Optimizing Rating Scale Category Effectiveness

John M. Linacre
MESA Psychometric Laboratory
University of Chicago

Journal of Applied Measurement 3:1 2002 p.85-106.

Abstract

Rating scales are employed as a means of extracting more information out of an item than would be obtained from a mere “yes/no”, “right/wrong” or other dichotomy. But does this additional information increase measurement accuracy and precision? Eight guidelines are suggested to aid the analyst in optimizing the manner in which rating scales categories cooperate in order to improve the utility of the resultant measures. Though these guidelines are presented within the context of Rasch analysis, they reflect aspects of rating scale functioning which impact all methods of analysis. The guidelines feature 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.

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 & 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 & 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 (Rasch, 1960) 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)

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

where

P_{nik} is the probability that person *n*, on encountering item *i* would be observed (or would respond) in category *k*,

$P_{ni(k-1)}$ is the probability that the observation (or response) 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*.

The "step calibration", F_k, is a rating scale threshold defined as the location corresponding to the equal probability of observing adjacent categories *k-1* and *k*. This contrasts with a Thurstone threshold which is defined in terms of the equal probability of observing categories *0, k-1* and *k, m*.

Under "partial credit" conditions, this model is often parameterized as:

$$\log (P_{nik} / P_{ni(k-1)}) \bullet B_n - D_{ik} \qquad (2)$$

but, for convenience, we reparameterize D_{ik} as $D_i + F_{ik}$, made identifiable by the constraints: $\Sigma F_{ik}=0$ and $\Sigma D_{ik}=\Sigma D_i$ for $k=1, m$.

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 about here]

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 about here]

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 & Masters, 1982, p. 81).

[Figure 3 about here]

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 is to verify 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 measurement qualitites of a test. Mathematically it can be proven that, when the data fit the Rasch model according to one categorization, then the fit of any other categorization to the model must be inferior (Jansen & 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 & Linacre, 1992; Linacre, 1999). 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. Some guidelines are now presented to assist the analyst in examining and improving the quantitative functioning of a rating scale. The only consideration here is the numerical information which indicates to what extent the data produce coherent raw scores, i.e., raw scores that support the construction of Rasch measures. Indeed, the description of the characteristics of an ideal rating scale, presented above, suggests an explicit procedure for verifying useful functioning and diagnosing malfunctioning of an empirical rating scale. Consequently, the guidelines are presented in the order in which the analytical concerns are encountered. Later guidelines are typically only relevant to the analysis of instruments under development.

[Table 1 about here]

[Figure 4 about here]

[Table 2 about here]

Consider the Ratings of Creativity (Guilford, 1954), also discussed from a different perspective in Linacre (1989b). Table 1 contains results from an analysis by *Facets* (Linacre, 1989a). 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 & Linacre, 1991). The model category characteristic curves for the LFS data are shown in Figure 5.

[Figure 5 about here]

Preliminary Guideline: All items oriented with latent variable.

It is common practice on survey and assessment instruments to have blocks of items employ the same rating scale. Other items may employ rating scales unique to each item. In whatever way the rating protocol is designed, it is essential that during the analysis all items and rating scales cooperate to construct one shared latent variable. A conventional strategem, however, is to deliberately reverse the polarity of some items. This may not be noticed when rating scale statistics summarize a block of items. Accordingly, item-level indices of polarity, such as point-biserial correlations, must be inspected before investigation of the functioning of the rating scale begins. Items whose polarities contradict the general item consensus must be rescored to accord with that consensus. Since the rating scale categories may function differently for negatively-oriented items (Yamaguchi, 1997), separate summary statistics for the positively- and negatively-oriented forms of each rating scale are required.

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.

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 overall 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. A line of argument based on dichotomous items suggests that 25 or even 50 observations per category may be required depending on the degree of stability desired. This would make the minimum sample size for stability at least 25*(m+1) subjects, and perhaps as many as 100*(m+1) subjects if use of categories is not uniform across the rating scale.

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 can 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 categories with incidental zeroes as structural zeroes for this analysis, renumbering the observed categories so as to omit unobserved categories. For example, if only categories 1, 4 and 5 are observed. These are analyzed as though they are categories 1, 2 and 3. This approach is inferentially insecure because next time different categories may be observed.

ii) Impose a scale structure (by anchoring thresholds or otherwise) that includes these categories. For example, a reasonably initial assertion might be that each succeeding category is twice as challenging as its predecesor. Then the $\{F_k\}$ would increase sequentially by 1.1 logits. Since the $\{F_k\}$ also conventionally sum to zero, their values for a 5-category scale could be pre-set (anchored) at -1.65, -.55, .55 and 1.65.

iii) Use a mathematical device to keep intermediate zero categories in the analysis. Wilson (1991) avoids estimating the F_k for a zero category (which would be minus-infinity) and the adjacent F_{k+1} (which would be plus-infinity), by estimating $F_k + F_{k+1}$ by means of

$$\log (P_{ni(k+1)} / P_{ni(k-1)}) \bullet 2(B_n - D_i) - (F_k+F_{k+1}) \qquad (3)$$

Sheridan & Puhl (1996) fit a polynomial function, of lower order than the number of observed categories, to the {Fk}. This enables all coefficients of the polynomial to be estimated, and so all $\{F_k\}$ to be assigned finite values.

Applying this guideline is usually strightforward. 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 labelled “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 categories 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, totalling 33, also 4 and 5, totalling 39, and then 6, 7, 8, and 9, totalling 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:

$$B_n - D_i - \{F_k\} \bullet X_{ni} \qquad (3)$$

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 crucial diagnostic indicator is the average of the measure differences, $\{(B_n\text{-}D_i)\}$, 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 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 even 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 & 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, equivalent to a conventional statistical chi-square statistic, this misinformation may be confined to a few substantively explainable and easily remediable observations. Nevertheless large mean-squares do indicate that the 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 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 loosing information. Extreme category 2 with mean-square 1.46 is somewhat noisy. This high value is cause by a mere 6 of the 852 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 increasing probabilities of the respondent being observed in higher categories of the rating scale. Andrich (1996) asserts this not merely for the entire rating scale, but for individual categories. Thus, Andrich asserts that as measures increase, or as individuals with incrementally higher measures are observed, each category of the scale in turn must be (or has been) designed to be most likely to be chosen. This assertion 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 requirement for this type of 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 being used in the rating process. Thus, step disordering degrades the interpretability of the resulting measures. It can indicate that a category represents too narrow a segment of the latent variable or corresponds to 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,

$$F_k \bullet \log (T_{k-1}/T_k) - \qquad B_k + B_{k-1} \tag{4}$$

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.

Sheridan and Puhl (1996) take this consideration further. They reparameterize F_k thus:

$$F_k = x k(m-k) + y k(m-k)(2k-m) + z k(m-k)(5x^2-5xm+m^2+1) \qquad (5)$$

where x is the spread of the rating scale parameters, y is their skewness and z is their kurtosis. F_k, k and m are as above.

Using spread, skewness and kurtosis parameters can be productive in bridging over missing categories, and interpreting the structure of long indistinct rating scales. Thus it simplifies analysis and interpretation of the rating scale used by respondents in response to the question, “On a scale from 1 to 10, how much pain do you feel?” On the other hand, it obscures analysis of rating scales in which each category is specifically defined.

Ordered steps also imply that the rating scale can be expressed as sets of dependent Guttman dichotomies. Consider the rating scale with m+1 categories, scored, for convenience, 0 to m. If the m step calibrations are ordered, then this is equivalent to m dichotomies. A rating of k is equivalent to success on the first k dichotomies and failure on the last m-k. The difficulty of the kth dichotomy is D_i+F_k.

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 about here]

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

[Table 3 about here]

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.

[Figure 7 about here]

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 or 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 somewhat fragile, but still acceptable. 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 independent dichotomous items is always 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 independent dichotomous items under specific conditions (Huynh, 1996).

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.

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.

[Table 4 about here]

Conclusion

Table 4 summarizes the relevance of the guidelines to aspects of test development, analysis and inference. 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 & 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. & Roskam, E.E. (1984) The polytomous Rasch model and dichotomization of graded responses. p. 413-430. in E. Degreef & 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. (1989a) *Facets computer program for many-facet Rasch measurement.* Chicago: MESA Press.

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

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

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

Linacre J.M. (1999) Investigating rating scale category utility. *Journal of Outcome Measurement,* 3:2, 103-122.

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. & 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.

Sheridan, B. & Puhl, L. (1996) Evaluating an indirect measure of student literacy competencies in HIgher Education using Rasch measurement. Chap. 2 in G. Engelhard, Jr. & M. Wilson (Eds.), Objective Measurement: Theory into Practice. Vol. 3. Norwood, NJ: Ablex.

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

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

Stone M. H. & 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. & Linacre, J.M. (1992) Combining and splitting categories. *Rasch Measurement Transactions*, 6, 3, 233-235.

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

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

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

Yamaguchi J. (1997) Positive vs. Negative Wording. Rasch Measurement Transactions 11:2 p. 567.

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

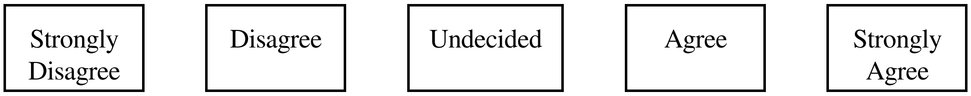


Figure 1. Prototypical Likert scale as presented to the respondent.

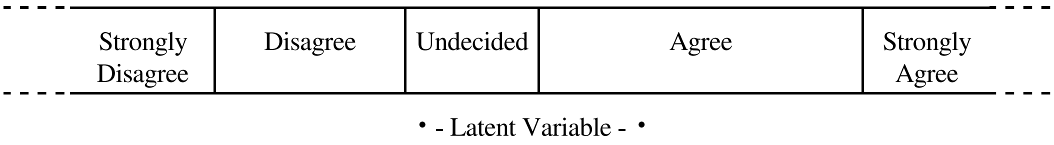


Figure 2. Prototypical Likert scale from a measurement perspective.

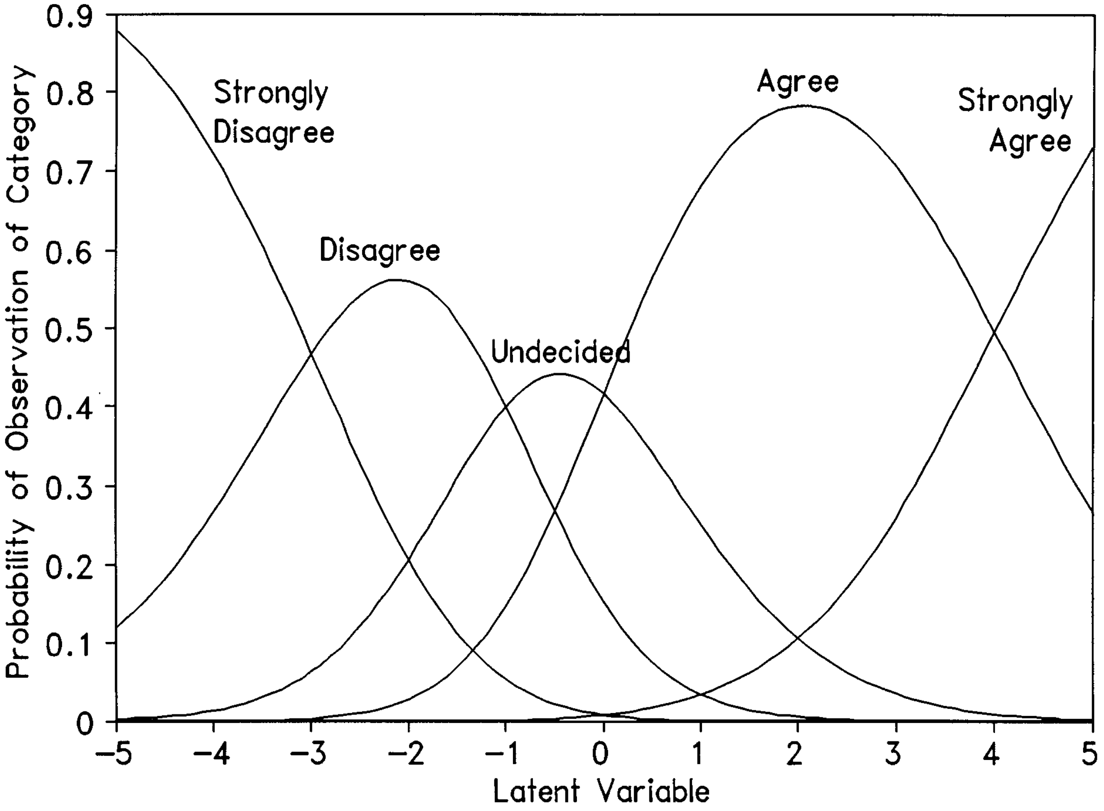


Figure 3. Category probability curves for 5 category Likert scale.

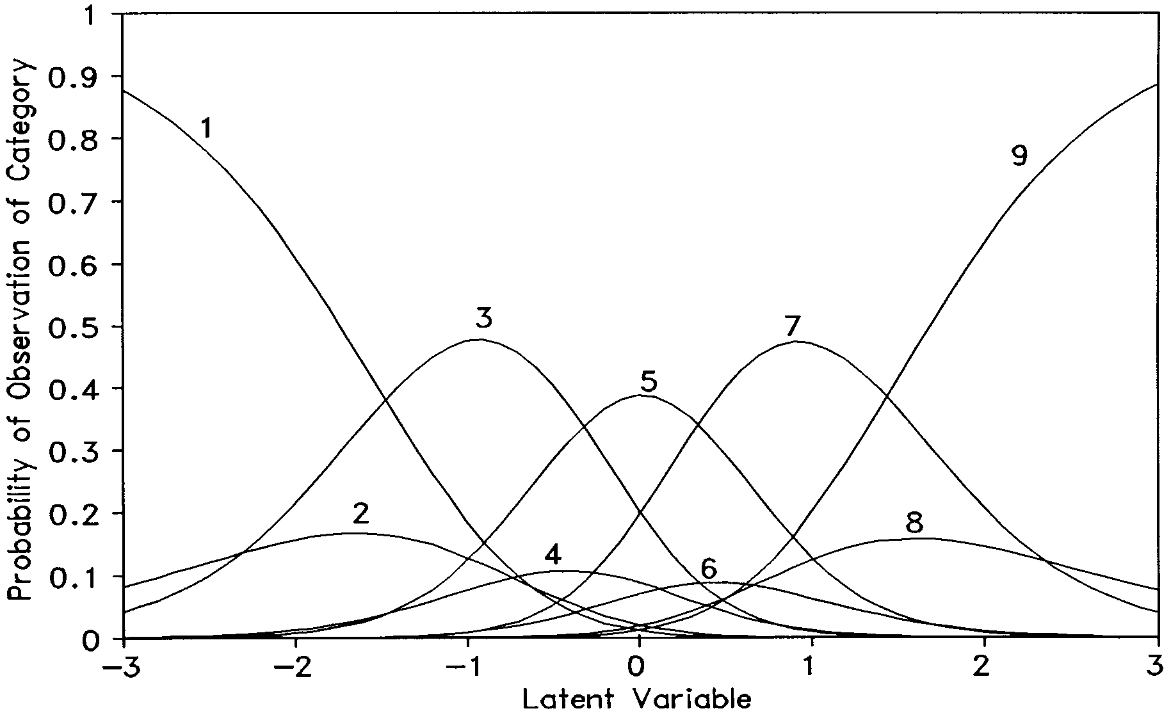


Figure 4. Rasch model category probability curves for Guilford's (1954) scale.

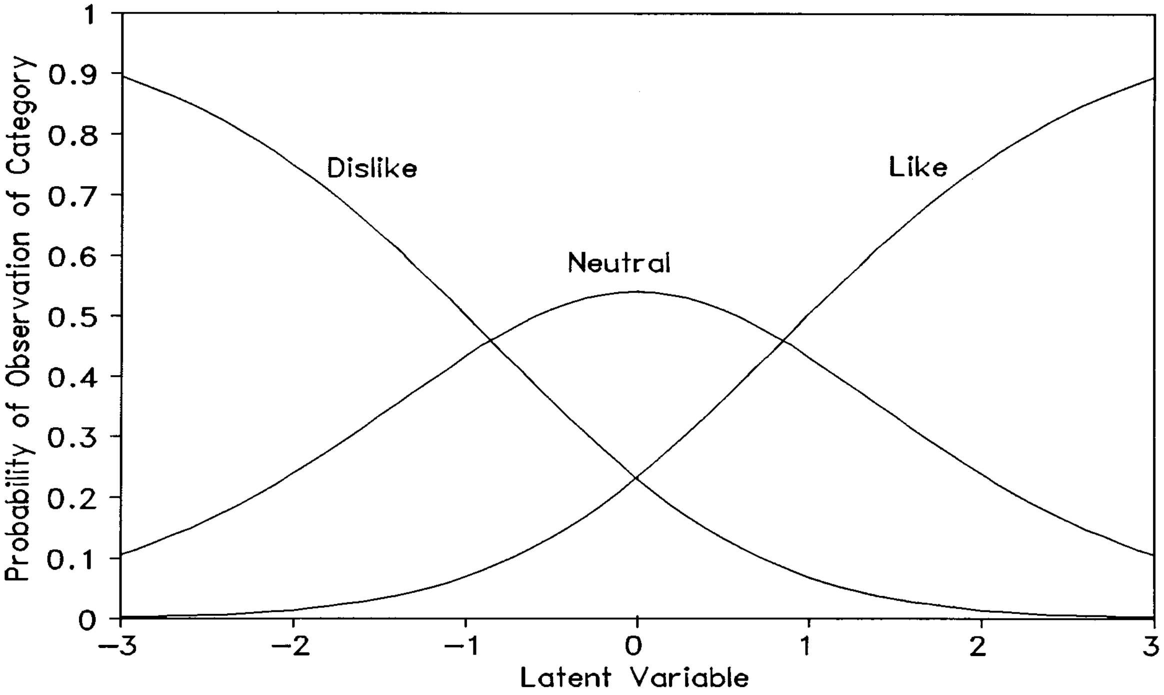


Figure 5. Model probability characteristic curves for LFS rating scale.

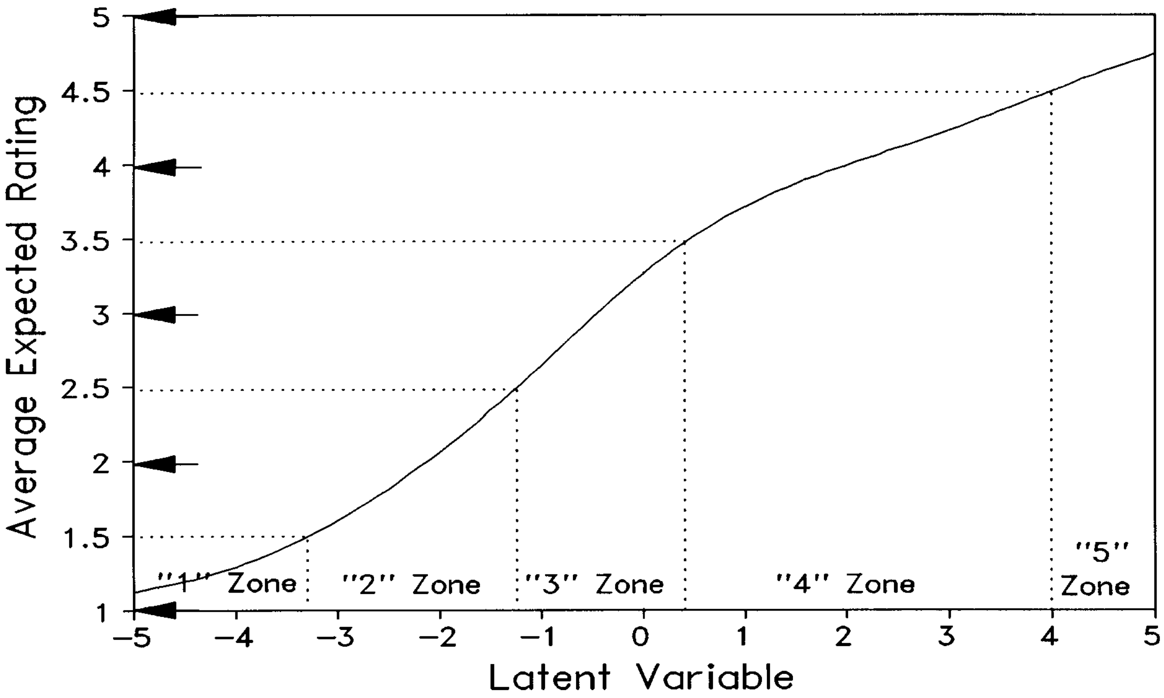
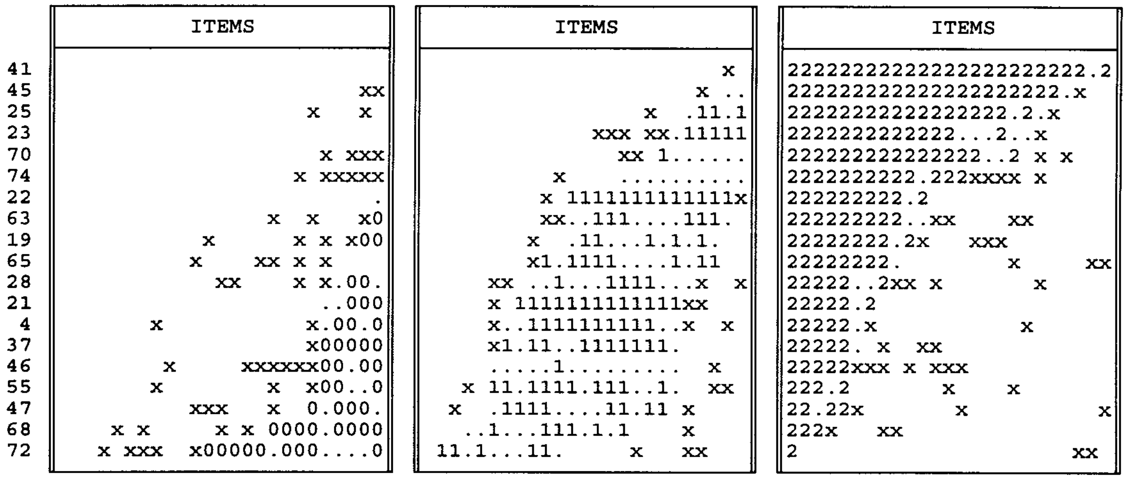


Figure 6. Expected score ogive for 5 category Likert scale showing rating-measure zones.



```
            ITEMS                     ITEMS                     ITEMS
41                                                  x    22222222222222222222222.2
45                        xx                     x ..    222222222222222222222.x
25                    x   x                  x  .11.1    22222222222222222.2.x
23                                       xxx xx.11111    2222222222222...2..x
70                     x xxx               xx 1......    222222222222222..2 x x
74                   x xxxxx          x    ..........    2222222222.222xxxx x
22                         .         x 1111111111111x    222222222.2
63                 x  x   x0         xx..111....111.     222222222..xx    xx
19            x      x x x00        x  .11...1.1.1.      22222222.2x   xxx
65           x    xx x x            x1.1111....1.11      22222222.        x     xx
28             xx    x x.00.     xx ..1...1111...x  x    22222..2xx x       x
21                     ..000     x 1111111111111xx       22222.2
 4        x           x.00.0     x..1111111111..x  x     22222.x           x
37                    x00000     x1.11..1111111.         22222. x  xx
46         x     xxxxxx00.00     .....1.........  x      22222xxx x xxx
55        x        x  x00..0   x 11.1111.111..1.  xx     222.2       x    x
47           xxx   x  0.000.  x  .1111....11.11 x        22.22x       x          x
68     x x     x x 0000.0000   ..1...111.1.1    x        222x   xx
72    x xxx  x00000.000....0  11.1...11.     x   xx      2                     xx
```

Figure 7. Excerpt from scalograms of LFS data. “x” indicates out-of-zone observations.

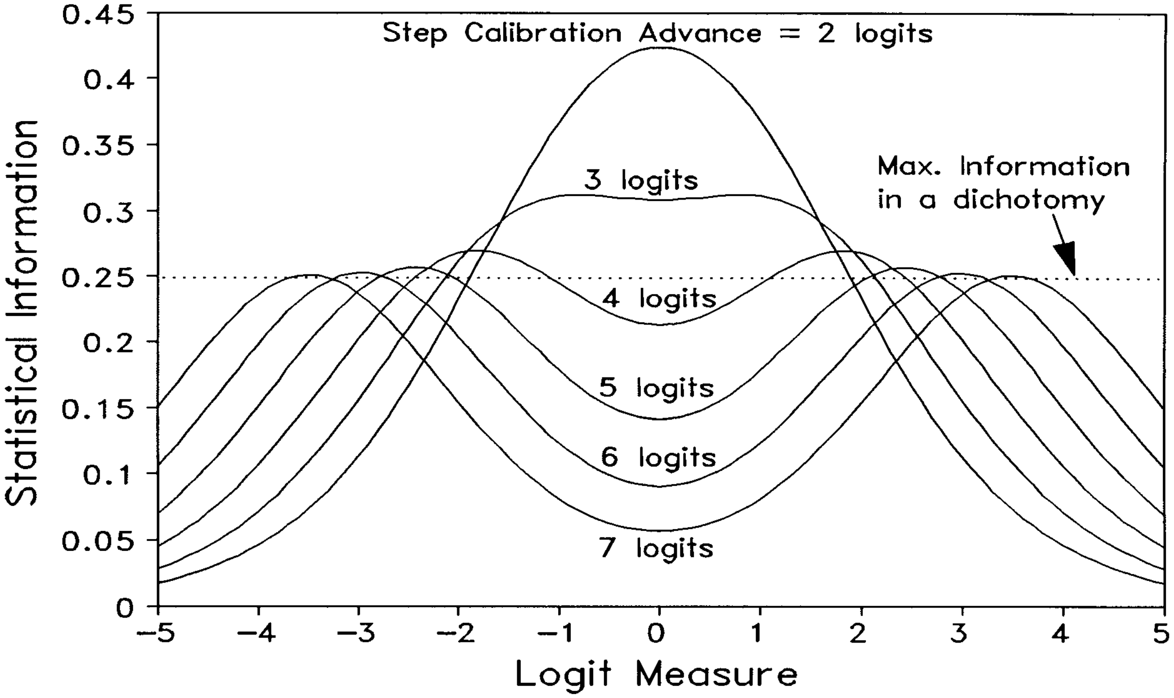


Figure 8. Information functions for a three-category rating scale.

Category Label	Category Count	Category %	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

Table 1. Analysis of Guilford's (1954) rating scale.

Category Label	Category Count	Average Measure	Expected Measure	OUTFIT MNSq	Step Calibration	Coherence M->C	Coherence C->M	Zone: From	Zone: To
0	378	-.87	-1.03	1.19	-	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	+∞

Table 2. Analysis of LFS rating scale data.

<table>
<tr><th></th><th>Observed Rating
<b>in</b> Category</th><th>Observed Rating
<b>outside</b> Category</th></tr>
<tr><td>Observed Measure
<b>in</b> Zone</td><td>ICIZ
(Rating in Figure 7)</td><td>OCIZ
(“.” in Figure 7)</td></tr>
<tr><td>Observed Measure
<b>outside</b> Zone</td><td>ICOZ
(“x” in Figure 7)</td><td>(included in other categories and zones)</td></tr>
<tr><td colspan="3">M->C = In Category & Zone / All in Zone =
100 * ICIZ / (ICIZ + OCIZ) %</td></tr>
<tr><td colspan="3">C->M = In Category & Zone / All in Category =
100 * ICIZ / (ICIZ + ICOZ) %</td></tr>
</table>

Table 3. Coherence of Observations.

	Guideline	Measure Stability	Measure Accuracy (Fit)	Description of this sample	Inference for next sample
Pre.	Scale oriented with latent variable	Essential	Essential	Essential	Essential
1.	At least 10 observations of each category.	Essential	Helpful		Helpful
2.	Regular observation distribution.	Helpful			Helpful
3.	Average measures advance monotonically with category.	Helpful	Essential	Essential	Essential
4.	OUTFIT mean-squares less than 2.0.	Helpful	Essential	Helpful	Helpful
5.	Step calibrations advance.				Helpful
6.	Ratings imply measures, and measures imply ratings.		Helpful		Helpful
7.	Step difficulties advance by at least 1.4 logits.				Helpful
8.	Step difficulties advance by less than 5.0 logits	Helpful			

Table 4. Summary of Guideline Pertinence.

<End of Manuscript>