Review of Educational Research
Spring 2001, Vol. 71, No. 1, pp. 105–131

Rescaling Ordinal Data to Interval Data in Educational Research

Michael R. Harwell
University of Minnesota
Guido G. Gatti
University of Pittsburgh Medical Center

Many statistical procedures used in educational research are described as requiring that dependent variables follow a normal distribution, implying an interval scale of measurement. Despite the desirability of interval scales, many dependent variables possess an ordinal scale of measurement in which the differences among values composing the scale are unequal in terms of what is being measured, permitting only a rank ordering of scores. This means that data possessing an ordinal scale will not satisfy the assumption of normality needed in many statistical procedures and may produce biased statistical results that threaten the validity of inferences. This article shows how the measurement technique known as item response theory can be used to rescale ordinal data to an interval scale. The authors provide examples of rescaling using student performance data and argue that educational researchers should routinely consider rescaling ordinal data using item response theory.

Many statistical procedures used in educational research are described as requiring that dependent variables be normally distributed, which implies that these variables possess an interval scale of measurement. The advantage of an interval scale is that relative differences among values composing the scale are assumed to be equal in terms of what is being measured, allowing arithmetic operations (e.g., addition, multiplication) to be used unambiguously. For example, suppose that scores on a 30-item test were computed by summing the number of items answered correctly, where higher scores are intended to reflect greater proficiency. If the test score variable possesses an interval scale, then the difference in proficiency reflected in scores of 10 and 15 is exactly the same as the difference in proficiency reflected in scores of 15 and 20.

Despite the desirability of interval-scaled variables, we share the view of Clogg and Shihadeh (1994, p. 140) that "perhaps most of the variables specified as dependent variables in social research are of this general kind [ordinal]." In an ordinal scale, the relative differences among values composing the scale are unequal in terms of what is being measured, permitting only a rank ordering of scores. For example, a Likert scale in which parents respond to a question about their support for current school attendance policy by marking "1 = none," "2 = weak," "3 = moderate," or

"4 = strong" reflects a rank ordering and is probably an ordinal scale because the difference between 1 and 2 (one unit) probably does not, in terms of intensity of parental support, equal the one-unit difference between 2 and 3. Detailed descriptions of ordinal and interval scales can be found in Siegel (1956), Stevens (1951), and Zumbo and Zimmerman (2000).

Some readers may quarrel with our characterization of ordinal and interval scales, but we suspect that few would quarrel with our preference for interval-scaled variables over those showing an ordinal scale. Our interest in this topic stems from our perception that educational researchers frequently employ ordinal-scaled dependent variables in statistical procedures that assume that these variables possess an interval scale of measurement.

Measurement Scales and Statistical Procedures

Concerns over ordinal data are not new and have been the subject of considerable debate in the methodological literature, much of it centered on the work of Stevens (1946, 1951, 1968, 1971). Stevens (1946) described four scales of measurement for variables—nominal, ordinal, interval, and ratio—that are familiar to most educational researchers. This hierarchy links the scale of a variable with the kinds of statistical analyses that can be performed and is based on the idea of permissible and impermissible scale transformations. Following Stevens, a scale transformation is an arithmetic operation that transforms or rescales the original data. For example, suppose that a constant, say, 20, was added to each Y_i score using the expression $Y_i^* = Y_i + 20$. This represents a linear transformation because the rescaling factor (20) is raised to a power of one. On the other hand, taking the logarithm of each score using the expression $Y_i^* = \log(Y_i)$ represents a monotonic transformation because rank order is preserved; that is, the examinee with the largest Y_i also has the largest Y_i^*, the examinee with the second largest Y_i also has the second largest Y_i^*, and so on.

According to Stevens, permissible transformations for ordinal data are those that are monotonic; thus, for example, log and linear transformations are permissible. For an interval scale, permissible transformations are those that preserve relative differences. Since linear transformations preserve relative differences, they are permissible and can safely be performed on interval data. On the other hand, log transformations do not preserve relative differences and are inappropriate for interval-scaled data.

As an example, consider the results reported by Resnick and Harwell (1998) for the New Standards reference exams (New Standards, 1997), in which examinees receive one of five grades for each of four English language arts clusters. Assignment of a grade to each examinee for each standards cluster was based on examinees' responses to dichotomously scored and constructed-response items. The possible grades for each cluster were as follows: achieved the standard with honors = 5, achieved the standard = 4, nearly achieved the standard = 3, below the standard = 2, and little evidence of achievement = 1. These values were then summed to produce an overall English proficiency score for each examinee with a possible score range of 4 to 20. These values are similar to those in many educational research studies, for example, Grolnick, Benjet, Kurowski, and Apostoleris (1997), and clearly represent ordinal data.

To relate scores on this variable to Stevens's scales of measurement, suppose that the New Standards English proficiency scores for three examinees were

$Y_1 = 6$, $Y_2 = 12$, and $Y_3 = 18$, producing a rank ordering of $Y_1 < Y_2 < Y_3$. Suppose also that a log transformation was performed on these scores, producing 2.079, 2.773, and 2.890, respectively. If Y possesses an ordinal scale, this transformation is permissible (following Stevens) because rank order is maintained, that is, $\log (Y_1) < \log (Y_2) < \log (Y_3)$; if Y is interval scaled, the log-transformed values 2.079, 2.773, and 2.890 no longer permit an inference of relative differences (i.e., differences among the log-transformed values are not equal along the log scale and the transformation is inappropriate). Stevens's scale typology implies that the operations of addition, subtraction, and multiplication can be performed for ordinal data because the rank order of the rescaled scores is preserved.

Stevens (1951) then coupled statistical procedures with measurement scales, which amounted to prescribing the use of certain statistical procedures with certain scales of measurement. According to Stevens, statistics that can be computed for ordinal data are those whose meanings are preserved when linear or monotonic transformations are applied to the data; only statistics whose meaning is unchanged when a linear transformation is applied are legitimate for interval-scaled data. Following this prescription, statistical procedures such as t tests and F tests should be used only for interval-scaled data; ordinal data are appropriately analyzed with procedures that require rank-order information, such as nonparametric statistical procedures. Taken literally, Stevens's prescription implies that the data analyses in many published studies that employ ordinally scaled dependent variables are inappropriate.

Stevens's coupling of measurement scales and statistical procedures provoked a firestorm among researchers in a variety of disciplines (see, e.g., Anderson, 1961; Baker, Hardyck, & Petrinovich, 1966; Borgatta & Bohrnstedt, 1980; Gaito, 1980; Gregoire & Driver, 1987; Hand, 1996; Labovitz, 1970; Michell, 1986; Townsend & Ashby, 1984; Zumbo & Zimmerman, 1993, 2000) and has been quite influential. For example, Stevens's prescriptions appear in numerous introductory statistics texts and published journal articles, as well as computer programs designed to assist researchers in choosing appropriate statistical analyses (Velleman & Wilkinson, 1993). Stevens's coupling of data analyses and the scale of measurement of a variable has garnered strong support among many researchers (Luce, Krantz, Suppes, & Tversky, 1990; Narens & Luce, 1986; Townsend & Ashby, 1984), and good summaries of Stevens's position appear in Townsend and Ashby (1984) and Maxwell and Delaney (1985).

Opponents of Stevens's position have argued that such prescriptions are unrealistic and that statistical techniques should not be held hostage to measurement scales because there is no requirement underlying these procedures that ties them to such scales. These authors point out that research literatures contain numerous examples in which something valuable was learned from data analyses that ignored Stevens's scale prescriptions (e.g., F test of means performed on ordinal data).

Along the same lines, Lord (1953) wrote a scathing criticism of scale prescriptions that argued that what counts is the meaningfulness of a statistical analysis, which depends on the questions the analysis is trying to answer. Tukey (1961) offered similar criticisms of scale prescriptions. Velleman and Wilkinson (1993) argued that scale type is rarely fixed but, rather, depends on the questions being asked or the introduction of new information. They used the example of the number of cylinders in a car engine to point out that the scale of this variable could be treated as nominal for questions involving miles per gallon and as interval for questions involving the average number of cylinders in U.S.-produced cars. Velleman and Wilkinson also reiterated one of the most potent criticisms of Stevens's prescriptions: They preclude the use of monotonic data transformations (log, square root) with interval-scaled data even though such transformations often increase the extent to which underlying statistical assumptions are satisfied by making distributions less skewed, variances more nearly equal, and so forth.

One of the more interesting techniques for ordinal data was described by Gautam, Kimeldorf, and Sampson (1996), who illustrated a transformation that produces the smallest and largest possible *F* tests for means for any assignment of values to an ordinal scale. These authors pointed out that the usual assignment of integer values to an ordinal scale (e.g., 1, 2, 3, etc.) is only one of many possibilities and that the assignment of other values can affect statistical tests. For example, it is possible that the values for a 5-point Likert-type scale that produce the largest statistical test are not 1, 2, 3, 4, and 5 but another collection such as 0, 0.23, 0.46, 0.90, and 1. Gautam et al. showed that if the minimum and maximum possible *F* tests, determined by examining sets of scale values, in a one-way fixed-effects analysis of variance (ANOVA) were both statistically significant, then one can conclude that evidence supporting the presence of a treatment effect exists and does not depend on the choice of ordinal scale values. The use of the *F* test in the Gautam et al. method would be inappropriate under Stevens's prescriptions, however.

Cogent summaries of arguments against Stevens's prescriptions appear in Gaito (1980), Zumbo and Zimmerman (1993, 2000), Borgatta and Bohrnstedt (1980), and Velleman and Wilkinson (1993). Still other writers have attempted to describe differences between those who support and those who are opposed to Stevens's prescriptions, finding strengths and weaknesses in both positions (Hand, 1996; Michell, 1986). Given the concerns with using ordinal data, it seems natural to ask how frequently ordinal data are used in educational research.

How Frequently Are Ordinal-Scaled Dependent Variables Used in Statistical Procedures Requiring Interval-Scaled Data in Educational Research?

We expected the percentage of ordinally scaled dependent variables in educational research to be high because our experience with the research of colleagues and students suggests that (a) dependent variables are often constructed to meet the needs of the research, and these variables are likely to possess an ordinal scale, and (b) when existing tests, questionnaires, and so forth are used, the scale of the resulting data is typically ordinal.

Many readers may be able to cite their own experiences in which statistical analyses usually described as requiring interval-scaled data were performed with dependent variables possessing an ordinal scale. Still, we attempted to estimate the prevalence of this practice by surveying three prominent educational research journals: *American Educational Research Journal, Sociology of Education*, and *Journal of Educational Psychology*. While not definitive, we believe that the empirical studies appearing in these journals are probably representative of a large class of quantitative research studies in education.

We classified a dependent variable in these studies as possessing an ordinal scale if there was evidence that the scores did not possess the property of equal relative differences but could be rank ordered. For example, Grolnick et al. (1997) studied predictors of parent involvement in children's schooling using a variety of dependent variables constructed for their study. Among these variables were measures of school involvement based on responses to five items using Likert scales with three, four, and five ordered-response categories. Grolnick et al. offered no evidence to support their treatment of these data as possessing an interval scale, and it seems unlikely that relative differences were equal (e.g., that the difference between scores of 3 and 6 reflected the same level of parent involvement as the difference between scores of 12 and 15).

On the other hand, a study by Rowan, Chiang, and Miller (1997) relied on estimates of students' mathematics proficiency obtained using a measurement technique known as item response theory (IRT). Under some conditions described later, IRT produces interval-scaled data, which means that relative differences of estimates of proficiency should be constant across the scale. Thus, the relative difference between examinees with proficiency estimates of, say, 0.5 and 1 (standard deviations) should reflect the same difference in mathematics proficiency as that between examinees with scores of 1.5 and 2 standard deviations.

In surveying the published articles in these three journals during the period 1993–1997, we were surprised that more than 85% of the studies reporting empirical results used hierarchical linear modeling (HLM) or structural equation modeling (SEM), although ANOVA-based procedures were also used. Overall, more than 100 studies published in these journals between 1993 and 1997 used statistical analyses usually described as requiring interval-scaled data. We felt that 100 studies was unmanageable and decided to limit the survey to articles published in 1997. Of the studies surveyed in 1997, HLM and SEM were again the most commonly used statistical techniques, followed by ANOVA-based procedures.

The *American Educational Research Journal* published 14 papers in 1997 in which statistical tests that required a formal assumption of normality were reported, usually with more than one dependent variable being analyzed. Of a total of 166 dependent variables, we judged that 88% showed an ordinal scale. Similarly, in 1997 *Sociology of Education* published 18 papers reporting results based on statistical tests requiring a normality assumption, with 56% of the 71 dependent variables appearing to show an ordinal scale; for the *Journal of Educational Psychology*, 70 papers and 71% of the 471 dependent variables showed an ordinal scale.

Overall, 73% of the dependent variables used in the articles published in these three journals in 1997 appeared to be measured using an ordinal scale. All of the dependent variables categorized as ordinal in these studies used Likert scales, and many were constructed to meet the needs of the particular study. Although there is room for disagreement about whether a specific variable in the articles surveyed was measured using an ordinal or interval scale, we believe that the overwhelming majority of these variables would be judged to represent ordinal scales of measurement regardless of whether liberal or conservative definitions are applied.

The fact that so many of the dependent variables showed an ordinal scale but were treated as interval suggests that a case can be made that educational researchers regularly employ ordinal-scaled dependent variables in analyses typically described as requiring these variables to be interval scaled. This practice can lead to problems in many statistical procedures. Before describing some of these problems, we distinguish between manifest and latent variables.

Manifest and Latent Variables

A manifest variable (Y) is a measured (observed) variable, whereas a latent variable (θ) cannot be measured directly and is typically assumed to be interval scaled (Borgatta & Bohrnstedt, 1980). For example, parental support for a current school attendance policy could be conceptualized as an interval-scaled latent variable ranging from strong to weak support, and the ordinal 1 to 4 scale described earlier could serve as a manifest variable that provides information about the latent variable. Similarly, the parental involvement variable of Grolnick et al. (1997) could be conceptualized as an interval-scaled latent variable. According to measurement theory, θ shows no measurement error and is completely reliable (Lord & Novick, 1968, pp. 27–28). In settings in which θ and Y are distinguished, examinees' observed scores Y_i ($i = 1,2,\ldots,N$) are treated as indicators of their true status or proficiency θ_i.

The distinction between manifest and latent variables allows ordinal and interval scales to be further delineated. Suppose that θ represents the latent variable English language arts proficiency and Y the manifest variable of total proficiency score. To relate the two, it is necessary to specify some statistical function linking θ and Y, say, $f(Y|\theta)$, where $Y|\theta$ indicates that the manifest data depend on θ. As Maxwell and Delaney (1985) show, if $f(Y|\theta)$ relates θ and Y in a linear manner, so that unit changes (relative differences) in Y reflect unit changes in θ, then Y is interval scaled. On the other hand, if θ and Y are monotonically related through $f(Y|\theta)$, then unit changes in Y do not reflect unit changes in θ and Y has an ordinal scale. This implies that if θ and Y are both interval scaled, then a monotonic transformation (rescaling) of the manifest data will destroy the link between the two. On the other hand, if θ possesses an interval scale but Y possesses an ordinal scale, a monotonic transformation of the manifest data would not destroy the link between the two because rank ordering would be maintained. If θ is ordinal, Y can never be rescaled to interval data; if θ is interval, however, it may be possible to transform an ordinal Y such that the rescaled scores possess an interval scale.

The distinction between manifest and latent variables also allows us to comment on how our presentation relates to that of Davison and Sharmu (1988, 1990, 1994). These authors consider the case of an ordinal (manifest) Y and an interval-scaled (latent) θ when testing statistical hypotheses associated with procedures such as ANOVA, analysis of covariance, and multiple regression. The Davison and Sharmu framework results in two statistical null hypotheses associated with a statistical test, one for Y and one for θ. These authors describe conditions under which statistical null hypotheses are or are not corrupted when θ is interval scaled but Y is ordinal.

For example, for a one-way, fixed-effects ANOVA with J groups, the underlying linear statistical model is

$$Y_{ij} = \mu_y + \tau_j + \epsilon_{ij}, \tag{1}$$

where Y_{ij} is the observed score of the ith subject in the jth group, μ_y is a grand population mean, τ_j represents a treatment effect defined as $\mu_j - \mu_y$ where μ_j is the mean of the jth population, and ϵ_{ij} represents an error term (Kirk, 1995, p. 32). The assumptions for the associated F test to be valid are that the Y_{ij} in the population are independent, are normally distributed, and show a common variance.

In the Davison and Sharmu approach, it is necessary to specify separate statistical null hypotheses for Y [H_0: $\mu_1(Y) = \mu_2(Y) = \ldots = \mu_J(Y)$] and θ [H_0: $\mu_1(\theta) = \mu_2(\theta) \ldots = \mu_J(\theta)$] and to then decide whether inferences based on an ordinal Y apply to an interval-scaled θ. One problem with the Davison and Sharmu approach is that it involves testing a statistical null hypothesis containing parameters [$\mu_J(\theta)$] based on θ when θ does not appear in the underlying statistical model. This undesirable practice is similar to what Marascuilo and Levin (1970) called a Type IV error. Researchers are prone to Type IV errors when they make inferences about parameters that do not appear in the statistical model assumed to underlie the data. Another problem is that the criteria used by Davison and Sharmu to determine whether inferences for Y apply to θ are not available for many statistical procedures, such as multivariate analysis of variance and HLM. A third problem is that these criteria are rather complex to apply.

Our preference is to rescale the ordinal Y values in the sample to an interval measure prior to statistical estimation and hypothesis testing. (We assume that the population distribution of Y values would also be [hypothetically] rescaled to an interval measure.) This means that, for hypothesis-testing purposes, θ does not exist or $Y \approx \theta$ because Y is assumed to be measured without error. For a one-way ANOVA, this results in a single statistical null hypothesis of equal population means for Y that is consistent with the parameters in Equation 1 [H_0: $\mu_1(Y) = \mu_2(Y) = \ldots = V_J(Y)$]. The presence of a single statistical null hypothesis containing parameters appearing in the statistical model assumed to underlie the data also means that Type IV errors of the kind associated with the Davison and Sharmu approach are avoided.

Another advantage of rescaling is that it is not tied to any particular statistical analysis and may even be useful in statistical procedures in which θ is explicitly modeled (i.e., Y is assumed to be measured with error), such as SEM models with latent variables. Finally, although rescaling procedures have their share of complexities, we believe that they are no more difficult to use than the Davison and Sharmu criteria, and they are probably easier.

Problems With Ordinal Data in Many Statistical Procedures

Many of the statistical procedures favored by educational researchers (e.g., ANOVA, HLM, SEM) have desirable properties if the assumptions underlying those procedures are satisfied. These desirable properties typically include estimates of parameters that are unbiased (i.e., are neither over- or underestimated) and statistical tests that reject a true or false statistical null hypothesis according to what is expected by statistical theory (i.e., Type I error rate and statistical power of a test). We consider two problems associated with the use of ordinal data that can compromise these properties.

Nonnormality

One problem with ordinal data in many statistical procedures is that these values cannot be assumed to be normally distributed, since this requires interval-scaled (continuous) data. Although some writers have argued that equating normal distributions with interval scales is inappropriate (e.g., Stine, 1989), we subscribe to the view of Guilford (1954, p. 17), Gaito (1959), Lord and Novick (1968, p. 22), and others that a normally distributed variable possesses at least an interval scale of measurement. This means that the population distribution of the 5-point Likert-type variable used in Grolnick et al. (1997) would have only five values and could not possibly be normally distributed.

Using ordinal data with statistical procedures requiring normality may produce parameter estimates that are biased and tests that fail to reject a statistical null hypothesis according to what is expected by statistical theory. For example, the use of ordinal data can lead to serious problems when one is estimating and testing variances and/or covariances or correlations. Bollen and Barb (1981) showed that estimates of the Pearson correlation coefficient were biased when computed for ordinal data, which implies that a test of the statistical null hypothesis H_0: $\rho_{xy} = 0$, where ρ_{xy} represents the population Pearson correlation coefficient, will reject H_0 more or less than expected by statistical theory. Martin (1973) and Wylie (1976) reported similar findings. Russell, Pinto, and Bobko (1991) showed that using a 5-point Likert-type variable as the dependent variable in moderated regression produced a substantial loss of statistical power relative to what was expected by statistical theory. Embretson (1996) showed that spurious interaction effects can emerge in factorial ANOVAs using ordinal *Y* data. Studies of the effects of using ordinal data in SEM when the variables are assumed to be interval scaled have also shown that estimation of relationships can be distorted and statistical test results can be misleading (Babakus, Ferguson, & Joreskog, 1987; Boomsa, 1983; Browne, 1984; Cobham & Applegate, 1999; Hu, Bentler, & Kano, 1992; Muthen & Kaplan, 1985).

Other studies have compared the Type I error rate and power of ANOVA-based *F* tests and nonparametric tests for ordinal data. These results generally indicate that *F* tests control Type I error as well as nonparametric tests for (nonnormal) ordinal data (Baker et al., 1966; Hsu, 1968; Nanna & Sawilowsky, 1998; Rasmussen, 1989), but that they often show less statistical power than their nonparametric competitors (Blair & Higgins, 1985; Nanna & Sawilowsky, 1998; Rasmussen, 1989). On the whole, the fact that ordinal data are not continuous and cannot be normally distributed creates problems for many statistical procedures.

Incoherence of the Y *and* θ *Scales*

Another problem is that using parameter estimates based on ordinal data to make inferences about an interval-scaled θ can introduce bias because of scale incoherence. The presence of bias means that a statistic overestimates or underestimates a parameter. For example, the estimation of the mean of the latent parental involvement variable of Grolnick et al. (1997) by their manifest ordinal data introduces more bias than would be present if *Y* and θ shared an interval scale. The implication is that the bias introduced by the incoherence of the *Y* and θ scales threatens the validity of one's interpretations (see, e.g., Bollen & Barb, 1981; Embretson, 1996).

One way to handle scale incoherence is to correct the parameter estimates for the bias induced using an ordinal *Y*. Krieg (1999) derived formulas for means, variances, covariances, correlations, and reliability coefficients that quantify the bias that is introduced when manifest and latent variables are on different scales. These equations can be used to correct the bias in parameter estimates computed for an ordinal *Y*, which in turn imposes coherence on the *Y* and θ scales.

The complexity of the bias computations reduces their attractiveness, however. For example, bias equations for the mean and variance require the calculation of probabilities that depend on the distribution of θ and the function relating *Y* and θ, $f(Y \mid \theta)$. If θ is assumed to be normally distributed, computing the needed probabilities requires numerically evaluating integrals over a normal probability function. Moreover, the bias correction must be applied for each statistic, so data analyses involving more parameters and, hence, more statistics require more computational labor. This computational labor is greater than it may appear at first glance because statistics often of interest in data analyses in educational research, such as partial regression coefficients in multiple regression, are computed from other statistics that would first have to be corrected for bias.

We think that rescaling ordinal data offers a simpler way to impose coherence on the *Y* and θ scales. Rescaling an ordinal *Y* to an interval scale eliminates the bias captured by the equations of Krieg (1999), which in turn should allow for more meaningful inferences.

Despite our enthusiasm for rescaling, it is important to point out that these methods typically introduce error into the rescaled data. This occurs for two reasons. First, rescaling methods rely on statistics computed using sample data that contain sampling error, which is transmitted to the rescaled values. A second potential source of error arises when the assumptions underlying rescaling techniques are not satisfied. These problems speak to the importance of using rescaling methods only when the error introduced by these techniques is minimal. We revisit this topic when IRT models are introduced.

What Are the Options for Handling Ordinal Data?

After reviewing the methodological literature, it is clear that educational researchers faced with analyzing ordinal data have two options. One is to employ statistical methods explicitly designed to analyze such data, which include nonparametric procedures (Conover, 1980; Marascuilo & McSweeney, 1977), contingency table analysis (Agresti, 1990; Clogg & Shihadeh, 1994), regression models for ordinal data (Clogg & Shihadeh, 1994; McCullagh, 1980), and specialized SEM models (Muthen, 1984).

A second option, mentioned earlier, is to rescale ordinal data to an interval scale and then employ standard statistical procedures to analyze the interval data. In our view, rescaling is often the most attractive and practical method for handling ordinal data that can resolve the dilemma of coupling measurement scales with statistical analyses. If *Y* is an ordinally scaled indicator of θ, and θ is interval scaled, then rescaling *Y* to an interval scale may allow the assumption of normality to be met. Of course, normality may not hold even for the rescaled data simply because the measured variable does not follow a normal distribution. In such cases, it is common to attempt to find a nonlinear transformation that produces data that are at least approximately normally distributed or to turn to statistical procedures not requiring normality, such as nonparametric techniques. But a *Y* variable showing an ordinal scale ensures that normality will not be met. Rescaling *Y* will also impose coherence on the *Y* and θ scales that in turn should enhance the validity of inferences.

Among the techniques available to rescale data, we believe that IRT is often the preferred method. Various rescaling techniques are available, such as multidimensional scaling, but we prefer IRT because it addresses many of the concerns of the arguments for and against Stevens's (1951) position. Under some conditions described later, IRT produces interval-scale data, satisfying measurement-based arguments. Yet, it does not explicitly link statistical analyses and the measurement scales of variables, satisfying many statistically based arguments.

The methods of IRT have been available for several years, and some educational research studies have used dependent variables that resulted from an application of IRT (e.g., Lee, Smith, & Croninger, 1997; Rowan et al., 1997). In general, though, the use of IRT to rescale data in educational research appears to be rare. Our presentation focuses on rescaling dependent variables, but arguments for rescaling ordinal independent/predictor variables using IRT follow the same reasoning. Next, we briefly describe the classical test theory (CTT) and IRT measurement models. Following the work of Fischer (1995), Embretson (1996), Maxwell and Delaney (1985), and others, we describe the conditions under which ordinal data can be transformed to an interval scale using IRT methods.

Classical Test Theory

Many variables used in educational research are created by adding together item responses for a test or instrument (e.g., questionnaire, survey) to produce a total or subscale score. This process can be thought of as a way of scaling data whose measurement properties are typically justified by appeals to CTT (Embretson, 1996). CTT refers to traditional methods and techniques for test design and analysis, most of which have a strong correlational flavor, and is distinguished by the dominant role played by a test score and the relatively minor role played by test items and their characteristics (Baker, 1992, p. 1). The CTT measurement model dominated measurement for decades, and understanding its deficiencies provides a powerful argument for considering rescaling ordinal data using IRT.

The deficiencies of CTT include its inability to produce an interval scale for test scores and its failure to take the characteristics of items into account or to provide information about the reliability of estimated scores or proficiencies (Embretson, 1994, 1996; Fischer, 1995; Hambleton & Swaminathan, 1985, pp. 1–14). Summing item responses, questionnaire responses, and so forth to create a total or subscale score for each examinee produces variables that possess an ordinal scale and that fail to reflect key characteristics of items, such as their difficulty. For example, in the Grolnick et al. (1997) study of parent involvement in children's schooling, teachers rated behavioral involvement in the school for various activities using a 1 to 5 Likert-type scale. Total scores were then created by summing responses to the items, giving a possible total score range of 5–25, with larger values interpreted as greater parental involvement. In addition to possessing an ordinal scale, total scores computed in this way weight each item equally and ignore the different characteristics of the items.

Item Response Theory

IRT approaches test design and analysis quite differently from CTT. In IRT, the items and their characteristics play the dominant role. Although there are several well-documented advantages of IRT over CTT (Embretson & Reise, 2000, pp. 13–43; Hambleton & Swaminathan, 1985, pp. 1–14), we focus on the ability of IRT to generate an interval scale for manifest ordinal data that takes into account item characteristics and to provide estimates of the reliability of each examinee's estimated proficiency.

In IRT, the key role of items and their characteristics appears in the statistical model linking examinees' responses to an item and their "true" proficiency on the latent variable, that is, $f(Y \mid \theta)$. The characteristics of items are taken into account in IRT models through the use of item parameters that capture these properties.

Three characteristics of items can be examined. One is item difficulty, which represents the point on the proficiency scale at which the probability of a correct response is one half. Difficulty parameters are typically expressed in standard deviation units and assumed to have a standard-normal distribution with a mean of zero and a standard deviation of one. Thus, items of average difficulty have a difficulty parameter near zero, those that are more difficult have positive difficulty parameter values (e.g., 1.5), and those that are easier have negative difficulty values (e.g.,–1.2).

A second characteristic is item discrimination, which reflects the ability of an item to discriminate among examinees with more or less proficiency. Estimated item discrimination parameters typically range between 0.5 and 2, with larger values indicative of greater discrimination. The third item characteristic that is sometimes examined reflects the extent to which examinees are likely to guess the answer to an item correctly.

Note that IRT is generally used as a measurement model whose focus is examining how items are performing; however, we use IRT as a scaling model whose focus is on transforming observed item or questionnaire responses to an estimated score on a latent variable. It is important to emphasize that either application of IRT involves the use of relatively complex models with rigorous assumptions of their own and that failure to satisfy these assumptions makes using these models inadvisable. In the following, we describe some IRT models and their assumptions.

Steps in Using IRT Models to Rescale Data

To rescale manifest ordinal data using IRT, it is necessary to perform four steps; interval-scaled proficiency estimates appear in Step 3.

Step 1: Identify an Appropriate IRT Model

IRT models can be distinguished in several ways. Thissen and Steinberg (1987) offered a detailed taxonomy of IRT models, but we rely on two categories to characterize IRT models: the nature of the item responses and the kinds of item parameters appearing in the IRT model. The first category consists of (a) nominal item responses, which occur when responses are categorical, including the common case of dichotomously scored (correct, incorrect) items, and (b) graded responses, which involve a rank ordering of responses to an item, such as those obtained using raters to score constructed-response items. IRT models can also be distinguished by the kinds of parameters used to capture the characteristics of an item. In keeping with common practice, we assume that logistic IRT models are used (Embretson & Reise, 2000, pp. 66–72).

The simplest IRT model for responses that have been dichotomously scored is the one-parameter Rasch model. This model uses a single difficulty parameter (β_k, $k = 1,2, \ldots, K$ items) for each item and assumes equal discrimination across items and no guessing. The Rasch model can be written as

$$P_{ik}(Y_{ik} = 1 \mid \theta_i) = \exp[(\theta_i - \beta_k)]/\{1 + \exp[(\theta_i - \beta_k)]\}, \qquad (2)$$

where $P_{ik}(Y_{ik} = 1 \mid \theta_i)$ is the probability of an examinee with proficiency θ_i answering the kth item correctly and exp represents the exponentiation operator (Fischer, 1995). Suppose that θ is normally distributed with a mean of zero and standard deviation of one, that $\beta_1 = -0.5$ (a slightly easy item), and that $\theta_1 = 1$ (above-average proficiency). This arrangement of values indicates that the probability that this examinee will answer this item correctly should be high. Using the Rasch model in Equation 2, the probability is

$$\begin{aligned} P_{11}(Y_{11} = 1 \mid \theta_1 = 1) &= \exp\{[1 - (-0.5)]\}/\{1 + \exp[1 - (-0.5)]\} \\ &= \exp(1.5)/[1 + \exp(1.5)] \\ &= 4.48/(1 + 4.48) \\ &= 0.82. \end{aligned} \tag{3}$$

As expected, it is quite likely that an examinee with this proficiency will answer this item correctly. Suppose that β_1 was still -0.5 but that θ_2 was -1.25, indicating an examinee with below-average proficiency. We would expect the probability of responding correctly to this item for this examinee to be much lower than 0.82:

$$\begin{aligned} P_{21}(Y_{21} = 1 \mid \theta_2 = -1.25) &= \exp\{[-1.25 - (-0.5)]\}/\{1 + \exp[-1.25 - (-0.5)]\} \\ &= \exp(-0.75)[1 + \exp(-0.75)] \\ &= 0.472/1.472 \\ &= 0.32. \end{aligned} \tag{4}$$

As expected, an examinee with $\theta_2 = -1.25$ has a relatively small probability of answering this item correctly. Good sources for additional information about the Rasch model are Andrich (1988) and Embretson and Reise (2000).

Two other IRT models are available for dichotomously scored item responses in addition to the Rasch model. If the items are expected to differ in difficulty but to have similar discriminating power, the Rasch model is preferred; if items are also expected to vary in their discriminating power, then a two-parameter model is appropriate; settings in which guessing needs to be taken into account require a three-parameter model. Choosing among these three models depends on the item parameters needed to describe the responses, although to some extent it is possible to leave the choice of models to the data since one can compare the models statistically to determine which (if any) best captures patterns in the item responses (Thissen, Steinberg, & Gerard, 1986). Good introductions to the two- and three-parameter IRT models for dichotomous responses include Embretson and Reise (2000, pp. 65–79), Hambleton and Swaminathan (1985, pp. 33–48), and Harris (1989). A variety of IRT models are available for nominal responses with more than two categories and for graded-response models (Baker, 1992, pp. 222–287; Embretson & Reise, 2000, pp. 95–124). Table 1 provides information about software for performing the analyses associated with various IRT models.

Step 2: Estimate Item Parameters

Once an IRT model has been selected, it is necessary to fit that model to the data for each item, which entails estimating the item parameters. Various estimation methods are reviewed in Baker (1987, 1992), and Harwell, Stone, Hsu, and Kirisci (1996) review computer programs available to estimate item parameters, includ-

TABLE 1
Software for analyzing data for item response theory models

Computer program	IRT model and nature of data	Chi-square fit test provided	Residuals provided
MULTILOG	1-,2-,3-PAR nominal data 1-,2-PAR graded response	Yes (for the test as a whole only)	Yes (for the test as a whole only)
BILOG	1-,2-,3-PAR dichotomous	Yes	Yes
PARSCALE	1-,2-,3-PAR nominal 1-,2-,3-PAR graded response	Yes	Yes
RASCAL	1-PAR dichotomous	Yes	No
QUEST	1-PAR dichotomous 1-PAR graded response	Yes	No
ASCAL	2-,3-PAR dichotomous	Yes	No

Note. More information about these programs can be obtained from the vendor at http://www.assess.com/softmenu.html. 1-PAR represents a Rasch IRT model with difficulty parameters, 2-PAR a two-parameter IRT model with difficulty and discrimination parameters, and 3-PAR a three-parameter model with difficulty, discrimination, and guessing parameters.

ing some of those in Table 1. Parameter estimation difficulties are, in general, least likely for the one-parameter Rasch model and increasingly likely for more complex IRT models such as the three-parameter model. The use of specialized estimation algorithms, such as those available in the computer program BILOG (Mislevy & Bock, 1997), has eased many of these difficulties, and there is evidence that item parameters can be reliably estimated for various IRT models and varying sample sizes and numbers of items (see, e.g., Drasgow, 1989; Hambleton, Jones, & Rogers, 1993; Harwell & Janosky, 1991; Hulin, Lissak, & Drasgow, 1982; Koch, 1983; Reise & Yu, 1990; Stone, 1992). Still, there are cases in which the use of complex IRT models such as the three-parameter model for dichotomous response or a graded-response model may lead to problems in estimating item parameters. Good sources to consult for parameter estimation in IRT include Baker (1987, 1992), Embretson and Reise (2000), and Hambleton and Swaminathan (1985).

Step 3: Estimate Proficiency Parameters

Next, proficiency parameters are estimated for each examinee (θ_i). Although the term *proficiency* is common in IRT, we remind readers that these models can be used to estimate scores for any interval-scaled latent variable; for example, parental involvement in children's schooling.

The various methods of estimating proficiency parameters that are available in the software listed in Table 1 (e.g., maximum likelihood estimation [MLE], expected a posteriori [EAP], maximum a posteriori [MAP]) tend to give similar results as long as the number of items/questions exceeds 20. For less than 20 items/questions, the EAP method is often recommended (Mislevy & Stocking, 1989), although this method, like the others, produces biased estimates. Good sources for ability estimation methods in IRT include Baker (1992), Embretson and Reise (2000), and Hambleton and Swaminathan (1985).

Fischer (1995) showed that assuming that θ is interval scaled and Y is ordinal produces estimated proficiencies θ_i that possess an interval scale if the Rasch IRT model is used for dichotomously scored data and adequately captures patterns in the data. Thus, using the Rasch model to estimate proficiencies can rescale ordinal data to an interval scale. A variation of the Rasch model in Equation 2 has been widely used in the psychometric ratings literature in which rating data are transformed to an interval scale (Englehard, 1992, 1994, 1996). Proficiency estimates for other IRT models, such as the two- and three-parameter models for dichotomous responses, have not been shown to possess an interval scale. This represents an important gap in the IRT rescaling literature since adequately capturing the response patterns in many tests or questionnaires may require the use of more complex IRT models.

Item response theory also provides evidence of the reliability of estimated proficiencies through the standard error of each θ_i. Suppose that θ_1 is 1 and that the associated standard error is 0.05. These values tell us that this estimated proficiency is well above average and is quite reliable because smaller standard errors indicate proficiency estimates of greater reliability. If $\theta_1 = 1$ but the associated standard error is 0.50, then this estimated proficiency is far less reliable. Alternatively, the so-called test information function can be used to provide evidence of reliability (Baker, 1992, pp. 83–84).

Step 4: Assess Model-Data Fit

Finally, the adequacy with which an IRT model captures patterns in item responses must be evaluated for each item; if these patterns are adequately captured by the IRT model, then estimates of item parameters and estimated proficiencies are credible. In the IRT literature, this is referred to as assessing the adequacy of the model-data fit. Model-data fit in IRT is usually assessed by examining the chi-square goodness-of-fit test produced for each item and by examining the residuals obtained in the process of fitting an IRT model to item response data. Inadequate model-data fit means that estimated item parameters and proficiencies cannot be credibly interpreted and usually leads to modifying and/or deleting some items. Good sources of information about model-data fit in IRT include Embretson and Reise (2000, pp. 233–238) and Hambleton and Swaminathan (1985, pp. 151–169).

In sum, the use of the Rasch model in IRT to produce proficiency estimates (a) rescales ordinal data to an interval scale that may allow the assumption of normality associated with many statistical procedures to be satisfied and (b) can eliminate the bias introduced when using an ordinal Y to make inferences about an interval θ. An IRT approach also incorporates item characteristics into the estimation of θ and provides evidence of the reliability of each examinee's estimated proficiency.

Before continuing, we reemphasize that IRT models are not simple "add-ons" that can be used without regard for their complexity and underlying assumptions. The two key assumptions underlying commonly used IRT models are (a) unidimensionality of proficiency, meaning that a single latent variable accounts for variation common to items, and (b) local independence, which means that, controlling for proficiency, item responses are independent of one another (Embretson & Reise, 2000, pp. 226–227). Failure to satisfy these assumptions makes rescaling data with these models inadvisable. Along the same lines, researchers need to select an IRT model that captures the operating characteristics of items, because failure to do so will probably lead to inadequate model-data fit, resulting in proficiency estimates that cannot be trusted. In short, researchers need to carefully attend to the complexity and assumptions of IRT models before using these models to rescale ordinal data. Good introductions to these assumptions can be found in Embretson and Reise (2000, pp. 226–233), Goldstein (1980), and Hambleton and Swaminathan (1985, pp. 15–31).

Illustrations of the Use of IRT for Rescaling Ordinal Data

We illustrate the use of IRT to rescale data with two examples. Our first example uses the responses of 1,000 fourth-grade examinees to 30 dichotomously scored items. After creating a total-correct score for each examinee, we illustrate the use of the Rasch model to show how these data can be rescaled to an interval scale. In a second example, we show how to rescale graded-response data using IRT.

Example 1: Rescaling Dichotomously Scored Data

The responses of 1,000 fourth-grade examinees to 30 dichotomously scored items for the New Standards exams were used to estimate English language proficiency (θ). Under CTT, we could sum the number correct to produce a total score for each examinee (Y), with a possible range of 0 to 30. Unfortunately, this also produces an ordinal scale of measurement and fails to take the characteristics of the items into account or to provide information about the reliability of the θ_i. To overcome these problems, we use IRT and follow the four steps described earlier. We describe each step in sufficient detail to allow readers to use this methodology for their own data via the computer programs in Table 1.

Step 1

First, we identified an appropriate IRT model. As is common practice, we assumed that each examinee had a true (latent) English language arts proficiency θ_i, that θ followed a normal distribution with a mean of zero and a standard deviation of one, and that the Y_i represented manifest values that can be used to estimate θ_i. Because the responses to the $k = 30$ items were dichotomously scored and we had no reason to believe that the items varied in discrimination or that guessing needed to be modeled, we selected the Rasch model in Equation 2.

Step 2

Next, we estimated the 30 β_k difficulty parameters using the BILOG for Windows (Mislevy & Bock, 1997) program in Table 1 to fit the model in Equation 2 to the data for each of the 30 items and to estimate the proficiency parameters. (A description of the BILOG commands used to perform the analysis is available upon request.) The results of these analyses are summarized in Table 2. For example, for Item 1, the estimated difficulty was –2.03, indicating that this item was relatively easy. The associated standard error of 0.087 indicates that this estimated difficulty is fairly reliable. The fact that almost all of the estimated difficulty parameters are negative indicates that this test was fairly easy for these examinees.

Step 3

In the third step, we used the BILOG computer program to estimate the $\hat{\theta}_i$, which are plotted against the Y_i in Figure 1. Note that the relationship between the Y_i and $\hat{\theta}_i$ is reasonably linear, a pattern often observed with the Rasch model. How-

TABLE 2
IRT results for dichotomously scored data

Item	Estimated β_k	Standard error of β_k	Chi-square test of fit
1	−2.03	0.087	15.1*
2	−0.41	0.056	2.8
3	−1.83	0.084	20.2*
4	−0.16	0.054	4.4
5	0.12	0.055	3.8
6	−0.99	0.063	5.70
7	−1.79	0.084	24.6*
8	−0.42	0.061	27.8*
9	−0.74	0.057	20.6*
10	−1.51	0.072	13.8
11	−0.19	0.054	5.4
12	0.16	0.053	5.6
13	−0.58	0.063	12.3
14	−0.12	0.056	3.5
15	−1.44	0.074	19.9*
16	−1.08	0.062	12.5
17	−1.43	0.071	5.5
18	−0.60	0.061	10.3
19	−0.36	0.061	9.2
20	−0.49	0.061	13.0
21	−0.58	0.062	18.8*
22	−1.76	0.080	4.9
23	−0.31	0.060	14.4*
24	−0.82	0.059	12.9
25	−0.03	0.059	5.1
26	−1.29	0.069	8.5
27	−1.11	0.069	8.8
28	−0.12	0.056	10.3
29	−1.10	0.071	68.6*
30	−1.03	0.069	46.3*

Note. Each chi-square test of model-data fit was performed using $\alpha = .05$.
* $\alpha = .05$.

ever, unlike the Y_i, the $\hat{\theta}_i$ possess an interval scale (assuming adequate model-data fit) and take varying item difficulties into account. For example, the difference between $\hat{\theta}_i$ of 0.80 and 0.30 (0.50 standard deviations) reflects the same difference in English language arts proficiency as the difference between, for example, $\hat{\theta}_i$ of −1.1 and −1.60. The standard errors of the estimated proficiencies are reported in Table 3 for 10 examinees. Examinee 1 had an estimated proficiency of 0.4406, which is above average. But the associated standard error of 0.2414 indicates that this is not a particularly reliable estimate.

Note that the number of unique $\hat{\theta}_i$ values equals the number of unique item response patterns, not the number of possible scores of Y. For example, a dichotomously scored test consisting of two items has $2^2 = 4$ possible response patterns, that is, 00; 01; 10; 11, where 1 is a correct response and 0 an incorrect response. Assum-

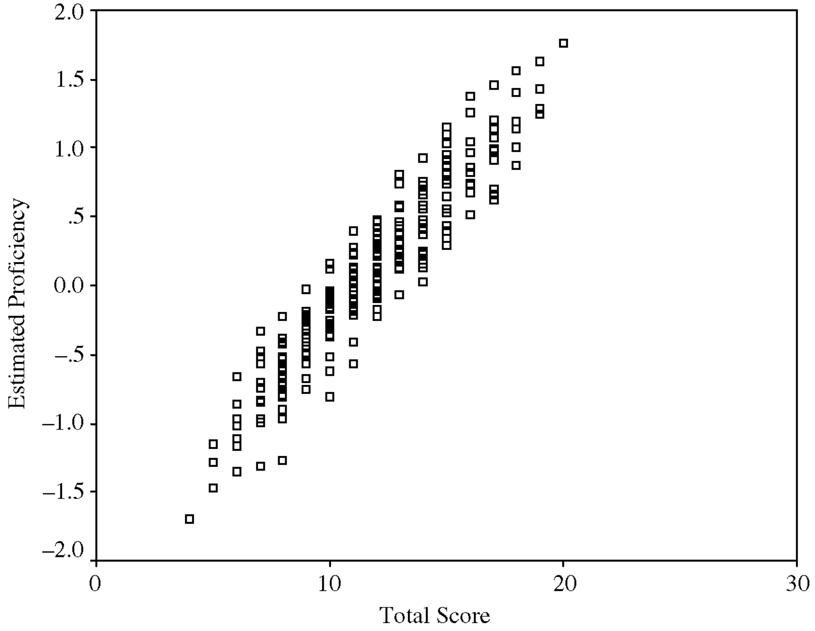

FIGURE 1. *Plot of total scores and estimated proficiencies for the Rasch model.*

ing that all four response patterns were present in the data, rescaling would produce four unique $\hat{\theta}_i$. For the 30-item, dichotomously scored test of English proficiency, there are $2^{30} = 1{,}073{,}741{,}824$ unique $\hat{\theta}_i$ values possible, although in practice a much smaller number of these patterns would be expected to appear in the data (101 were present in the English proficiency data). Because the estimation of proficiency uses item response patterns and the estimated item parameters, it is possible that examinees with the same total-correct score could have different $\hat{\theta}_i$ values.

TABLE 3
Estimated proficiencies using the Rasch model and the standard errors for 10 examinees

Y_i	Estimated proficiency θ_i	Standard error of θ_i
24	0.4406	0.2414
15	−0.5818	0.3230
21	−0.0581	0.4418
17	−0.4538	0.1426
27	0.9434	0.4726
21	−0.0581	0.4418
25	0.5268	0.2928
11	−1.2898	0.2067
6	−1.8240	0.4460
26	0.6848	0.4065

Step 4

In the fourth step, we assessed the fit between the Rasch model and the data for each item. Initially, we assessed model-data fit using chi-square tests output by BILOG. A nonsignificant chi-square fit test at some chosen level of significance α provides evidence of adequate model-data fit, whereas a statistically significant test signals that the item needs special attention and may need to be modified or omitted. Several of the chi-square fit tests reported in Table 2 are significant at $\alpha = 0.05$, suggesting that model-data fit for these items is inadequate.

We also examined the model residuals for each item output by BILOG. These residuals are standardized and are assumed to follow a standard-normal distribution (mean = 0, standard deviation = 1). Residuals are calculated in BILOG by assuming that the interval-scaled θ distribution is represented by discrete values in the ±4 range. Although some information is lost in this process (e.g., all $\hat{\theta}_i$ between 1.47 and 1.89 are grouped together), it greatly simplifies checking of model-data fit. Typically, the bulk of any model-data misfit is in the tails of the θ distribution (high and low proficiencies), because these proficiencies are the most difficult to estimate reliably. Rules of thumb vary, but residuals greater than ±2.5 certainly provide evidence of significant misfit.

As is common practice, we represented the θ distribution in BILOG with 10 values, meaning that for each item there were 10 residuals. For Item 1, the standardized residuals were –6.297, 0.210, –2.075, 0.763, 0.168, 0.335, 0.880, 0.341, and 0.098, suggesting adequate model-data fit with the exception of the first residual, which reflects a poor fit for very low proficiency. For the second item, the residuals were 2.731, 2.419, –2.226, 2.155, –0.655, –0.459, 1.474, –0.130, –0.818, and –0.630, suggesting adequate model-data fit with the exception of the first residual, again associated with low proficiency. This process was repeated for the remaining items.

Combining the information from the chi-square model-data fit tests and the standardized residuals for each item, 10 of the items showed inadequate fit and should be examined for problems in their wording, content, and so forth. It is possible that the poor fit of some items may be attributable to a failure to model item discrimination and/or the likelihood of guessing, and that using more complex IRT models may produce better model-data fit. In many cases, researchers may choose to delete those items showing inadequate model-data fit and base the estimation of proficiency on responses to items showing adequate model-data fit. For illustrative purposes, we assume that model-data fit is adequate for each of the 30 items so that responses to all of the items were used to estimate proficiency.

Once $\hat{\theta}$ is available, it can serve as an interval-scaled dependent variable in many statistical analyses. Transforming an ordinal Y to an interval-scaled θ means that statistical null hypotheses, parameter estimates, and so forth are all expressed on the θ scale. As with any rescaling or data transformation, researchers must be able to specify meaningful statistical null hypotheses based on the rescaled scores and to interpret the statistical results. Otherwise, it is better not to rescale and to employ a statistical procedure that assumes that Y possesses an ordinal scale.

Example 2: Rescaling Graded-Response Data

Recall that each examinee received a grade for each of the English language arts standards clusters that were scored 1 to 5, which represents an ordinal scale. We use the four IRT steps described earlier to rescale these responses in a way that takes item characteristics into account. Psychometric rating models such as those described in Englehard (1992, 1994, 1996) could also be used here.

Step 1

First, we identified an appropriate IRT model. Because of the ordinal nature of the responses, we selected a graded-response model that assumes that the response categories show a rank ordering. A detailed description of graded-response models is beyond the scope of this article, and interested readers are referred to Baker (1992, pp. 222–250), Embretson and Reise (2000, pp. 97–119), and Van der Linden and Hambleton (1996).

One option is to select a Rasch graded-response model (Englehard, 1992, 1994, 1996; Linacre, 1989) in which only difficulty parameters are estimated. We chose Samejima's (1969) graded-response model because it allows us to illustrate the use of a model that takes into account both the difficulty and discrimination of items. Suppose that we have $k = 4$ items, each rated on a 1 to 5 scale. Samejima's model essentially creates four dichotomous response models for each item. In this setup, there are a total of t–1 dichotomous response models (t = total number of response categories) and t–1 difficulty parameters [$\beta_{(t-1)}k$] that must be estimated. There is also a common discrimination parameter for the dichotomous response models. Within each item, this allows us to estimate the difficulty associated with each dichotomous response model as well as the item's discriminating power.

Although the creation of t–1 dichotomous response models for each item allows the $\beta_{(t-1)}$ to be estimated, it complicates their interpretation. The difficulty parameters β_1 and $\beta_{(t-1)}$ reflect the point on the proficiency scale at which the probability that the response will appear in the first or last response category is one half; the remaining difficulty parameters must be averaged in a particular fashion before the usual interpretation of difficulty is appropriate (Baker, 1992, pp. 228–229).

Step 2

Next, the item parameters were estimated. Because BILOG does not handle graded-response models, we estimated the $(t-1)k = 4 \times 4 = 16$ difficulty parameters associated with the four items using the MULTILOG program listed in Table 1 (Thissen, 1991). (A description of the MULTILOG commands used to perform the analysis is available upon request.) The estimated difficulty and discrimination parameters and their standard errors are reported in Table 4. For example, Item 1 shows moderate discrimination (1.38) and is more difficult than the other items. Item 3, on the other hand, has little discriminating power (0.52) and, in practice, would probably be modified or omitted.

Step 3

In the third step, each examinee's proficiency was estimated. The estimated examinee proficiencies and the Y_i are plotted in Figure 2. Note that Figure 2 shows less similarity in the patterns of total-correct responses and estimated proficiencies than was the case for the Rasch model in Figure 1, indicating that the effects of rescaling were more pronounced using the graded-response model. Although $5^4 =$ 625 unique response patterns are possible, 29 appeared in the sample, meaning that rescaling produced 29 unique θ values.

TABLE 4
Item analysis results for graded-response data

Item	Estimated ω	Estimated $\beta_{(t\text{-}1)}$
1	1.38 (.11)	−0.19 (.07)
		0.13 (.06)
		0.48 (.08)
		1.84 (.14)
2	1.05 (.12)	−2.51 (.25)
		0.94 (.12)
		1.09 (.13)
		1.15 (.13)
3	0.52 (.08)	−1.84 (.32)
		0.23 (.16)
		1.98 (.34)
		2.20 (.37)
4	1.15 (.10)	−1.53 (.12)
		−0.92 (.09)
		−0.69 (.09)
		1.01 (.11)

Note. ω represents the discrimination parameter. Values in parentheses are standard errors.

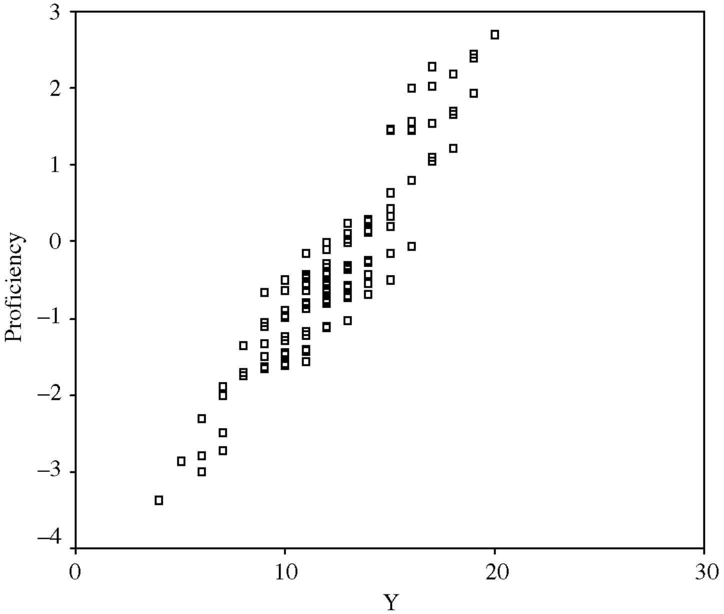

FIGURE 2. *Plot of total scores and estimated proficiencies for Samejima's model*

Step 4

We examined a test of model-data fit for Samejima's model as well as model residuals for the four items combined because MULTILOG does not output results for each item. The overall chi-square test was statistically significant, and an examination of the residuals for the four items suggested that model-data fit might not be adequate (Figure 3) because of the large positive residuals. As before, these large residuals signal a need to examine such factors as item wording and content.

Computer Simulation Study

Unfortunately, the $\hat{\theta}$ for Samejima's graded-response model has not been proven to generate an interval scale. In response to the absence of such information, we performed a small computer simulation study to investigate the extent to which the $\hat{\theta}$ appeared to possess an interval scale under Samejima's model. The advantage of performing a simulation study is that all factors are under the investigator's control; this allowed us to simulate item response data with known characteristics.

We used the GENIRV (Baker, 1986) computer program to simulate graded-response data for 1,000 hypothetical examinees who were assumed to have θ values that followed a normal distribution (mean = 0, variance = 1). We limited the range of θ to ±3 standard deviations and simulated data for a test composed of four graded-response items with five response categories per item. Ordinal Y_i values were produced by summing the four scores for each simulated examinee. We chose

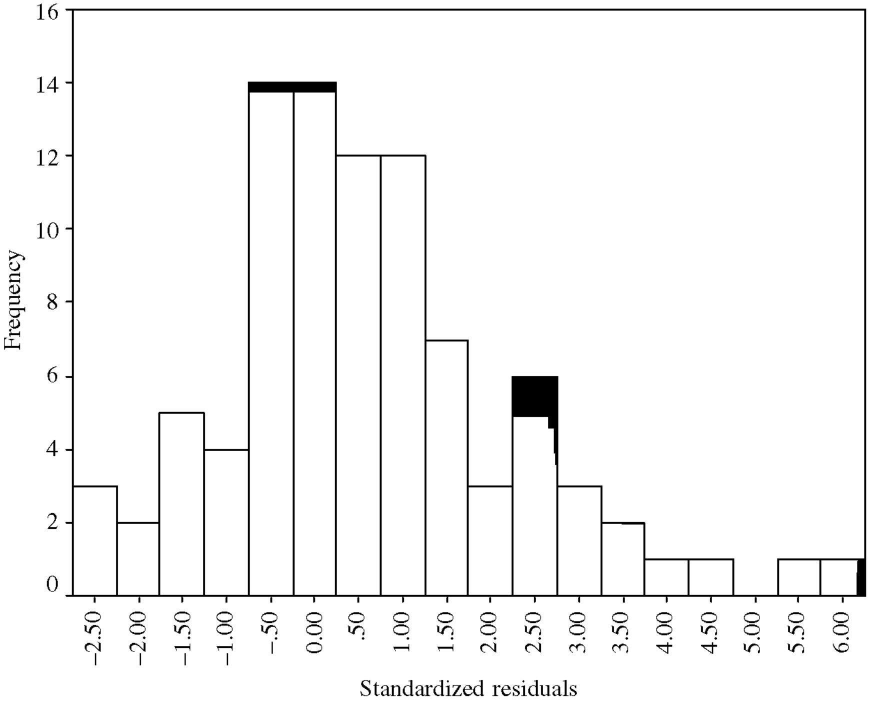

FIGURE 3. *Standardized model residuals for* k = *4 items using Samejima's model*

this configuration because it matched the New Standards English language arts clusters, but in practice it would be wise to use tests with more items since this should produce more reliable proficiency estimates. Our use of simulated data ensured that the latent variable θ had an interval scale, that the true proficiency of each simulated examinee was known (θ_i), and that the manifest Y_i values possessed an ordinal scale.

Once the graded-response data had been simulated by GENIRV, they were submitted to the MULTILOG program to be analyzed under Samejima's model. This produced 16 estimated difficulty parameters, 4 estimated discrimination parameters, and 1,000 estimated proficiencies. We then compared each $\hat{\theta}_i$, estimated using ordinal data, with the associated true proficiency θ_i, which possessed an interval scale of measurement. If the θ_i were similar to the $\hat{\theta}_i$, there would be evidence supporting the use of IRT to rescale ordinal data to an interval scale using Samejima's model.

We began by examining the 1,000 differences ($\hat{\theta}_i - \theta_i$), which appear in Figure 4. Virtually all of the differences were within ±2 standard deviations of the mean of the differences of 0.02 (i.e., were within a range attributable to sampling error). In fact, the largest difference was –2.40, and the next largest was –2.03; however, most of the differences were relatively close to zero. Thus, the differences in Figure 4 provide preliminary evidence that Samejima's IRT model can be used to rescale graded-response data to an interval scale. However, substantial additional work is needed that examines a more comprehensive set of conditions, such as modeling a broader range of θ values.

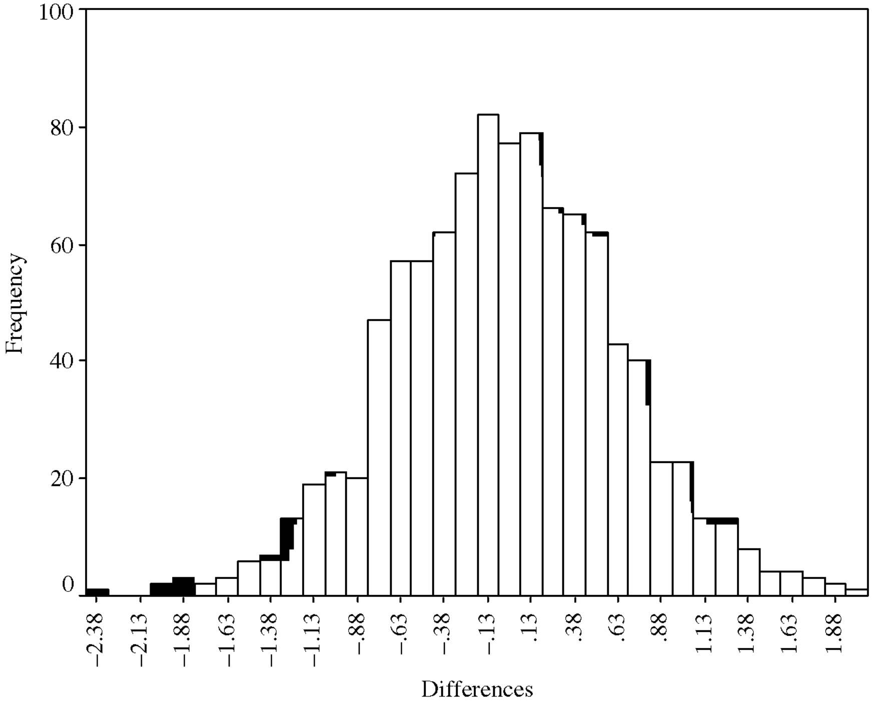

FIGURE 4. *Plot of residuals for the computer simulation study*

Conclusions

Educational research is replete with examples of latent variables that are hypothesized to represent an interval scale but manifest variables that show an ordinal scale. The inability of manifest ordinal variables to follow a normal distribution creates problems when these data serve as dependent variables in many statistical analyses popular in educational research. The bias introduced in using an ordinal dependent variable to make inferences about an interval-scaled latent variable also creates difficulties. Among the options available to educational researchers, rescaling manifest variables measured using an ordinal scale to an interval scale using item response theory is particularly attractive. Under the assumption that the latent variable possesses an interval scale, use of the Rasch IRT model can produce estimated proficiencies that represent an interval scale and that reflect item characteristics. The estimated proficiencies may also follow a normal distribution, an important assumption for many statistical procedures, and will avoid the bias associated with using ordinal data. Information about the reliability of each estimated proficiency can be obtained as well. Put simply, the use of IRT to rescale manifest ordinal data circumvents many of the traditional problems associated with these values.

Still, the use of IRT models to rescale data comes with rigorous assumptions that must be satisfied for the models to be of value. In the event that these assumptions are not satisfied, we advise researchers to consider other options for handling ordinal data, for example, adopting statistical models designed to handle ordinal data (e.g., nonparametric techniques) or employing another rescaling technique such as multidimensional scaling. Unfortunately, these alternatives lack the ability to take item characteristics into account in producing interval-scaled data or to provide information about the reliability of each estimated proficiency.

Clearly, additional work is needed to demonstrate that the estimated proficiencies for a variety of IRT models and item types show an interval scale. One option is to follow Fischer's (1995) approach in which proficiencies under the Rasch model were proved to possess an interval scale. This is the most attractive approach, but such proofs are difficult beyond the case of the Rasch model for dichotomous responses. Alternatively, computer simulation studies could be performed under realistic data conditions to provide evidence of the scale properties of estimated proficiencies, for example, for graded-response models. At a minimum, these studies need to examine factors that would be expected to contribute to scale properties, such as range of proficiency and number and type of items, and to employ a comprehensive statistical strategy for determining whether the estimated proficiencies follow a normal distribution. The availability of computer simulation results of the scale properties of estimated proficiencies will help to clarify the settings in which the use of IRT models to rescale ordinal data to an interval scale is appropriate.

Acknowledgements

This work was supported by the High Performance Learning Communities Project, which is funded by the Office of Educational Research and Improvement of the Department of Education (Grant RC-96-137002). We wish to thank Bruno Zumbo for his suggestion that a computer simulation be performed and Laura D'Amico for her careful reading of the manuscript. We also thank the reviewers for suggestions that were helpful in restructuring the manuscript.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Anderson, N. H. (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin, 58*, 305–316.

Andrich, D. (1988). *Rasch models for measurement.* Beverly Hills, CA: Sage.

Babakus, E., Ferguson, C. E., & Joreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *37*, 72–141.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' prescriptions on statistics. *Educational and Psychological Measurement*, *26*, 291–309.

Baker, F. B. (1986). *GENIRV: A computer program for generating item responses.* Unpublished manuscript, University of Wisconsin, Madison.

Baker, F. B. (1987). Methodology review: Item parameter estimation under the one-, two-, and three-parameter logistic model. *Applied Psychological Measurement*, *11*, 111–141.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Marcel Dekker.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t-test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, *97*, 119–128.

Bollen, K. A., & Barb, K. H. (1981). Pearson's r and coursely categorized measures. *American Sociological Review*, *46*, 232–239.

Boomsa, A. (1983). *On the robustness of LISREL (maximum likelihood) against sample size and non-normality*. Amsterdam: Sociometric Research Foundation.

Borgatta, E. F., & Bohrnstedt, G. W. (1980). Level of measurement: Once over again. *Sociological Methods and Research*, *9*, 147–160.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.

Clogg, C. C., & Shihadeh, E. S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.

Cobham, I., & Applegate, B. (1999, April). *The effects of subject-to-variable ratio, measurement scale, and number of factors on the stability of the factor model.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Conover, W. J. (1980). *Practical nonparametric statistics* (2nd ed.). New York: Wiley.

Davison, M. L., & Sharma, A. R. (1988). Parametric statistics and levels of measurement. *Psychological Bulletin, 104*, 137–144.

Davison, M. L., & Sharma, A. R. (1990). Parametric statistics and levels of measurement: Factorial designs and multiple regression. *Psychological Bulletin*, *107*, 394–400.

Davison, M. L., & Sharma, A. R. (1994). ANOVA and ANCOVA of pre- and post-test ordinal data. *Psychometrika*, *59*, 593–600.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*, 77–90.

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss, (Eds.), *Modern theories of measurement: Problems and issues*. Ottawa, Ontario, Canada: University of Ottawa, Faculty of Education.

Embretson, S. E. (1996). Item response theory and spurious interaction effects in factorial ANOVA designs. *Applied Psychological Measurement*, *20*, 201–212.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Englehard, G. E. (1992). The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, *5*, 171–191.

Englehard, G. E. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, *33*, 93–112.

Englehard, G. E. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement*, *33*, 56–70.

Fischer, G. H. (1995). Derivations of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Gaito, J. (1959). Non-parametric methods in psychological research. *Psychological Reports*, *5*, 115–125.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564–567.

Gautam, S., Kimeldorf, G., & Sampson, A. R. (1996). Optimized scorings for ordinal data for the general linear model. *Statistics & Probability Letters*, *27*, 231–239.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, *33*, 234–246.

Gregoire, T. G., & Driver, B. L. (1987). Analysis of ordinal data to detect population differences. *Psychological Bulletin*, *101*, 159–165.

Grolnick, W. S., Benjet, C., Kurowski, C. O., & Apostoleris, N. H. (1997). Predictors of parent involvement in children's schooling. *Journal of Educational Psychology, 89*, 538–548.

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

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society*, *Series A*, *159*, 445–492.

Harris, D. (1989). Comparison of 1-, 2-, and 3-parameter IRT models. *Educational Measurement: Issues and Practice*, *8*, 35–41.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior distribution variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*, 279–291.

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125.

Hsu, T. C. (1968). *An empirical investigation of the effect of the length of the score scale on the significance of the F-test*. Unpublished doctoral dissertation, University of Iowa, Iowa City.

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*, 351–362.

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

Kirk, R. E. (1995). *Experimental design* (3rd ed.). New York: Wiley.

Koch, W. R. (1983). Likert scaling using the graded-response model. *Applied Psychological Measurement*, *7*, 15–32.

Krieg, E. F. (1999). Biases induced by course measurement scales. *Educational and Psychological Measurement*, *59*, 749–766.

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review*, *35*, 515–524.

Lee, V. E., Smith, J. B., & Croninger, R. G. (1997). How high school organization influences the equitable distribution of learning in mathematics and science. *Sociology of Education*, *70*, 128–150.

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

Lord, F. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750–751.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luce, R. D., Krantz, D. H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement* (Vol. 3). New York: Academic Press.

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance: The elimination of Type IV errors. *American Educational Research Journal*, *7*, 392–421.

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole.

Martin, W. S. (1973). The effects of scaling on the correlation coefficient: A validity study. *Journal of Marketing Research*, *10*, 316–318.

Maxwell, S. E., & Delaney, H. (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin*, *97*, 85–93.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society*, *Series B*, *42*, 109–142.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin*, *100*, 398–407.

Mislevy, R. J., & Bock, R. D. (1997). *PC-BILOG: Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Stocking, M. L. (1989). A consumer's guide to LOGIST and BILOG. *Applied Psychological Measurement*, *13*, 57–75.

Muthen, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115–132.

Muthen, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*, 171–189.

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert data in disability and medical rehabilitation research. *Psychological Methods*, *3*, 55–67.

Narens, L., & Luce, R. D. (1986). Measurement: The theory of numerical assignments. *Psychological Bulletin*, *99*, 166–180.

New Standards. (1997). *New Standards Reference Examination technical summary*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Rasmussen, J. L. (1989). Analysis of Likert-scale data: A reinterpretation of Gregoire and Driver. *Psychological Bulletin*, *105*, 167–170.

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

Resnick, L., & Harwell, M. (1998). *High performance learning communities: District 2 achievement*. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Rowan, B., Chiang, F., & Miller, R. J. (1997). Using research on employees' performance to study the effects of teachers on students' achievement, *Sociology of Education*, *70*, 256–284.

Russell, C. J., Pinto, J. K., & Bobko, P. (1991). Appropriate moderated regression and inappropriate research strategy: A demonstration of information loss due to scale coarseness. *Applied Psychological Measurement*, *15*, 257–266.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs*, No. 17.

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw-Hill.

Stevens, S. S. (1946). On the theory of measurement. *Science*, *103*, 677–680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: Wiley.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science*, *161*, 849–856.

Stevens, S. S. (1971). Issues in psychophysical measurement. *Psychological Review*, *78*, 426–450.

Stine, W. W. (1989). Meaningful inference: The role of measurement in statistics. *Psychological Bulletin*, *105*, 147–155.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*, 1–16.

Thissen, D. (1991). *MULTILOG Version 6 user's guide.* Mooresville, IN: Scientific Software.

Thissen, D., & Steinberg, L. (1987). Taxonomy of item response models. *Psychometrika*, *51*, 567–577.

Thissen, D., Steinberg, L., & Gerard, M. (1986). Beyond group mean differences: The concept of item bias. *Psychological Bulletin*, *99*, 118–128.

Townsend, J. T., & Ashby, F. G. (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, *96*, 394–401.

Tukey, J. W. (1961). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In L. V. Jones (Ed.), *The collected works of John W. Tukey* (Vol. 3, pp. 391–484). Belmont, CA: Wadsworth.

Van der Linden, W. J., & Hambleton, R. (1996). *Handbook of modern item response theory*. New York: Springer.

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician, 47*, 65–72.

Wylie, P. B. (1976). Effects of coarse grouping and skewed marginal distributions on the Pearson product moment correlation coefficient. *Educational and Psychological Measurement*, *36*, 1–7.

Zumbo, B. D., & Zimmerman, D. W. (1993). Is the selection of statistical methods governed by level of measurement? *Canadian Psychology*, *34*, 390–400.

Zumbo, B. D., & Zimmerman, D. W. (2000). Levels (scales) of measurement and the relation between parametric and nonparametric statistical tests. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 6). Greenwich, CT: JAI Press.

Authors

MICHAEL R. HARWELL is Professor of Research Methodology, Department of Educational Psychology, 323 Burton Hall, 178 Pillsbury Drive, SE, University of Minnesota, Minneapolis, MN 55455; harwe001@umn.edu. Specializations include the behavior of parametric and nonparametric statistical tests under assumption violations, applications of meta-analysis in methodological research, and optimal methods for handling missing data.

GUIDO G. GATTI is Senior Research Specialist, Center for Biomedical Informatics, Suite 8084 Forbes Tower, 200 Lothrop Street, University of Pittsburgh Medical Center, Pittsburgh, PA 15213-2582. Specializations include the application of structural equation modeling to research in the social sciences and the effects of assumption violations on frequently used statistical models.