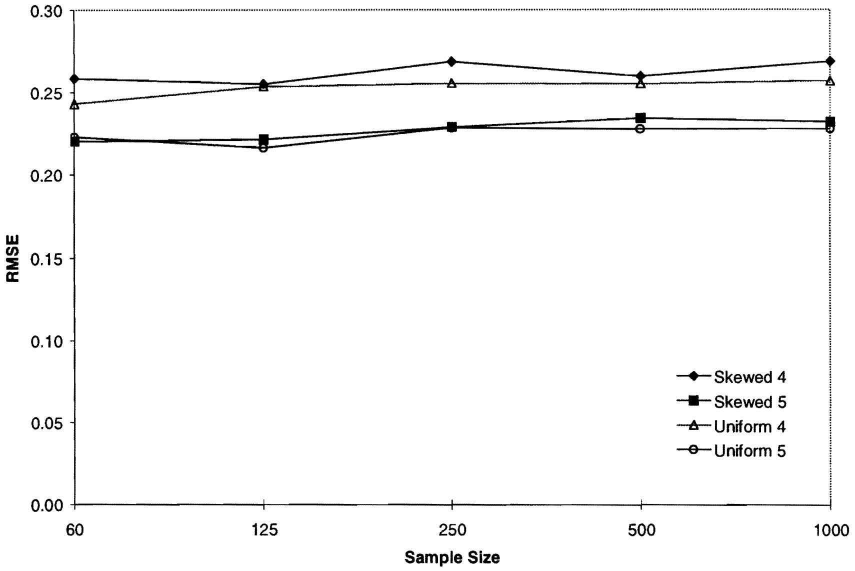


Figure 2. RMSEs for the trait parameter estimates averaged across replications for the normal distribution of known trait level condition.

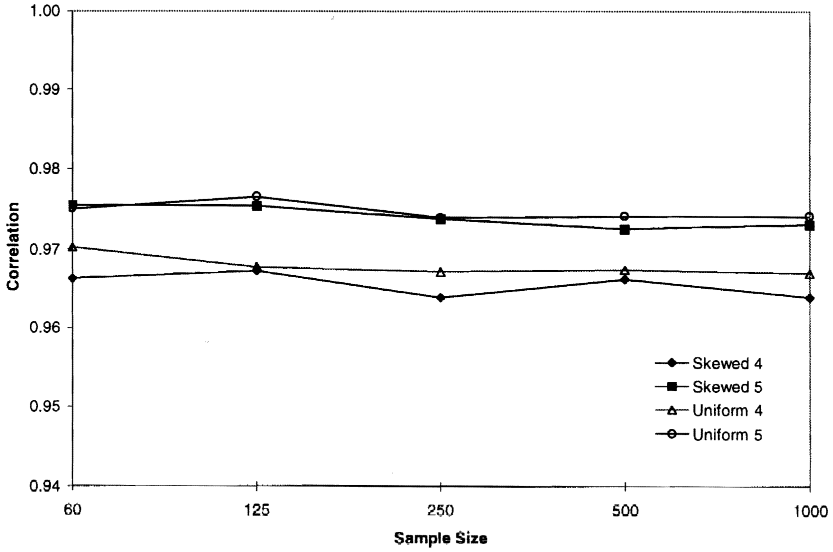


Figure 3. Correlation between estimated and true trait parameters averaged across replications for the normal distribution of known trait level condition.

Table 4

Main Factor Indices for Trait Recovery for the Normal Distribution Known Trait Level Condition

	RMSE	Correlation	Bias
Skewed	0.2448	0.9698	9.33E-07
Uniform	0.2386	0.9713	1.24E-06
Four Categories	0.2575	0.9667	1.23E-06
Five Categories	0.2259	0.9744	9.36E-07
Sample Size: 60	0.2362	0.9717	1.38E-06
125	0.2366	0.9718	7.60E-07
250	0.2453	0.9697	1.32E-06
500	0.2441	0.9701	9.50E-07
1000	0.2462	0.9695	1.02E-06

Table 5

RMSE, Bias, and Correlation Indices for Trait Parameter Recovery Averaged Across Replications for the Skewed Distribution of Known Trait Level Condition

Distribution of Items	Number of Categories	Number of Simulees	RMSE	Correlation	Bias
Skewed	**4**	**60**	0.2173	0.9630	-3.04E-05
		125	0.2146	0.9488	-8.52E-06
		250	0.2297	0.9538	1.34E-05
		500	0.2305	0.9542	1.99E-06
		1000	0.2295	0.9548	4.55E-06
	5	**60**	0.2051	0.9670	-2.78E-05
		125	0.2064	0.9625	-6.99E-06
		250	0.2078	0.9622	1.39E-05
		500	0.2080	0.9627	1.23E-06
		1000	0.2047	0.9641	5.57E-06
Uniform	**4**	**60**	0.2591	0.9475	-2.84E-05
		125	0.2515	0.9445	-8.47E-06
		250	0.2503	0.9451	1.21E-05
		500	0.2474	0.9472	-2.11E-08
		1000	0.2536	0.9448	5.94E-06
	5	**60**	0.2043	0.9676	-2.91E-05
		125	0.2272	0.9546	-7.94E-06
		250	0.2225	0.9566	1.47E-05
		500	0.2240	0.9568	2.43E-06
		1000	0.2259	0.9562	6.00E-06

ence in RMSEs is .0062). The number of response categories variable does show differences that indicate that five response categories are better than four response categories for theta parameter estimation. The average difference between the four and five response category RMSEs was .03490 for the skewed scale value conditions and .0285 for the uniform scale value conditions. Results of the correlation analysis support the conclusions drawn from the analysis of the RMSE data. For all practical purposes, the Bias index was functionally zero for all experimental conditions.

Skewed datasets. Table 5 presents the RMSE, correlation, and Bias indices averaged across replications for the trait level parameters in the skewed distribution of trait level condition. Figure 4 shows the RMSE results graphically, while Figure 5 displays the correlation coefficients. The average RMSE, correlation coefficients, and Bias indices for each of the main effects are presented in Table 6.

As was the case with the normal datasets, theta estimation was not greatly influenced by sample size. The difference between the average RMSEs for the skewed and uniform distributions of scale values was quite small (.0185). The theta parameters were estimated better in the five response category conditions than the four response category conditions. The average difference between the RMSEs for the four and five category

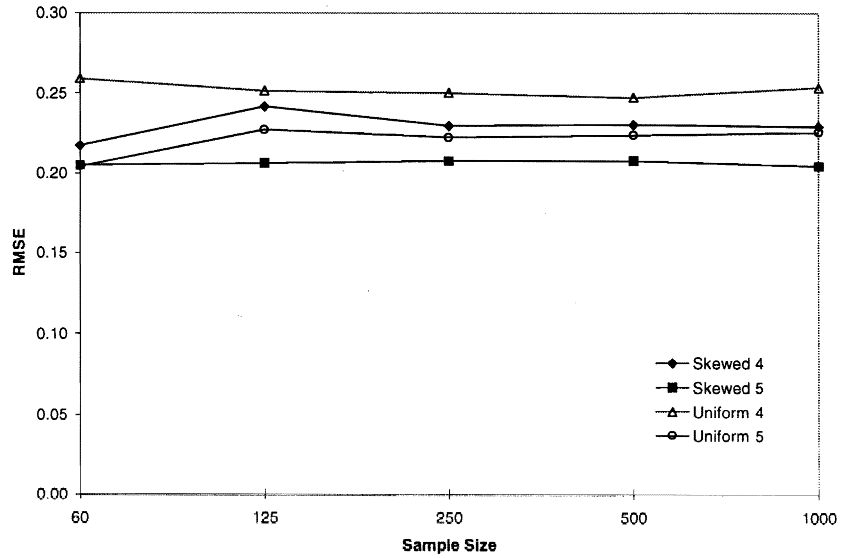


Figure 4. RMSEs for the trait parameter estimates averaged across replications for the skewed distribution of known trait level condition.

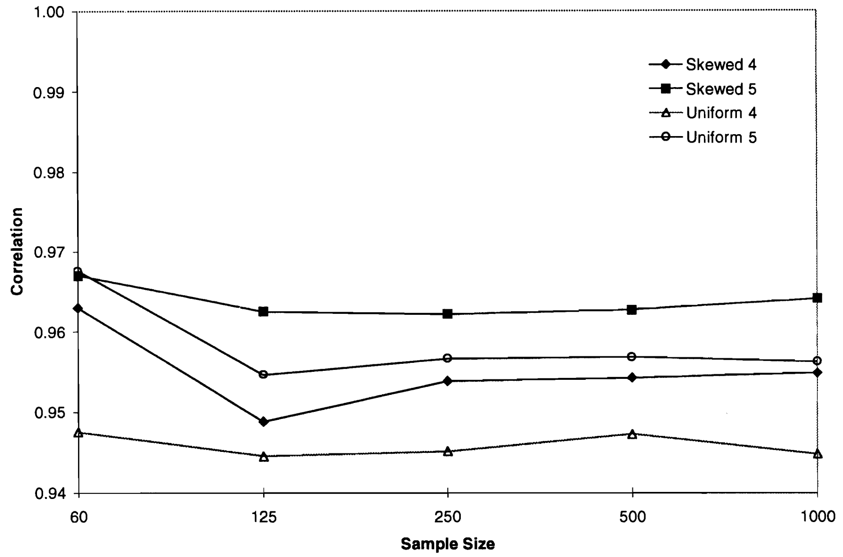


Figure 5. Correlation between estimated and true trait parameters averaged across replications for the skewed distribution of known trait level condition.

Table 6

Main Factor Indices for Trait Recovery for the Skewed Distribution of Known Trait Level Condition

	RMSE	Correlation	Bias
Skewed	0.2181	0.9597	-3.31E-06
Uniform	0.2366	0.9527	2.40E-06
Four Categories	0.2410	0.9507	-3.78E-06
Five Categories	0.2136	0.9613	-2.80E-06
Sample Size: 60	0.2214	0.9620	-2.89E-05
125	0.2317	0.9531	-7.98E-06
250	0.2276	0.9548	1.35E-05
500	0.2275	0.9556	1.41E-06
1000	0.2284	0.9555	5.52E-06

conditions was .0179 for the skewed distribution of scale values and .0316 for the uniform distribution of scale values. Unlike the finding for the normal datasets, the average RMSE was slightly smaller for the skewed distribution of scale values condition than the uniform distribution of scale values condition. In general, the correlation coefficients mirrored the RMSE results. As was the case with the normal datasets, the Bias index was essentially zero for all conditions.

Item Parameters: Scale Value (b_i)

Normal datasets. Table 7 contains the RMSE, correlation, and Bias indices averaged across replications for the scale value parameters in the normal distribution of trait level condtions. Table 8 shows the same results collapsed across the main effects of the variables manipulated in the study for the normal datasets. Results of the RMSE analysis indicate that sample size was the most important factor affecting scale value recovery. Across all conditions, RMSEs ranged from .2074 (uniform, 4 response categories, N=60) to .0388 (skewed, 5 response categories, N=1000). In general, the skewed item pool showed better recovery at sample sizes of 60 and 125, while results for larger sample sizes show no evidence that either item pool distribution provides an advantage over the other.

In terms of the number of response categories, RMSEs in all cases of the five response category conditions provided better recovery than

Table 7

RMSE, Bias, and Correlation Indices for Scale Value Recovery Averaged Across Replications for the Normal Distribution of Known Trait Level Condition

Distribution of Items	Number of Categories	Number of Simulees	RMSE	Correlation	Bias
Skewed	**4**	**60**	0.1805	0.9723	-0.0021
		125	0.1267	0.9870	0.0062
		250	0.0905	0.9930	0.0126
		500	0.0613	0.9969	0.0057
		1000	0.0450	0.9983	0.0102
	5	**60**	0.1698	0.9798	-0.0322
		125	0.1040	0.9912	0.0004
		250	0.0793	0.9947	0.0083
		500	0.0555	0.9976	-0.0009
		1000	0.0388	0.9988	0.0080
Uniform	**4**	**60**	0.2074	0.9868	-0.0043
		125	0.1355	0.9945	0.0108
		250	0.0900	0.9973	-0.0057
		500	0.0652	0.9986	-0.0043
		1000	0.0487	0.9992	0.0055
	5	**60**	0.1634	0.9918	0.0183
		125	0.1132	0.9959	0.0054
		250	0.0770	0.9982	-0.0035
		500	0.0546	0.9991	0.0050
		1000	0.0409	0.9995	-0.0029

Table 8

Main Factor Indices for Scale Value Recovery for the Normal Distribution of Known Trait Level Condition

		RMSE	Correlation	Bias
Skewed		0.0951	0.9910	0.0016
Uniform		0.0996	0.9961	0.0024
Four Categories		0.1051	0.9924	0.0035
Five Categories		0.0896	0.9947	0.0006
Sample Size:	**60**	0.1803	0.9827	-0.0051
	125	0.1198	0.9922	0.0057
	250	0.0842	0.9958	0.0029
	500	0.0592	0.9980	0.0014
	1000	0.0434	0.9990	0.0052

their four response category analogs. The largest differences in recovery between the four and five response category conditions occur at sample sizes of 60, while the smallest occur at sample sizes of 1000. Figure 6 presents the RMSE results graphically.

Figure 7 contains the results for the average correlation coefficients and shows a trend. As was found for the RMSE index, the magnitude of the average correlation coefficients was a function of sample size and the number of response categories. Unlike in the findings for the RMSEs, the

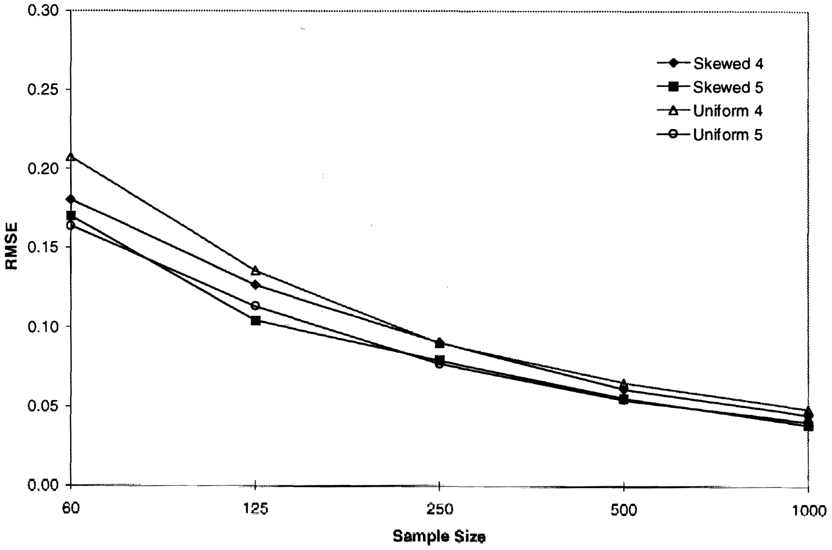


Figure 6. RMSEs for the scale value parameter estimates averaged across replications for the normal distribution of known trait level condition.

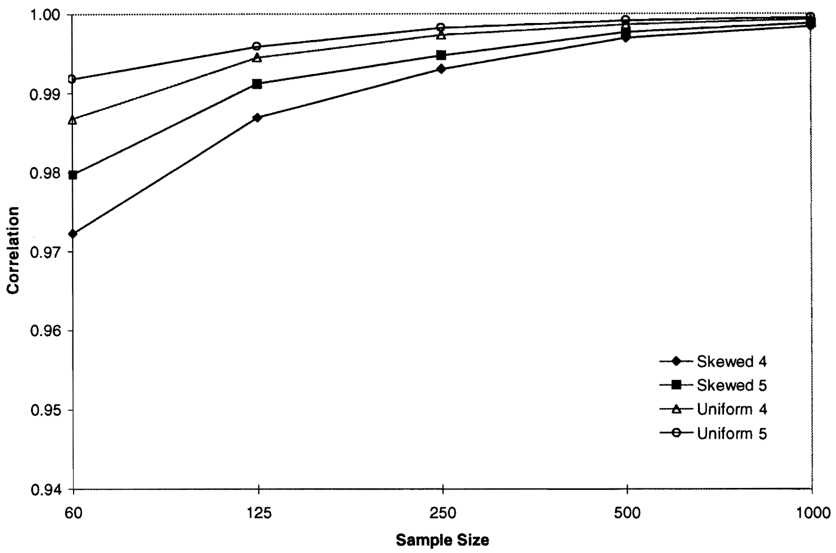


Figure 7. Correlation between estimated and true scale value parameters averaged across replications for the normal distribution of known trait level condition.

average correlation coefficients point to a superiority of the uniform item pool. Correlation coefficients ranged from a low of .9723 (skewed, 4 response categories, N=60) to a high of .9995 (uniform, 5 response categories, N=1000). The Bias index ranged from -.0322 to .0126. No clear trend emerged from analysis of the Bias indices.

Skewed datasets. The RMSE, correlation, and Bias indices averaged across replications for the scale value parameters in the skewed distribution of trait level condition are presented in Table 9. The same results collapsed according to the main effects of the experimental conditions are presented in Table 10. As was found for the normal datasets, sample size had a greater impact on the scale value parameter recovery than did the distribution of the scale values or the number of response categories. The RMSEs ranged from .0347 (skewed, 5 response categories, N = 1,000) to .1909 (uniform, 4 response categories, N = 60). The RMSEs were smaller for the five response category conditions than the corresponding four response category conditions. The smallest differences occurred for sample sizes of 1,000, while the largest differences were found for sample sizes of 60. These results are depicted graphically in Figure 8.

The results obtained for the correlation coefficients are presented in Figure 9. Sample size and number of response categories influenced the magnitude of the correlation coefficients.

Table 9

RMSE, Bias, and Correlation Indices for Scale Value Recovery Averaged Across Replications for the Skewed Distribution of Known Trait Level Condition

Distribution of Items	Number of Categories	Number of Simulees	RMSE	Correlation	Bias
Skewed	**4**	**60**	0.1556	0.9780	2.48E-05
		125	0.1110	0.9888	2.70E-05
		250	0.0812	0.9940	2.74E-05
		500	0.0592	0.9968	4.34E-05
		1000	0.0390	0.9986	2.73E-05
	5	**60**	0.1417	0.9817	1.46E-05
		125	0.1027	0.9903	3.21E-05
		250	0.0639	0.9963	3.02E-05
		500	0.0479	0.9979	3.68E-05
		1000	0.0347	0.9989	2.00E-05
Uniform	**4**	**60**	0.1909	0.9874	-1.52E-05
		125	0.1329	0.9939	1.63E-05
		250	0.0940	0.9970	9.93E-06
		500	0.0635	0.9986	3.02E-06
		1000	0.0504	0.9991	3.53E-06
	5	**60**	0.1692	0.9901	-5.43E-06
		125	0.1079	0.9960	-3.09E-06
		250	0.0808	0.9978	-3.85E-06
		500	0.0537	0.9990	-1.73E-05
		1000	0.0401	0.9994	1.22E-05

Table 10

Main Factor Indices for Scale Value Recovery for the Skewed Distribution of Known Trait Level condition

		RMSE	Correlation	Bias
Skewed		0.0837	0.9950	2.84E-05
Uniform		0.0983	0.9974	1.10E-08
Four Categories		0.0978	0.9957	1.67E-05
Five Categories		0.0843	0.9969	1.16E-05
Sample Size:	**60**	0.1644	0.9850	4.69E-06
	125	0.1136	0.9928	1.81E-05
	250	0.0800	0.9965	1.59E-05
	500	0.0561	0.9982	1.65E-05
	1000	0.0410	0.9990	1.58E-05

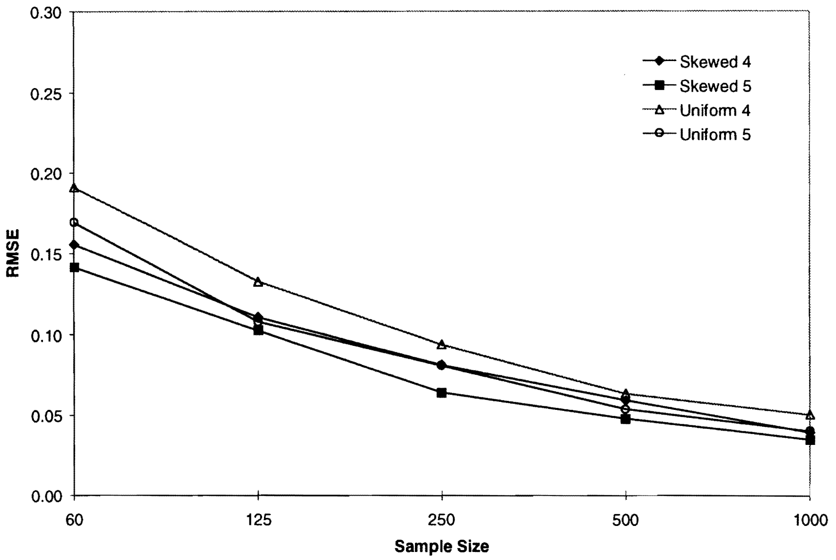


Figure 8. RMSEs for the scale value parameter estimates averaged across replications for the skewed distribution of known trait level condition.

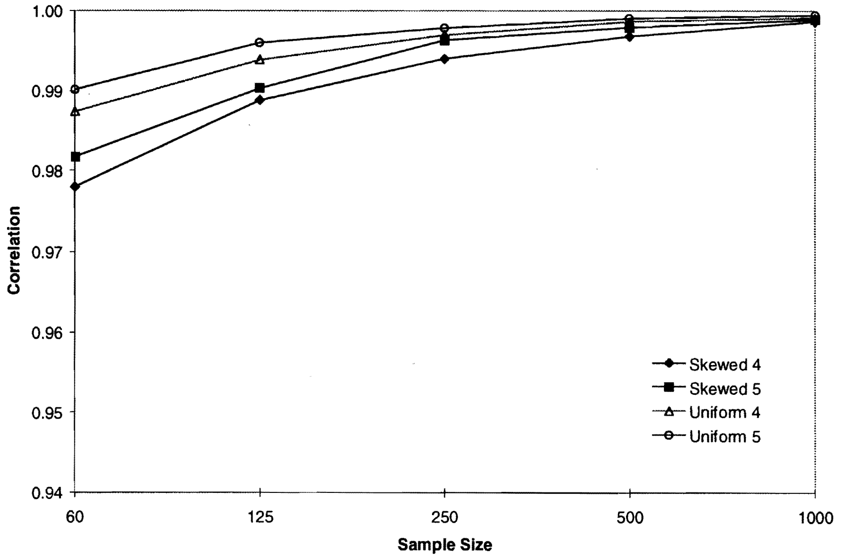


Figure 9. Correlation between estimated and true scale value parameters averaged across replications for the skewed distribution of known trait level condition.

Item Parameters: Thresholds (t_{ik})

Normal Datasets. Tables 11 and 12 contain the results from the analyses of the threshold recovery in the normal distribution of known trait level condition. The only index calculated for the thresholds was RMSE because each threshold was treated separately, and therefore only ten estimates existed per condition. Table 11 contains the RMSE average across the ten replications for each experimental condition. Table 12 contains the main effects for each of the manipulated variables. Results show that as sample size increases from 60 to 1000, the estimation of all thresholds generally improves: from a high of .1426 (skewed, 5 response categories, N=60) to .0070 (uniform, 5 response categories, N=1000). For the first threshold, RMSEs provide no evidence of an item pool distribution effect upon recovery. For all sample sizes, the four response category condi-

Table 11

RMSEs for Threshold Recovery Averaged Across Replications for the Normal Distribution of Trait Level Condition

Distribution of Items	Number of Categories	Number of Simulees	RMSE			
			t_1	t_2	t_3	t_4
Skewed	**4**	**60**	0.0547	0.0664	0.0749	–
		125	0.0481	0.0623	0.0959	–
		250	0.0360	0.0246	0.0475	–
		500	0.0305	0.0323	0.0237	–
		1000	0.0134	0.0171	0.0188	–
	5	**60**	0.1050	0.0745	0.1426	0.1170
		125	0.0555	0.0206	0.0749	0.1098
		250	0.0509	0.0253	0.0511	0.0721
		500	0.0223	0.0369	0.0232	0.0452
		1000	0.0191	0.0201	0.0179	0.0211
Uniform	**4**	**60**	0.0562	0.0614	0.0812	–
		125	0.0456	0.0336	0.0547	–
		250	0.0278	0.0347	0.0436	–
		500	0.0248	0.0311	0.0241	–
		1000	0.0216	0.0205	0.0171	–
	5	**60**	0.0872	0.0913	0.0773	0.0742
		125	0.0679	0.0757	0.0414	0.0624
		250	0.0560	0.0487	0.0339	0.0662
		500	0.0314	0.0259	0.0262	0.0214
		1000	0.0262	0.0159	0.0209	0.0070

Table 12

Main Factor RMSEs for Threshold Recovery for the Normal Distribution of Known Trait Level Condition

		RMSE			
		t_1	t_2	t_3	t_4
Skewed		0.0436	0.0380	0.0570	0.0730
Uniform		0.0445	0.0439	0.0420	0.0462
Four Categories		0.0359	0.0384	0.0482	–
Five Categories		0.0522	0.0435	0.0509	0.0596
Sample Size:	**60**	0.0758	0.0734	0.0940	0.0956
	125	0.0543	0.0480	0.0667	0.0861
	250	0.0427	0.0333	0.0440	0.0692
	500	0.0272	0.0316	0.0243	0.0333
	1000	0.0201	0.0184	0.0187	0.0140

tions show lower RMSEs than the five response category conditions. The RMSEs for the second threshold also indicate little impact of item pool distribution upon threshold parameter recovery. Additionally, the effect of the number of response categories is less clear than for the first threshold. RMSEs for the third threshold show only weak effects due to item pool distribution and the number of response categories. Finally, the fourth threshold results show a superiority of the uniform distribution relative to the skewed distribution of scale values.

Skewed Datasets. Table 13 reports the average RMSE for each experimental condition for the skewed known trait levels. Most of the RMSE values are quite small. Inspection of Table 14, which presents the same results for each main effect of the study that was manipulated for the skewed distribution of known trait levels, shows that the type of distribution of scale values appears to have little impact on the threshold parameter recovery. The RMSE is about the same for the uniform and skewed distribution of scale values conditions for each threshold parameter, respectively. No clear pattern emerges concerning the number of response categories or the sample size variables.

Discussion

Consistent with the findings of the studies by Reise and Yu (1990) and Choi et al. (1997), sample size did not affect the recovery of the trait parameters. Similar to the findings of other recovery studies, it was also found that the presence of more response categories resulted in slightly

Table 13

RMSEs for Threshold Recovery Averaged Across Replications in the Skewed Distribution of Known Trait Level Condition

Distribution	Number of	Number of	RMSE			
of Items	**Categories**	**Simulees**	**t_1**	**t_2**	**t_3**	**t_4**
Skewed	**4**	**60**	-0.0549	0.0205	0.0344	–
		125	-0.0395	-0.0178	0.0573	–
		250	-0.0660	0.0007	0.0652	–
		500	-0.0544	0.0027	0.0517	–
		1000	-0.0513	-0.0019	0.0533	–
	5	**60**	-0.0918	-0.0225	0.0429	0.0714
		125	-0.0597	0.0032	0.0054	0.0511
		250	-0.0301	-0.0389	0.0077	0.0613
		500	-0.0463	-0.0188	0.0048	0.0603
		1000	-0.0571	-0.0247	0.0145	0.0673
Uniform	**4**	**60**	-0.0665	-0.0395	0.1060	–
		125	-0.0497	0.0136	0.0360	–
		250	-0.0838	0.0101	0.0737	–
		500	-0.0584	0.0018	0.0566	–
		1000	-0.0521	0.0057	0.0464	–
	5	**60**	-0.0847	-0.0318	0.0290	0.0875
		125	-0.0440	-0.0342	-0.0138	0.0921
		250	-0.0346	-0.0325	0.0208	0.0463
		500	-0.0496	-0.0321	0.0192	0.0625
		1000	-0.0475	-0.0227	0.0124	0.0577

more accurate theta estimation than in the case of items with fewer response categories. The difference found in the current study was so small, however, that it is doubtful that there is a practical importance. It is quite possible that the number of response categories could have made a larger difference if more levels of the variable had been studied. Given the levels of variables investigated in the current study, however, it can be concluded that PARSCALE provided good trait estimation.

The PARSCALE program also provided good estimation of the scale values across most conditions researched. The highest RMSE under any condition for scale value recovery was 0.2074 and the lowest correlation was 0.9723. Each of these values in an absolute sense is competitive with the RMSE and correlation values reported in previous studies. However, as pointed out by some of those previous studies, rather than look strictly at the values of the recovery indices, one should look at sample size–to–parameter ratios. Even looking at the results of the current study in this

Table 14

Main Factor RMSEs for Threshold Recovery for the Skewed Distribution of Known Trait Level Condition

		RMSE			
		t_1	t_2	t_3	t_4
Skewed		-0.0551	-0.0098	0.0337	0.0623
Uniform		-0.0571	-0.0162	0.0386	0.0692
Four Categories		-0.0577	-0.0004	0.0581	–
Five Categories		-0.0545	-0.0255	0.0143	0.0658
Sample Size:	**60**	-0.0745	-0.0183	0.0531	0.0795
	125	-0.0482	-0.0088	0.0212	0.0716
	250	-0.0536	-0.0152	0.0419	0.0538
	500	-0.0522	-0.0116	0.0331	0.0614
	1000	-0.0520	-0.0109	0.0137	0.0625

light, one could conclude that PARSCALE recovered the parameters of Andrich's rating scale model very well. As noted above, the worst recovery indices were obtained under four response category conditions with a sample size–to–parameter ratio of 1.82:1. The only conditions with a lower sample size–to–parameter ratio (1.76:1) actually resulted in better recovery indices. This result appears to contradict that found by Choi et al. (1997), who concluded that for a fixed number of parameters, an increase in the number of response categories per item requires larger sample sizes. One explanation for these results could be that the lower ratio, which was associated with five response categories, allowed for better estimation of the scale values, because the scale value of an item is a location parameter for the item and is similar to the *b* parameter of the dichotomous models.

The threshold parameters, on the other hand, confirmed the findings of Choi et al. (1997) in that the thresholds for the five response category conditions were worse than those for the four response category conditions. Thus, the warning by Choi et al. that spreading out a fixed number of people over a larger number of response categories reduces the accuracy of item parameter estimation appears to apply to the threshold parameter estimates of the rating scale model but not to the scale value parameter estimates.

De Ayala (in press) recommended a ratio of 5:1 for accurate estimation with the nominal model. Walker-Bartnick (1990) recommended a ratio of 2:1 for the partial credit model, and Choi et al. (1997) warned that sample size–to–parameter ratio considerations are more complex than previously thought and concluded that finding a good "rule of thumb" would require more research.

The results of this study certainly do not conflict with the recommendations of these other researchers, but they do illustrate some of the complexity involved in determining how the parameter estimates are affected by various real world constraints. The parameters for Andrich's rating scale model were estimated well with considerably smaller number of persons-to-item parameters ratios (as small as 1.76:1) than the ratios recommended for other item response theory models that have been investigated. It should be noted that this finding could be due to Andrich's rating scale model being a special case of the partial credit model and which means there are fewer item parameters to estimate per item. Certainly, however, more research is necessary before a general rule can be applied across item response models.

References

Andrich, D. (1978a). Application of a psychometric model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*, 581-594.

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

Bock, D. R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

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

Choi, S. W., Cook, K. F., and Dodd, B. G. (1997). Parameter recovery for the partial credit model using MULTILOG. *Journal of Outcome Measurement*, *1*, 114-142.

De Ayala, R. J. (in press). Item Parameter recovery for the nominal response model. *Applied Psychological Measurement.*

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive attitude measurement using the rating scale model. *Applied Psychological Measurement*, *14*, 355-366.

Hambleton, R. K., Jones, R. W., and Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement*, *30*, 143-155.

Hulin, C. L., Lissak, R. I., and Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

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

Muraki, E. and Bock, D. R.. (1993). *PARSCALE [computer program].* Mooreville, IN: Scientific Software.

Reise, S. P. and Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133-144.

Samejima, F. (1969). Estimation of latent ability using a response of graded scores. *Psychometrika Monograph Supplement, 17.*

Thissen, D. (1986). *MULTILOG [computer program].* Mooreville, IN: Scientific Software.

Thissen, D. (1991). *MULTILOG [computer program].* Mooreville, IN: Scientific Software.

Walker-Bartnick, L. A. (1990). *An investigation of factors affecting invariance of item parameter estimates in the partial credit model.* Unpublished doctoral dissertation, University of Maryland, College Park.

Wright, B., Congdon, R., and Shultz, M. (1988). *MSTEPS partial credit analysis [computer program].* Chicago: MESA Psychometric Laboratory, University of Chicago.

Yen, W. M. (1987). A comparison of efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 52*, 275-291.

CONTRIBUTOR INFORMATION

Content: *Journal of Outcome Measurement* publishes refereed scholarly work from all academic disciplines relative to outcome measurement. Outcome measurement being defined as the measurement of the result of any intervention designed to alter the physical or mental state of an individual. The *Journal of Outcome Measurement* will consider both theoretical and applied articles that relate to measurement models, scale development, applications, and demonstrations. Given the multi-disciplinary nature of the journal, two broad-based editorial boards have been developed to consider articles falling into the general fields of Health Sciences and Social Sciences.

Book and Software Reviews: The *Journal of Outcome Measurement* publishes only solicited reviews of current books and software. These reviews permit objective assessment of current books and software. Suggestions for reviews are accepted. Original authors will be given the opportunity to respond to all reviews.

Peer Review of Manuscripts: Manuscripts are anonymously peer-reviewed by two experts appropriate for the topic and content. The editor is responsible for guaranteeing anonymity of the author(s) and reviewers during the review process. The review normally takes three (3) months.

Manuscript Preparation: Manuscripts should be prepared according to the *Publication Manual of the American Psychological Association* (4th ed., 1994). Limit manuscripts to 25 pages of text, exclusive of tables and figures. Manuscripts must be double spaced including the title page, abstract, text, quotes, acknowledgments, references, and appendices. On the cover page list author name(s), affiliation(s), address(es), telephone number(s), and electronic mail address(es). On the second page include a 100 to 150 word abstract. Place tables on separate pages. Include photocopies of all figures. Number all pages consecutively.

Authors are responsible for all statements made in their work and for obtaining permission from copyright owners to reprint or adapt a table or figure or to reprint a quotation of 500 words or more. Copies of all permissions and credit lines must be submitted.

Manuscript Submission: Submit four (4) manuscript copies to Richard M. Smith, Editor, *Journal of Outcome Measurement*, Rehabilitation Foundation Inc., P.O. Box 675, Wheaton, IL 60189 (e-mail:JOMEA@rfi.org). Prepare three copies of the manuscript for peer review by removing references to author(s) and institution(s). In a cover letter, authors should indicate that the manuscript includes only original material that has not been previously published and is not under review elsewhere. After manuscripts are accepted authors are asked to submit a final copy of the manuscript, original graphic files and camera-ready figures, a copy of the final manuscript in WordPerfect format on a 3 ½ in. disk for IBM-compatible personal computers, and sign and return a copyright-transfer agreement.

Production Notes: Manuscripts are copy-edited and composed into page proofs. Authors review proofs before publication.

SUBSCRIBER INFORMATION

Journal of Outcome Measurement is published four times a year and is available on a calendar basis. Individual volume rates are $35.00 per year. Institutional subscriptions are available for $100 per year. There is an additional $24.00 charge for postage outside of the United States and Canada. Funds are payable in U.S. currency. Send subscription orders, information requests, and address changes to the Subscription Services, Rehabilitation Foundation, Inc. P.O. Box 675, Wheaton, IL 60189. Claims for missing issues cannot be honored beyond 6 months after mailing date. Duplicate copies cannot be sent to replace issues not delivered due to failure to notify publisher of change of address. Back issues are available at a cost of $12.00 per issue postpaid. Please address inquiries to the address listed above.