GAIN SCORE IN ITEM RESPONSE THEORY AS AN EFFECT SIZE MEASURE

WEN-CHUNG WANG
National Chung Cheng University

WU CHYI-IN
Academia Sinica

Because of the requirement of reporting effect sizes and in the interest of measurement of change within the item response theory framework, their combination becomes a new issue. In the present study, repeated measures are decomposed as an initial ability and one or more modifiabilities (gain score) using a multidimensional Rasch model. The modifiability can be directly interpreted in terms of logit scale. The standardized mean modifiability is recommended for meta-analysis when test equating is not possible across studies. A simulation study was conducted to assess parameter recovery. It appeared that the point estimates were accurate whereas the error variances were underestimated. The bootstrap method was used and found appropriate for estimating the error variances. Implications and applications are illustrated through an empirical example.

Keywords: *Rasch model; multidimensional item response model; modifiability; change measurement; item response theory*

In recent years, reporting effect size has been increasingly recognized as a necessary practice. The editorial policies of 20 journals in education and psychology formally require effect size reporting (Capraro & Capraro, 2002). The number of such journals is certainly increasing. The fourth edition of the American Psychological Association (APA; 1994) *Publication Manual* notes, "You are encouraged to provide effect-size information" (p. 18). The

Correspondence concerning this article should be addressed to Wen-Chung Wang, Department of Psychology, National Chung Cheng University, Chia-Yi, Taiwan; e-mail: psywcw@ccu.edu.tw.

Educational and Psychological Measurement, Vol. 64 No. 5, October 2004 758-780
DOI: 10.1177/0013164404264118

APA Task Force on Statistical Inference cautions, "Always provide some effect size estimate when reporting a *p* value" (Wilkinson & the APA Task Force on Statistical Inference, 1999, p. 599). The fifth edition of the APA *Publication Manual* emphasizes, "For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship" (APA, 2001, p. 25). Many effect size measures have been developed over the decades. The variety of effect size measures can be classified as two broad categories: measures of effect size according to group mean differences and measures of association strength according to the proportion of variance accounted for (Maxwell & Delaney, 1990).

Although there are many studies on effect size measures, most discuss the measures within the framework of classical test theory (CTT; Lord & Novick, 1968) rather than item response theory (IRT). CTT is known to be problematic in many theoretical and practical testing aspects, such as mutual dependence for the estimation of item and person parameters (Lord, 1980). In addition, raw scores or their linear transformations do not necessarily yield an interval scale. Nowadays, IRT is widely used in educational and psychological tests, such as the Armed Services Vocational Aptitude Battery, the Scholastic Assessment Test, the Graduate Record Examinations, the Differential Ability Scales, the Woodcock-Johnson Psycho-Educational Battery, the Multidimensional Personality Questionnaire, the Beck Depression Inventory, and the Rosenberg Self-Esteem Scale (Embretson & Reise, 2000). Given the requirement of reporting effect sizes and the popularity of IRT, reporting IRT-based effect sizes becomes a new direction to explore. In particular, when tests or inventories are used to measure outcomes in experimental or observational studies and item response models are used to analyze the data, how can effect size measures be reported?

DeMars (2001) compared the estimates of Cohen's (1969, p. 18) standardized mean difference based on simulated groups under the Rasch (1960) and three-parameter logistic models (Birnbaum, 1968). The standardized mean difference between groups was computed using maximum likelihood estimates or expected a posteriori (EAP) estimates for individual persons. The population standardized mean difference was substantially underestimated, especially when tests were short. The underestimation, not explained by the author, is in fact due to measurement error in the estimates. The shorter the test, the larger the measurement error. The attenuation on effect sizes due to measurement error should be corrected for a better approximation of true effect sizes. To disattenuate, Wang and Chen (2004) proposed a procedure that takes measurement error into consideration in the model and directly estimates the mean difference between groups and the common variance of the two groups. They also proposed a disattenuation procedure based on the IRT test reliability to obtain the sampling variance of the standardized mean difference. Through simulations, it is found that the point estimate of the

standardized mean difference is practically unbiased, and its sampling variance can be accurately obtained.

Measurement error in scores and its potential influence on effect size measures have not received sufficient attention in research practice (Baugh, 2002; Henson, 2001; Thompson & Snyder, 1998; Thompson & Vacha-Haase, 2000; Vacha-Haase, Ness, Nilsson, & Reetz, 1999). In reality, response measures are always imperfect and contain measurement error that results in attenuation in effect size measures. Only when measurement error is absent or trivial (e.g., height and weight measures) can the usual effect size measures be applied safely; otherwise, disattenuation is more appropriate. Therefore, it is very reasonable to disattenuate the effect size measure when the measurement error is not negligible.

DeMars (2001) and Wang and Chen (2004) investigated the standardized mean difference between groups. In many areas of education and psychology, measuring change is of great interest. Often, a gain score is computed as the simple difference between two successive test scores, such as a pretest score and a posttest score. Bereiter (1963) noted three fundamental psychometric problems in measuring change: (a) paradoxical reliability, such that the higher the correlation between the pretest and posttest, the lower the reliability of gain scores; (b) scale incompatibility, such that change is not measured on the same scale for persons at different initial score levels; and (c) spurious relationship, such that gain scores have a spurious negative relationship to initial scores.

Within the CTT framework, observed scores X and Y are assumed to contain both a true score and an error. The difference between X and Y reflects two parts: (a) difference in true scores and (b) difference due to measurement error. The reliability of the gain score is represented as

$$\rho_{DD'} = \frac{\sigma_X^2 \rho_{XX'} + \sigma_Y^2 + \rho_{YY'} - 2\rho_{XX}\sigma_X\sigma_Y}{\sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y}, \quad (1)$$

where $\rho_{DD'}$ is the reliability of the gain score D (= $Y - X$), $\rho_{XX'}$ is the reliability of X (pretest score), $\rho_{YY'}$ is the reliability of Y, ρ_{XY} is the correlation between X and Y, σ_X^2 is the variance of X, and σ_Y^2 is the variance of Y. If X and Y are highly correlated, the true score part of X must overlap considerably with the true score part of Y. As a consequence, there will be hardly any difference between the true scores of these two variables. Differences between the scores on X and scores on Y thus will be due almost entirely to measurement error. Accordingly, the more highly correlated X and Y, the less reliable their difference will be.

This paradoxical reliability has implications for both research and testing applications. For instance, a comparison of pretest and posttest scores might be used to determine the progress made in a remedial program. The pretest

and posttest are usually comparable because it would not make sense to use completely different types of tests when assessing a person's progress. Therefore, scores on the two tests are likely to be positively correlated whereas gain scores show low reliability. The measure of individual differences in the amount progressed (posttest minus pretest) could be quite unreliable. Persons who appear to have progressed quite a bit could show little apparent gain in an equivalent remedial program. This does not reflect instability in their proficiency to progress; rather, it is a reflection of the potential unreliability of gain scores.

Lord (1956) and Cronbach and Furby (1970) warned educational and psychological researchers about potential problems in the analysis of unreliable gain scores and suggested abandoning them if at all possible. In fact, gain scores are not necessarily unreliable. They can be reliable, and their reliability coefficients are intermediate between those of the pretest and posttest in a large proportion of practical testing applications, even though the paradoxical reliability still exists (Rogosa & Willett, 1983; Williams & Zimmerman, 1996, 1999; Zimmerman & Williams, 1982, 1998). Several researchers have suggested a variety of statistical techniques for studying and measuring change (Collins & Horn, 1991; Cribbie & Jamieson, 2000; Dugard & Todman, 1995; Edwards, 1993, 1995; Geenen & van de Vijver, 1993; Hake, 1998; Jamieson, 1994; Malgady & Colon-Malgady, 1991; Overall & Tonidandel, 2002); nevertheless, strictly speaking, these problems seem irresolvable because the change measurements are based on CTT, in which the estimation of item and person parameters is mutually confounded. In addition, classical methods of gain scores suffer from certain fundamental disadvantages. For example, all gain scores are considered as equally precise indicators of true change. It is, however, the test scores on different time points of different persons and their corresponding gain scores that can have different degrees of precision, depending on the location of the persons on the latent dimension relative to the location of the items. Moreover, both the measurement of change and the assessment of its precision as defined within the CTT context refer to the "manifest scale," the interval property of which is postulated but not empirically or theoretically substantiated. A compression of the scale is bound to occur near the boundaries of the score domain (Fischer, 2003).

Raw score measures are ordinal rather than interval. Because of the ordinal nature of raw scores, the difficulty of pretest can bias the amount of gain (in raw score unit) observed in groups that differ in initial achievement (May & Nicewander, 1998). To compute gain scores by subtracting a pretest score from a posttest score, interval scale measurement is required. The Rasch scale (Rasch, 1960) has been recognized as possessing this property of interval measurement (Andrich, 1988; Bond & Fox, 2001; Embretson & Reise, 2000; Fischer, 1995; Perline, Wright, & Wainer, 1977; Scheiblechner, 1999;

Wright & Stone, 1979) and applied to change measurement. For example, Fischer and Pazer (1991) and Fischer and Ponocny (1994) extended the rating scale model (Andrich, 1978) and partial credit model (Masters, 1982) into the linear rating scale model and linear partial credit model, respectively, and demonstrated their applications in the measurement of change. Fischer (2003) provided confidence intervals for gain scores on the latent dimension under the partial credit model (Masters, 1982). Andersen (1985) developed a Rasch measurement model for longitudinal latent structure between repeated testings, which combines the values of the latent dimension at several occasions into a multidimensional latent density and directly estimates the variance-covariance matrix among the values. Likewise, Embretson (1991) presented a multidimensional Rasch model for measuring learning and change. A simplex structure was postulated to link item responses to an initial ability and one or more modifiabilities (learning abilities). This model, unlike Andersen's, decomposes the effective ability involved in the latter occasion into an initial ability and one or more modifiabilities. The modifiability represents individuals' gain across occasions.

Embretson (1991, 1993) demonstrated how the multidimensional Rasch model resolves Bereiter's (1963) fundamental problems by conceptualizing change as a latent dimension. In particular, the paradoxical reliability results from conceptualizing the multiple measurements as influenced by only one dimension, which not only obscures the inherent multidimensional nature of the change concept but also is unable to accommodate situations in which changes in performance result from qualitative changes in psychological processes, as noted by Cronbach and Furby (1970). On the contrary, the multidimensional Rasch model allows changes in processes to be represented as separate dimensions (i.e., initial ability and modifiabilities). For the problem of scale incompatibility, the interval scale measurement cannot be easily justified for raw scores and their linear transformations. In contrast, interval-level scaling can be justified directly by the IRT measurement models, particularly the family of Rasch models. Thus, change measured on the Rasch scale has a constant meaning for performance when measured from different initial levels, whereas it is not for raw scores. With respect to the problem of spurious relationship, a negative correlation between initial status and change can be expected for raw scores, when the scale of raw scores is compressed near the boundaries of the score domain and when a test has too little ceiling and floor to observe gains at high ability levels and loses to low abilities. Change would be underestimated at these extremes, which would create a negative bias in the correlation of initial status and change. The negative bias can be partially removed in item response models, because the IRT scale is not compressed near the boundaries of the score domain. When an adaptive testing procedure is used, floor and ceiling effects can be further eliminated, given that the item pool contains sufficient items of various difficulties.

Even though IRT permits a reconceptualization of these fundamental problems in gain scores, the above studies of IRT modeling have some drawbacks. First, the linear rating scale model and linear partial credit model are not suitable for measuring individual differences in change because all individuals are assumed to change by the same amount across occasions. Second, although gain scores of individual persons are applicable in Fischer's (2003) study, the item parameters have to be known in advance, rather than jointly calibrated from the whole data of item responses over time. Third, even though individual differences in change and item parameters can be jointly calibrated using Andersen's and Embretson's models, these models are limited to dichotomous items. To resolve these problems, Wang, Wilson, and Adams (1998) applied the multidimensional random coefficients multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) to the measurement of change with polytomous items. However, no specific implications about the modifiability within the context of effect size measures were provided, nor were the standard errors reported for estimating the confidence interval of the mean modifiability.

The present study formulates the modifiability within the context of IRT effect size measures. The standardized mean modifiability is recommended for meta-analysis when test equating is not possible across studies. The MRCMLM is used to estimate the initial ability and modifiability. A simulation study was conducted to assess whether the parameters can be adequately recovered. The bootstrap was used to approximate the sampling distributions of the standardized and unstandardized mean modifiabilities. Finally, an empirical example is given.

The MRCMLM

The MRCMLM, which is a multidimensional extension of the random coefficients multinomial logit model (Adams & Wilson, 1996), can be expressed as

$$\log\left(\frac{p(X_{ij}=1}{p(X_{i(j-1)}=1}\right)=(\mathbf{b}_{ij}^{'}-\mathbf{b}_{i(j-1)}^{'})\theta+(\mathbf{a}_{ij}^{'}-\mathbf{a}_{i(j-1)}^{'})\xi, \tag{2}$$

where $p(X_{ij} = 1; \xi|\theta)$ and $p(X_{i(j\ -\ 1)} = 1\xi|\theta;)$ denote the probabilities of responses to item i that are in categories j and $j - 1$, respectively, conditioned on ability vector θ; $\mathbf{b}_{ij}^{'}$ and $\mathbf{b}_{i(j-1)}^{'}$ denote the scoring vectors given to categories j and $j - 1$ in item i, respectively; and $\mathbf{a}_{ij}^{'}$ and $\mathbf{a}_{i(j-1)}^{'}$ denote the design vectors given to categories j and $j - 1$ in item i, respectively, to express the relationship among the elements in the item parameter vector $\boldsymbol{\xi}= (\xi_1, \ldots, \xi_p)$. Note that the item scoring vector $\mathbf{b}_{ij}$ in the MRCMLM is not a set of parameters but is known a priori. The MRCMLM belongs to the family of Rasch

measurement models, so that interpretation of the item parameters is simpler than for models in which discrimination parameters are present.

The rating scale model (Andrich, 1978), being commonly used to analyze Likert-type items, can be expressed as

$$\log\left(\frac{p(X_{ij}=1)}{p(X_{i(j-1)}=1)}\right) \equiv \text{log it} = \theta - (\delta_i + \tau_j), \tag{3}$$

where δ_i is the overall difficulty of item i and τ_j is the threshold difficulty of category j across items. Consider that Likert-type items are administered at K occasions, and the rating scale modeling is used. To implement Embretson's modeling of the initial ability and modifiability, the modeling at Occasion 1 is

$$\text{log it} = \theta_1 - (\delta_i + \tau_j), \tag{4}$$

where θ_1 denotes the ability at Occasion 1 (the initial ability). At Occasion 2, the modeling is

$$\text{log it} = \theta_1 + \theta_2 - (\delta_i + \tau_j), \tag{5}$$

where θ_2 denotes the modifiability at Occasion 2 (i.e., the change in ability from Occasions 1 to 2). At Occasion k, the modeling is

$$\text{log it} = \theta_1 + \ldots + \theta_{k-1} + \theta_k - (\delta_i + \tau_j), \tag{6}$$

where θ_k denotes the modifiability at occasion k (the change in ability from occasions $k-1$ to k). Equations 4 to 6 as a whole can be expressed in terms of Equation 2 by specifying appropriate scoring vectors and design vectors. Note that the item parameters δ_i and τ_j in Equations 4 to 6 remain unchanged across occasions, so that the abilities across occasions are automatically set on the same scale. Even when some of the items are different across occasions, the abilities across occasions can be put on the same scale, as long as the scales are equated through common items.

Construct invariance over time is the prerequisite of change measurement. If the test construct changes over time, the test would become useless, and any IRT or classical study on its reliability, validity, or gain scores would be rendered meaningless because the results would not be generalized to other time points and the meaning of gain scores is vague. To detect whether the construct changes over time within the IRT context, one may compare the item parameter estimates that are calibrated from different time points. If substantial variations in the parameter estimates over time are found, the construct does not hold constant over time (also called item parameter drift; Bock, Muraki, & Pfeiffenberger, 1988). Standard detection methods of dif-

ferential item functioning (Holland & Waner, 1993) can be carried out to detect item parameter drift over time.

The modifiability $\theta_k (k > 1)$ can be directly interpreted as an effect size: gain score from occasions $k-1$ to k on the logit scale. One point of gain on the logit scale indicates that the log odds of any item in the test at occasion k are 1 point higher than those at occasion $k-1$, or equivalent; the odds at occasion k are $e^1 (= 2.72^1)$ times of those at occasion $k-1$. The mean modifiability across persons can be interpreted as the mean gain on the logit scale. The mean modifiabilities across studies can also be compared directly and interpreted in the same way, once test equating is made across studies. The comparison of the mean modifiabilities across studies becomes complicated when different variances of modifiability exist across studies even after test equating. As an example, consider the following situation in which the mean modifiabilities are 1 logit for both studies after test equating but the variance of the modifiability in the first study is 1 and that in the second study is 2. Likewise, the variances of the modifiabilities in a study may be very different over occasions. This brings about the same issue why the mean difference has to be standardized (Cohen, 1969) to obtain a scale-free measure for comparison across studies, even though the mean modifiabilities across studies can still be interpreted in terms of log odds when test equating is made. In such a case, the standardized mean modifiability provides additional information about the effect sizes. Besides, when completely different items are used across studies and test equating is not possible, the standardized mean modifiability is the only choice for meta-analysis.

The standardized mean modifiability across persons is defined as

$$d_{\theta_k} = \frac{\mu_{\theta_k}}{\sigma_{\theta_k}}, \tag{7}$$

where μ_{θ_k} and σ_{θ_k} are the mean and standard deviation of θ_k, respectively. This standardized mean modifiability, like the standardized mean difference, can be interpreted as the percentage of overlap between the sampling distributions under the null hypothesis (H_0, no gain) and the alternative hypothesis (H_1), as depicted in Figure 1. The percentages of overlap are .85, .67, and .53 when the values of the standardized mean modifiability are .2, .5, and .8, respectively (Cohen, 1988). When the unstandardized mean modifiability is used, the interpretation should focus on the log odds of test items. If the standardized mean modifiability is used, the interpretation should focus on the percentage of overlap. In practice, reporting both kinds of modifiability indices is desirable.

The accompanying computer program ACER ConQuest (M. Wu, Adams, & Wilson, 1998) of the MRCMLM provides marginal maximum likelihood estimation (Bock & Leiberman, 1970) with an expectation/maximization

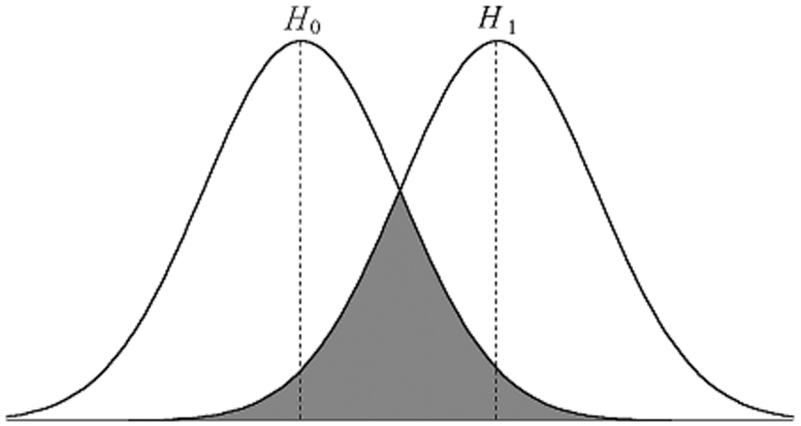

Figure 1. Overlap between the sampling distributions under H_0 and H_1.

algorithm (Bock & Aitkin, 1981) in which the person abilities are assumed to be a representative sample from a distribution, usually a multivariate normal distribution. As only the population parameters of the person ability distribution (i.e., mean vector and variance-covariance matrix for a multivariate normal distribution) are estimated, the individual person is usually given an estimate of the expected value of the marginal posterior distribution (EAP estimate) conditional on the person's responses (Bock & Mislevy, 1982). After obtaining the direct estimates of the mean initial ability, modifiabilities, and variance-covariance matrix, one can compute the estimate of the standardized mean modifiabilities as

$$\hat{d}_{\theta_k} = \frac{\hat{\mu}_{\theta_k}}{\hat{\sigma}_{\theta_k}}, \tag{8}$$

where $\hat{\mu}_{\theta_k}$ and $\hat{\sigma}_{\theta_k}$ are the maximum likelihood estimates.

Maximum likelihood estimates have several optimal properties: (a) consistency, convergence to the true value with increasing sample size; (b) efficiency, the relatively smallest standard error; and (c) asymptotically normal distribution of estimation error (Eliason, 1993, p. 18; Embretson & Reise, 2000, p. 166). Based on these properties, $\hat{\mu}_{\theta_k}$ and $\hat{\sigma}_{\theta_k}$ are efficient estimators of $\hat{\mu}_{\theta_k}$ and $\hat{\sigma}_{\theta_k}$, respectively. Therefore, $\hat{d}_{\theta_k}$ is a good estimator of d_{θ_k}. Moreover, because both $\hat{\mu}_{\theta_k}$ and $\hat{\sigma}_{\theta_k}$ are maximum likelihood estimators, and they are practically uncorrelated, $\hat{d}_{\theta_k}$ is expected to be asymptotically normally distributed with an expected value of d_{θ_k}. However, it is difficult to derive the theoretical sampling variance of $\hat{d}_{\theta_k}$.

The current version of ACER ConQuest does not compute error variances with the full information matrix of item parameters and person distributional parameters for the multidimensional forms of the model. Instead, it approximates error variances by ignoring the covariances in the parameter estimates. In doing so, the obtained error variances are usually underestimated. To resolve this problem, the bootstrap (Efron, 1979) is recommended.

The Simulation

Design

The design and the generating values were adopted from the following empirical example, in which 1,080 students completed a hostility scale four times, once every year (C.-I. Wu, 1999). The scale contained six 5-point Likert-type items. On the fourth administration, only three of the six items were given. The rating scale modeling together with Embretson's procedure was used. One mean initial ability and three mean modifiabilities were estimated for the four occasions. In addition, a 4 • 4 variance-covariance matrix for the four kinds of abilities was estimated. The item parameters contained five overall difficulties (the overall difficulty of the last item was constrained to be the negative sum of the five overall difficulties to make the mean overall difficulties zero) and three threshold difficulties (the fourth threshold difficulty is the negative sum of the three threshold difficulties). Altogether, 22 parameters were estimated, including 8 item parameters, 4 parameters for the mean vector, and 10 parameters for the 4 • 4 variance-covariance matrix. Five hundred replications were made.

Analysis

ACER ConQuest was used to calibrate parameters. The bias value, empirical sampling variance, and mean square error of the estimates across the 500 replications were computed as

$$Bias_{\hat{\varsigma}} = \frac{1}{500}\sum_{k=1}^{500}(\hat{\varsigma}_k - \varsigma), \tag{9}$$

$$S^2_{\hat{\varsigma}} = \frac{1}{499}\sum_{k=1}^{500}(\hat{\varsigma}_k - \bar{\hat{\varsigma}}_k), \tag{10}$$

$$MSE_{\hat{\varsigma}} = \frac{1}{500}\sum_{k=1}^{500}(\hat{\varsigma}_k - \varsigma)^2, \tag{11}$$

Table 1
Descriptive Statistics for the Simulation Study

Parameter	Generating	Bias	S^2	*MSE*
$\xi_1 \equiv \delta_1$	1.815	0.0077	0.0007	0.0007
$\xi_2 \equiv \delta_2$	0.746	0.0026	0.0005	0.0005
$\xi_3 \equiv \delta_3$	0.788	0.0055	0.0008	0.0008
$\xi_4 \equiv \delta_4$	0.693	0.0005	0.0008	0.0008
$\xi_5 \equiv \delta_5$	0.031	0.0003	0.0004	0.0004
$\xi_6 \equiv \tau_1$	1.477	0.0075	0.0009	0.0009
$\xi_7 \equiv \tau_2$	0.098	0.0021	0.0008	0.0008
$\xi_8 \equiv \tau_3$	0.502	0.0039	0.0015	0.0016
μ_{θ_1}	2.040	0.0044	0.0010	0.0010
μ_{θ_2}	0.304	0.0006	0.0011	0.0011
μ_{θ_3}	0.027	0.0024	0.0017	0.0017
μ_{θ_4}	1.312	0.0042	0.0069	0.0069
σ_{11}	0.380	0.0016	0.0007	0.0007
σ_{12}	0.304	0.0036	0.0009	0.0009
σ_{13}	0.044	0.0001	0.0010	0.0010
σ_{14}	0.108	0.0003	0.0026	0.0026
σ_{22}	0.909	0.0213	0.0041	0.0046
σ_{23}	0.383	0.0117	0.0029	0.0031
σ_{24}	0.448	0.0132	0.0064	0.0066
σ_{33}	0.965	0.0238	0.0050	0.0056
σ_{34}	0.547	0.0201	0.0092	0.0096
σ_{44}	2.579	0.0683	0.0477	0.0524
d_{θ_2}	0.3344	0.0112	0.0026	0.0027
d_{θ_3}	0.0280	0.0016	0.0020	0.0020
d_{θ_4}	0.5087	0.0190	0.0028	0.0032

respectively, where $\hat{\varsigma}$ denotes a particular estimator, ς denotes the generating value, and $\bar{\hat{\varsigma}}_k$ denotes the mean estimate across replications.

Results

Table 1 shows the generating value, bias value, empirical sampling variance, and mean square error for the 22 estimators. To test jointly whether the estimates for the 22 parameters were biased, the Hotelling T^2 test was used. The transformed F statistic was 7.175, with degrees of freedom 22 and 478 and a p value smaller than .001, indicating that the estimates were biased. The bias values were in the range of 0.068 and 0.020. This range was very small and negligible compared to the range of the generating values, 2.040 to 2.579.

The last three rows of Table 1 list the descriptive statistics for the three standardized mean modifiabilities. For these three estimators, the trans-

Table 2
Sampling Variances Obtained From the Bootstrap Samples and Simulated Replications and Their Ratios

Parameter	S^2 (Bootstrap)	S^2 (Simulation)	Ratio
μ_{θ_2}	0.0012	0.0011	1.09
μ_{θ_3}	0.0017	0.0017	1.00
μ_{θ_4}	0.0063	0.0069	0.91
d_{θ_2}	0.0026	0.0026	1.00
d_{θ_3}	0.0020	0.0020	1.00
d_{θ_4}	0.0025	0.0028	0.89

formed F statistic was 27.03, with degrees of freedom 3 and 497 and a p value smaller than .001. Therefore, the three estimates were also biased. The bias values were 0.011, 0.002, and 0.019 for the three generating values 0.334, 0.028, and 0.509, respectively. Likewise, the bias was not serious. The Shapiro-Wilk W test for normality (Shapiro & Wilk, 1965) was used to test whether the estimates of the three standardized mean modifiabilities were normally distributed. The W test can be viewed as being based approximately on the correlation coefficient between the ordered values and their expected values under normality. The closer W is to unity, the more plausible normality will be. The W statistics for the three standardized mean modifiabilities were .978 (p = .027), .988 (p = .829), and .974 (p = .001), respectively. As the W statistics were very close to unity, the nonnormality for the third mean modifiability was not serious. In brief, ACER ConQuest yielded very accurate point estimates for the model parameters as well as the standardized mean modifiabilities. The estimates of the standardized mean modifiabilities were approximately normally distributed.

The Bootstrap

The bootstrap, which has been used to approximate unknown sampling distribution, was used to obtain the sampling distributions of the (standardized) mean modifiabilities. Five hundred bootstrap samples were randomly resampled with replacement from the real data set (discussed later in the empirical example). These 500 bootstrap samples were then calibrated using ACER ConQuest. Table 2 shows the empirical sampling variances of the three mean modifiabilities and the three standardized mean modifiabilities obtained from the bootstrap samples and the above simulated replications. The ratio of these two kinds of empirical sampling variances was computed. The empirical sampling distribution obtained from the simulated replications can be treated as the best approximation to the theoretical sampling distribution. If the bootstrap was accurate, the ratio should be very close to unity. It was found that the ratios for the six kinds of mean modifiabilities were be-

tween 0.89 and 1.09, with a mean of 0.98. Accordingly, the bootstrap yielded very good approximations for the error variances.

Figure 2 presents the histograms of the estimates for the three unstandardized mean modifiabilities. Each mean modifiability has two histograms, one for the bootstrap samples and the other for the simulated replications. Likewise, Figure 3 presents those for the three standardized mean modifiabilities. All 12 empirical sampling distributions appeared to be normal. Given the practically unbiased point estimates from ACER ConQuest and the good approximations for the error variances from the bootstrap, confidence intervals for these parameters can be drawn accurately.

An Empirical Example

A hostility scale with six 5-point Likert-type items was administered to 1,080 seventh graders four times, once per year. At the fourth occasion, only three of the six items were administered. Table 3 shows the raw score means and standard deviations for the four occasions, the gain scores from Occasions 1 to 2 and from Occasions 2 to 3, and the respective standardized mean gains. The mean scores were increased from Occasions 1 to 2, then to Occasion 3. The mean gain scores from Occasions 1 to 2 and from Occasions 2 to 3 were 1.63 and 0.24, respectively. The respective standardized mean gains (mean gain/*SD*) were 0.41 and 0.06, respectively. Thus, the mean hostility levels were increased from the seventh to eighth grade quite a bit and then to ninth grade almost unnoticeably. The mean score at Occasion 4 was much smaller than those at the previous occasions, mainly because only half of the items were administered at Occasion 4. Because different numbers of items were administered at Occasions 3 and 4, a direct subtraction of the pretest score from the posttest score was inappropriate. Therefore, it was difficult to define the gain from Occasions 3 to 4 (from the ninth to tenth grade) using raw scores. Strictly speaking, when different items are used across occasions, the direct subtraction in raw scores is problematic.

For item response analysis, Andrichs rating scale modeling together with Embretsons procedure was applied, in which each item was modeled with one overall difficulty and three threshold difficulties. The mean overall difficulty of the six items was constrained to zero, which means that the overall difficulty of the last item was the negative sum of the five overall difficulties. As a result, 8 item parameters, including five overall difficulties and three threshold difficulties, were estimated. Because the persons were assumed to be a representative sample of a multivariate normal distribution, only the mean vector and variance-covariance matrix of the multivariate normal distribution were empirically estimated (i.e., empirical Bayes; Lee, 1997, p. 214). There were 4 parameters in the mean vector, including one mean initial ability and three mean modifiabilities. Likewise, there were 10 parame-

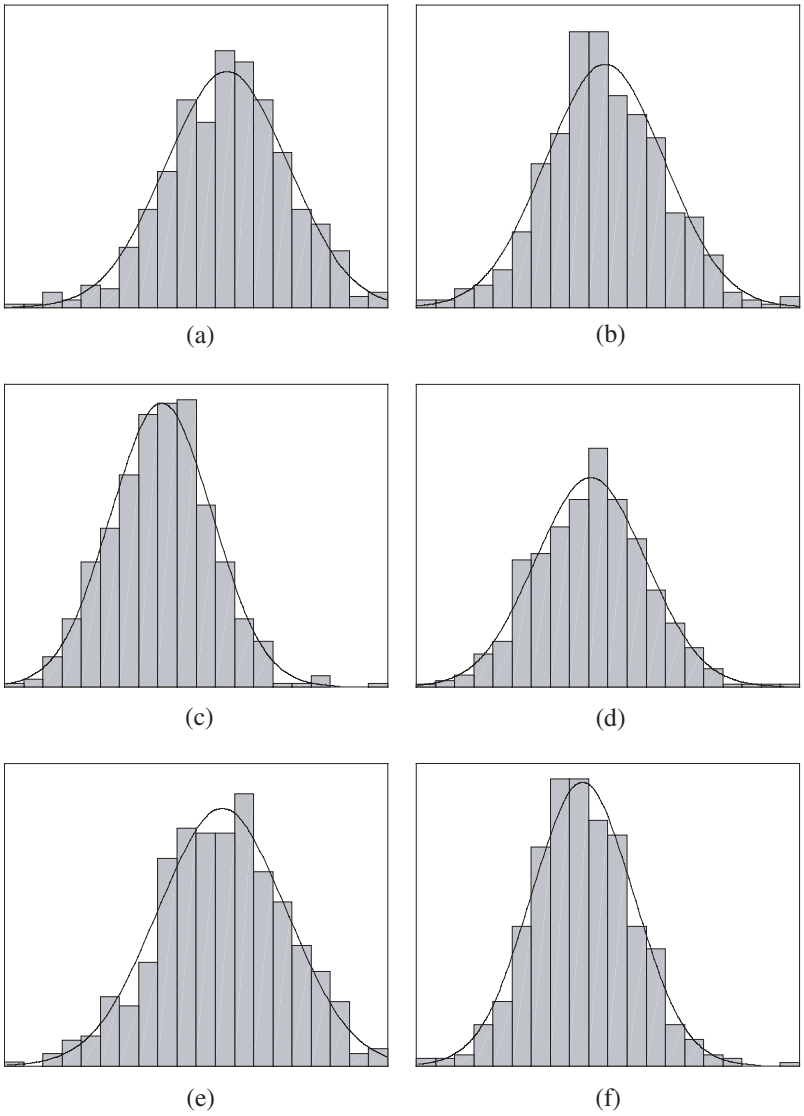

Figure 2. Histograms of the estimates for the three mean modifiabilities obtained from the bootstrap samples and simulated replications. (a) First modifiability from the bootstrap samples ($S^2 = 0.0012$). (b) First modifiability from the simulated replications ($S^2 = 0.0011$). (c) Second modifiability from the bootstrap samples ($S^2 = 0.0017$). (d) Second modifiability from the simulated replications ($S^2 = 0.0017$). (e) Third modifiability from the bootstrap samples ($S^2 = 0.0063$). (f) Third modifiability from the simulated replications ($S^2 = 0.0067$).

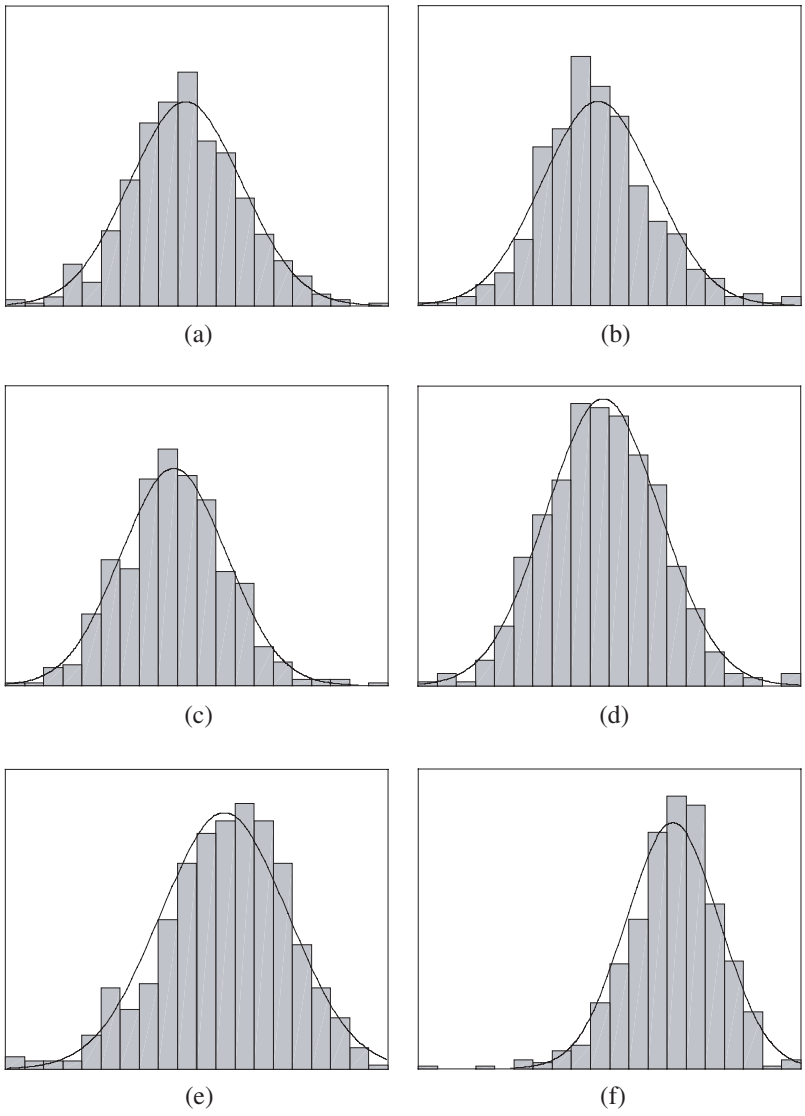

Figure 3. Histograms of the estimates for the three standardized mean modifiabilities obtained from the bootstrap samples and simulated replications. (a) First modifiability from the bootstrap samples ($S^2 = 0.0026$). (b) First modifiability from the simulated replications ($S^2 = 0.0026$). (c) Second modifiability from the bootstrap samples ($S^2 = 0.0020$). (d) Second modifiability from the simulated replications ($S^2 = 0.0020$). (e) Third modifiability from the bootstrap samples ($S^2 = 0.0025$). (f) Third modifiability from the simulated replications ($S^2 = 0.0028$).

Table 3
Descriptive Statistics of the Raw Scores for the Four Occasions, Gain Scores, and Standardized Mean Gains

Occasion	*M*	*SD*	Standardized Mean Gain
1	9.90	2.29	
2	11.53	4.58	
3	11.77	4.83	
4	3.79	1.41	
Gain ($1 \rightarrow 2$)	1.63	3.94	0.41
Gain ($2 \rightarrow 3$)	0.24	3.95	0.06

ters in the 4×4 variance-covariance matrix, including four variances and six covariances. Altogether, 22 parameters were estimated.

Table 4 lists the parameter estimates (the same as the generating values in Table 1), standard errors obtained from the ACER ConQuest printout, standard errors from the bootstrap samples, standard errors from the simulated replications, weighted mean square errors, and *Z* statistics for testing whether the item fitted the models expectation. The weighted mean square errors (M. Wu et al., 1998) for the eight item parameters were between 1.04 and 1.23, with a mean of 1.16. The *Z* statistics were between 1.04 and 4.95, with a mean of 3.44, indicating that the items did not fit the expectation of the model very well (note that the sample size of 1,080 was very large). However, as the weighted mean square errors were not far apart from unity, the item fit was still acceptable, although not ideal. The standard errors obtained from the bootstrap samples and simulated replications were very similar. Both were larger than those obtained from the ACER ConQuest printout, which ignored the covariances in the parameter estimates. In brief, the standard errors obtained from the bootstrap samples were larger and more accurate than those from the ACER ConQuest printout.

The estimate of the mean initial ability was 2.040, indicating that the average hostility level was very low, compared to the zero mean item difficulties. The first mean modifiability (i.e., mean gain from the 7th to 8th grade) was 0.304, indicating that the average hostility increased slightly. The second mean modifiability (i.e., mean gain from the 8th to 9th grade) was 0.027, indicating that the average hostility remained practically unchanged. The third mean modifiability (i.e., mean gain form the 9th to 10th grade) was 1.312, indicating that the average hostility reduced substantially. This substantial reduction might be because the students, just entering senior schools, were released from very competitive entrance examinations and attracted to the novelty of school environments and new friendship. The log odds of any items at Occasion 4 were 1.92 lower than those at Occasion 3, which were 0.0227 higher than those at Occasion 2, which in turn were 0.304 higher than those at Occasion 1.

Table 4
Parameter Estimates, Standard Errors, and Item Fit Statistics

Parameter	Estimate	SE (Printout)	SE (Bootstrap)	SE (Simulation)	WMSE	*Z*
δ_1	1.815	0.02	0.027	0.026	1.23	4.95
δ_2	0.746	0.02	0.023	0.022	1.22	4.84
δ_3	0.788	0.02	0.028	0.028	1.19	4.26
δ_4	0.693	0.02	0.030	0.028	1.10	2.34
δ_5	0.031	0.02	0.022	0.021	1.13	2.97
τ_1	1.477	0.02	0.030	0.029	1.14	3.02
τ_2	0.098	0.02	0.027	0.027	1.19	4.09
τ_3	0.502	0.03	0.036	0.039	1.04	1.04
μ_{θ_1}	2.040	0.018	0.029	0.031		
μ_{θ_2}	0.304	0.029	0.035	0.034		
μ_{θ_3}	0.027	0.030	0.041	0.041		
μ_{θ_4}	1.312	0.049	0.079	0.083		
σ_{11}	0.380	0.015	0.026	0.026		
σ_{12}	0.304	0.024	0.028	0.030		
σ_{13}	0.044	0.025	0.033	0.031		
σ_{14}	0.108	0.041	0.053	0.051		
σ_{22}	0.909	0.038	0.074	0.064		
σ_{23}	0.383	0.040	0.055	0.054		
σ_{24}	0.448	0.066	0.075	0.080		
σ_{33}	0.965	0.040	0.067	0.071		
σ_{34}	0.547	0.067	0.086	0.096		
σ_{44}	2.579	0.113	0.215	0.218		
d_{θ_2}	0.334	NA	0.051	0.051		
d_{θ_3}	0.028	NA	0.045	0.045		
d_{θ_3}	0.509	NA	0.050	0.053		

Note. WMSE = weighted mean square error; NA = not available.

After standardization, the estimates for the first, second, and third standardized mean modifiabilities were 0.334, 0.028, and 0.509, with standard errors (obtained from the bootstrap samples) of 0.051, 0.045, and 0.050, respectively. The 95% confidence intervals for the three standardized mean modifiabilities were (0.234, 0.434), (0.060, 0.116), and (0.607, 0.411), respectively. According to Cohen's (1988) criterion, the effect sizes were small for the first two modifiabilities and medium for the last modifiability.

The first and second standardized mean gains in raw scores were 0.41 and 0.06, respectively, as shown in Table 3. They were close to the first and second standardized mean modifiabilities of 0.334 and 0.028, respectively. The similarity is due to the scale-free nature of the standardized mean gain or modifiability. Note that it is not possible to obtain the third standardized mean gain using raw scores because only half of the items were administered at Occasion 4.

Table 5
Correlations Between the Initial Ability and Modifiabilities and the Item Response Theory (IRT) Reliabilities

	Initial	First Modifiability	Second Modifiability	IRT Reliability
Initial			.76	
First modifiability	.52			.71
Second modifiability	−.07	−.41		.55
Third modifiability	−.11	−.29	−.35	.50

Table 5 shows the correlations between the initial ability and three modifiabilities. They were between .41 and .52, indicating that the correlations were small to moderate. Note that the modifiabilities are gain scores. Therefore, the correlation between the first and second modifiabilities (.41) was the correlation between two gain scores, not the correlation between the initial status and its gain. The sum of the initial ability θ_1 and the first modifiability θ_2 is the initial ability for the second modifiability θ_3. The correlation between the second modifiability θ_3 and its initial ability $\theta_1 + \theta_2$ can be computed as

$$\rho_{(\theta_1+\theta_2)} = \frac{\mathrm{cov}(\theta_1+\theta_2+\theta_3)}{\sqrt{\mathrm{var}(\theta_1+\theta_2)\times\mathrm{var}(\theta_3)}} = \frac{\mathrm{cov}(\theta_1,\theta_3)+\mathrm{cov}(\theta_2,\theta_3)}{\sqrt{[\mathrm{var}(\theta_1)+\mathrm{var}(\theta_2)+2\,\mathrm{cov}(\theta_1,\theta_2)]\times\mathrm{var}(\theta_3)}}. \quad (12)$$

Likewise, the correlation between the third modifiability θ_4 and its initial ability $\theta 1 + \theta_2 + \theta_3$ is

$$\rho_{(\theta_1+\theta_2+\theta_3),\theta_4} = \frac{\mathrm{cov}(\theta_1+\theta_2+\theta_3+\theta_4)}{\sqrt{\mathrm{var}(\theta_1+\theta_2+\theta_3)\times\mathrm{var}(\theta_4)}} = \frac{\mathrm{cov}(\theta_1,\theta_4)+\mathrm{cov}(\theta_2,\theta_4)+\mathrm{cov}(\theta_3,\theta_4)}{\sqrt{[\mathrm{var}(\theta_1)+\mathrm{var}(\theta_2)+\mathrm{var}(\theta_3)+2\,\mathrm{cov}(\theta_1,\theta_2)+2\,\mathrm{cov}(\theta_1,\theta_3)+2\,\mathrm{cov}(\theta_2,\theta_3)]\times\mathrm{var}(\theta_4)}}. \quad (13)$$

According to the estimated variances and covariances in Table 4, $r_{(\theta_1+\theta_2)\theta_3}$ was .25 and $r_{(\theta_1+\theta_2+\theta_3)\theta_4}$ was .46.

Within the IRT framework, the measurement error is no longer homogeneous. If a single quantity is required to denote test reliability as a whole, an average reliability across persons can be used, which can be computed as

$$r_{\theta\theta} = \frac{\hat{\sigma}^2_{\mathrm{EAP}}}{\hat{\sigma}^2_{\theta}}, \quad (14)$$

where $\hat{\sigma}^2_{EAP}$ is the sample variance of the EAP estimates and $\hat{\sigma}^2_{\theta}$ is the estimated variance of the latent trait when marginal maximum likelihood estimation is implemented (Mislevey, Beaton, Kaplan, & Sheehan, 1992). This IRT reliability can be viewed as the counterpart of the classical test reliability. As shown in Table 5, the IRT reliability for the initial ability was .76, which was satisfactory for six 5-point Likert-type items. The IRT reliability for the third modifiability was .50, which was not low, compared to only three items administered at the fourth occasion and the low reliability nature of gain scores.

Conclusion

Because of the requirement of reporting effect sizes and the interest of change measurement within the IRT framework, a combination of these two becomes a new issue. The most serious fundamental problems with gain scores within the CTT framework are paradoxical reliability, scale incompatibility, and spurious relationship. Recent developments in IRT have made it applicable to resolve these problems. For gain scores to be adequate, interval scale measurement is required. The IRT scale, particularly the Rasch scale, has been recognized to have the property of interval measurement. In the present study, an initial ability and a set of modifiabilities are used to describe change over time. Item parameters and the initial ability and modifiabilities are jointly estimated using ACER ConQuest. A simulation study was conducted to investigate whether ACER ConQuest yields accurate estimates. It appeared that the point estimates are very accurate, although not unbiased. However, the error variances are underestimated. To correct the underestimation, the bootstrap was used and found to yield very good approximations. In practice, 25 to 200 bootstrap samples are often enough for estimating a standard error (Efron & Tibshirani, 1993, p. 52).

With the Rasch measurement, mean modifiability can be directly interpreted as the gain in the log odds of any item in the test. Moreover, if the logit scales across studies are equated, mean modifiabilities across studies can also be directly interpreted in terms of the log odds. However, when completely different items are used in different studies and test equating is impossible, mean modifiabilities across studies are not directly comparable. In such a case, the standardized mean modifiability, interpreted as the percentage of overlap between the sampling distributions under H_0 (no gain) and H_1, is more useful for comparison of effect sizes across studies. For both direct interpretation and meta-analysis, reporting both kinds of modifiability indices is recommended.

In addition to ACER ConQuest, the SAS NLMIXED procedure (SAS Institute, 1999) is an alternative for fitting many common nonlinear and generalized linear mixed models, including the MRCMLM model. The reader is

referred to Wolfinger and SAS Institute (n.d.) for details of the NLMIXED procedure. According to the authors experiences in applying the multidimensional approach, the NLMIXED procedure may take several days to converge (or sometimes may fail to converge) whereas ACER ConQuest takes only a few minutes.

The idea of standardized mean modifiability can be generalized to multiparameter item response models (e.g., Birnbaum, 1968; Samejima, 1969). To the authors knowledge, no multidimensional multiparameter item response models or commercial computer programs (e.g., Bguin & Glas, 2001; Bock, Gibbons, & Muraki, 1988; Fraser, 1988; McDonald, 1982; McKinley & Reckase, 1983; Reckase, 1985; Wilson, Wood, & Gibbons, 1991) are available for modeling modifiability. Future studies may be conducted to develop such models and corresponding computer programs. There are many other types of effect size measures, such as ρ^2, η^2, ω^2, the multivariate Roys Θ, and Pillai-Bartlett *V*, that are not considered in this study. Future studies may aim at generalizing these measures into an IRT context.

References

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch models as a mixed coefficients multinomial logit. In G. Engelhard & M. R. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 143-166). Norwood, NJ: Ablex.

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

American Psychological Association. (1994). *Publication manual of the American Psychological Association* (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3-16.

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

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Baugh, F. (2002). Correcting effect sizes for score reliability: A reminder that measurement and substantive issues are linked inextricably. *Educational and Psychological Measurement, 62*, 254-263.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541-562.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in measuring change*. Madison: University of Wisconsin Press.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Bock, R. D., Gibbons, R., & Muraki, E. J. (1988). Full information item factor analysis. *Applied Psychological Measurement*, *12*, 261-280.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 179-197.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*, 275-285.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Capraro, R. M., & Capraro, M. M. (2002). Treatments of effect sizes and statistical significance tests in textbooks. *Educational and Psychological Measurement*, *62*, 771-782.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Collins, L. M., & Horn, J. L. (1991). *Best methods for the analysis of change*. Washington, DC: American Psychological Association.

Cribbie, R. A., & Jamieson, J. (2000). Structural equation models and the regression bias for measuring correlates of change. *Educational and Psychological Measurement*, *60*, 893-907.

Cronbach, L. J., & Furby, L. (1970). How we should measure change or should we? *Psychological Bulletin*, *74*, 68-80.

DeMars, C. (2001). Group difference based on IRT scores: Does the model matter? *Educational and Psychological Measurement*, *61*, 60-70.

Dugard, P., & Todman, J. (1995). Analysis of pre-test post-test control group designs in educational research. *Educational Psychology*, *15*, 181-198.

Edwards, J. R. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, *36*, 1577-1613.

Edwards, J. R. (1995). Alternatives to difference scores as dependent variables in the study of congruence in organizational research. *Organizational Behavior and Human Decision Processes*, *64*, 307-324.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Eliason, S. R. (1993). *Maximum likelihood estimation: Logic and practice*. Newbury Park, CA: Sage.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495-515.

Embretson, S. E. (1993). Psychometric models for learning and cognitive processes. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Lawrence Erlbaum.

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

Fischer, G. H. (1995). Derivation of the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 15-38). New York: Springer-Verlag.

Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement, 27*, 3-26.

Fischer, G. H., & Pazer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, *56*, 637-651.

Fischer, G. H., & Ponocny, I. (1994). An extension of the partial credit model with an application to the measurement of change. *Psychometrika*, *59*, 177-192.

Fraser, C. (1988). NOHARM: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Armidale, New South Wales, Australia: University of New England, Centre for Behavioral Studies.

Geenen, R., & van de Vijver, F. J. R. (1993). A simple test of the law of initial values. *Psychophysiology*, *30*, 525-530.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*, 64-74.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Jamieson, J. (1994). Measurement of change and the law of initial values: A computer simulation study. *Educational and Psychological Measurement*, *55*, 38-46.

Lee, P. M. (1997). *Bayesian statistics: An introduction* (2nd ed.). New York: John Wiley.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421-437.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

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

Malgady, R. G., & Colon-Malgady, G. (1991). Comparing the reliability of difference scores and residuals in analysis of covariance. *Educational and Psychological Measurement*, *51*, 803-807.

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

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective*. Belmont, CA: Wadsworth.

May, K., & Nicewander, W. A. (1998). Measuring change conventionally and adaptively. *Educational and Psychological Measurement*, *58*, 882-897.

McDonald, R. P. (1982). Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement*, *6*, 379-396.

McKinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation*, *15*, 389-390.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133-161.

Overall, J. E., & Tonidandel, S. (2002). Measuring change in controlled longitudinal studies. *British Journal of Mathematical and Statistical Psychology*, *55*, 109-124.

Perline, R., Wright, B. D., & Wainer, H. (1977). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-255.

Rasch, G. (1960). *Probabilistic models for some intelligent and attainment tests*. Copenhagen, Denmark: Institute of Educational Research.

Reckase, M. D. (1985). The difficulty of test items that measure more than one dimension. *Applied Psychological Measurement*, *9*, 401-412.

Rogosa, D., & Willett, J. B. (1983). Demonstrating the reliability of the difference score. *Journal of Educational Measurement*, *20*, 335-343.

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

SAS Institute. (1999). SAS OnlineDoc (Version 8) [Computer software manual on CD-ROM]. Cary, NC: Author.

Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika*, *64*, 295-316.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611.

Thompson, B., & Snyder, P. A. (1998). Statistical significance and reliability analyses in recent JCD research articles. *Journal of Counseling and Development, 76*, 431-436.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195.

Vacha-Haase, T., Ness, C., Nilsson, J., & Reetz, D. (1999). Practices regarding reporting of reliability coefficients: A review of three journals. *Journal of Experimental Education, 67*, 335-341.

Wang, W.-C., & Chen, H.-C. (2004). The standardized mean difference within the framework of item response theory. *Educational and Psychological Measurement, 64,* 201-223.

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1998). Measuring individual differences in change with Rasch models. *Journal of Outcome Measurement*, *2*, 240-265.

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594-604.

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59-69.

Williams, R. H., & Zimmerman, D. W. (1999). Nonindependence of parameters of the validity and reliability of gain scores. *Perceptual & Motor Skills*, *88*, 679-682.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Mooresville, IN: Scientific Software.

Wolfinger, R. D., & SAS Institute. (n.d.). *Fitting nonlinear mixed models with the new NLMIXED procedure*. Retrieved August 17, 2003, from http://support.sas.com/rnd/app/papers/nlmixedsugi.pdf.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wu, C.-I. (1999). *The etiology of adolescents substance abuse: A social learning model* (Tech. Rep. No. DOH88-HR-621). Taipei, Taiwan: Academia Sinica.

Wu, M., Adams, R. J., & Wilson, M. R. (1998). ACER ConQuest: Generalised item response modeling software [Computer software]. Camberwell, Victoria: Australian Council for Educational Research.

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149-154.

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology*, *51*, 343-351.