Moderated Multiple Regression, Spurious Interaction Effects, and IRT

Sun-Mee Kang, California State University, Northridge
Niels G. Waller, Vanderbilt University

Two Monte Carlo studies were conducted to explore the Type I error rates in moderated multiple regression (MMR) of observed scores and estimated latent trait scores from a two-parameter logistic item response theory (IRT) model. The results of both studies showed that MMR Type I error rates were substantially higher than the nominal alpha levels when scale scores were composed of summed binary item responses (e.g., true/false, yes/no, disagree/agree items). Performing the regression analyses on estimated trait scores ($\hat{\theta}$) from a two-parameter logistic model improved the error detection rates considerably. That is, the Type I error rates for spurious interaction effects were similar to the nominal alpha levels under most conditions. These findings suggest that IRT provides a viable means of controlling an important source of spurious interactions in data sets that are well characterized by IRT models. *Index terms: moderated multiple regression, item response theory, spurious interaction, Type I error.*

In recent years, considerable attention has focused on the detection of nonspurious interactions in moderated multiple regression (MMR) (e.g., Aiken & West, 1991; Jaccard, Turrisi, & Wan, 1990; McClelland & Judd, 1993). Of the several factors that attenuate power in this context, three stand out as being most important: small sample sizes, measurement error (Busemeyer & Jones, 1983; Dunlap & Kemery, 1988), and information loss due to the use of coarse response scales (Aguinis & Stone-Romero, 1997; Russell, Pinto, & Bobko, 1991). To avoid these problems, some researchers have favored alternative models for detecting interactions, such as principal components regression, errors-in-variables regression, and nonlinear structural equation models (Anderson, Stone-Romero, & Tisak, 1996; Busemeyer & Jones, 1983; Champoux & Peters, 1987; Cronbach, 1987; Fisicaro & Tisak, 1994; Jaccard & Wan, 1995; Kenny & Judd, 1984; Paunonen & Jackson, 1988; Ping, 1996; Tisak, 1994).

Whereas much of the aforementioned research considers factors that reduce the power to detect true interactions, less attention has been paid to factors that increase the probability of detecting spurious interactions (Aiken & West, 1991; Busemeyer, 1980; Busemeyer & Jones, 1983; Cohen & Cohen, 1983). Nevertheless, a close reading of the literature suggests that at least three (nonindependent) factors fall in this second category: multicollinearity, inappropriate median splits, and the use of ordinal-level measures. This article focuses on the third factor. Specifically, it shows that psychological scales that are composed of binary items are at increased risk of producing spurious

Applied Psychological Measurement, Vol. 29 No. 2, March 2005, 87–105
DOI: 10.1177/0146621604272737

interaction effects in multiple regression models. This is especially true when the average difficulty of the test items is poorly matched to the trait levels of the examinees—a condition known as test inappropriateness (Embretson, 1996)[1]. The article then demonstrates how item response theory (IRT) can be used to reduce the Type I error rates for the interaction terms in MMR. To set the stage for the current study, three factors that produce spurious interaction effects are briefly reviewed. These potential sources of bias are (1) multicollinearity, (2) inappropriate median splits, and (3) ordinal-level measurement.

Factors Associated With Spurious Interaction Effects

Multicollinearity

It is well known that under many conditions, linear models can approximate nonlinear models (Ganzach, 1997; MacCallum & Mar, 1995). For instance, in regression models, when nonlinearity arises from product terms, quadratic effects are often well approximated by multiplicative terms when multicollinearity is high (Lubinski & Humphreys, 1990). To better understand this point, consider equations (1) and (2). Obviously, these equations will yield similar results when X_1 and X_2 are highly correlated.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \varepsilon, \tag{1}$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon. \tag{2}$$

Recognizing this problem, several investigators (e.g., Lubinski & Humphreys, 1990; Tellegen, Kamp, & Watson, 1982) have suggested that quadratic terms should always be included in models with interaction terms (for cautious notes concerning this recommendation, see MacCallum & Mar, 1995; Shepperd, 1991). Equation (3) illustrates this full model. Ganzach (1997) has written a particularly scholarly treatment of this issue.

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + \varepsilon. \tag{3}$$

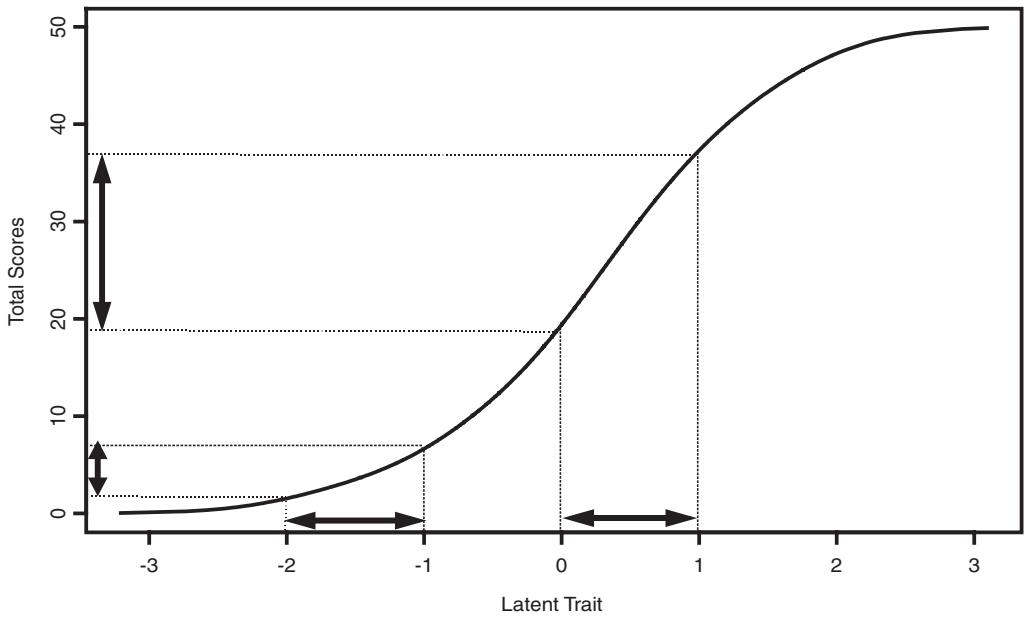
Inappropriate Median Splits

Inappropriate median splits represent a second source of spurious interactions in multiple regression models. This is especially true when the median splits have been performed on one or more independent variables (Bissonnette, Ickes, Bernstein, & Knowles, 1990; Maxwell & Delaney, 1993; Russell et al., 1991). The previous statement may surprise readers who learned that dichotomized variables are associated with reduced statistical power (Cohen, 1983; Peters & Van Voorhis, 1940; see also MacCallum, Zhang, Preacher, & Rucker, 2002). Although that point holds true in general, in some regression models, binary variables can increase the Type I error rates for interaction terms. Maxwell and Delaney (1993, p. 187) discussed this problem and showed that equation (2) approximates equation (1) when dichotomized independent variables are correlated. However, they also noted that equation (3) makes no sense in this context because the independent-variable correlation matrix is rank deficient. Specifically, if X is a 0/1 dichotomized variable, then $X = X^2$; moreover, X_1 and X_1^2, as well as X_2 and X_2^2, correlate 1.00.

Ordinal Measurement Level

A third factor associated with spurious interaction effects—and the problem considered most closely in this article—concerns the measurement scale of the independent and dependent variables. Because most questionnaires in the behavioral sciences yield ordinal-level measurements at best

Figure 1
Test Characteristic Curve Showing the Relationship Between the Latent Trait and the Expected Observed Scores



(Busemeyer & Jones, 1983; Cliff, 1989, 1992; Michell, 1999; Stevens, 1946), this problem is seemingly endemic to the field (see Embretson, 1996, for an insightful discussion of this problem in the context of ANOVA).

By definition, any monotonic function can be applied to ordinal data without disturbing the data ranks. Nevertheless, researchers should be aware that some rank-preserving transformations produce unintended consequences. For instance, a nonlinear function can transform an additive regression model into a multiplicative (i.e., interactive) model (Anderson, 1961; Bogartz, 1976; Embretson, 1996; Loftus, 1978). Moreover, when such functions are applied to interval-level data—or data that are approximately interval level—the resulting metric is no longer interval. Consider, for instance, a nonlinear function that is popular among psychometricians: the nonlinear regression of (expected) observed scores on latent trait scores under the logistic IRT model (Hambleton, & Swaminathan, 1985; Lord, 1980). Through this nonlinear transformation, the latent trait scores are stretched and squeezed at various points along the latent continuum via the test characteristic curve. Although psychometricians continue to debate the interval status of latent trait scores (e.g., Perline, Wright, & Wainer, 1979), all parties in this debate agree on one point. Namely, for virtually all psychological tests, the observed scores form an ordinal scale at best. Figure 1 illustrates a test characteristic curve (TCC) from a two-parameter logistic IRT model (Lord & Novick, 1968). Note that an important consequence of this nonlinear mapping is that the relative distances between the scores are modified as one moves from the latent trait scale (the x-axis) to the metric of the expected observed scores (the y-axis).

Results presented later demonstrate that under certain conditions, the transformation portrayed by the TCC can produce interaction effects in observed score models when such effects are not present among the latent traits. More specifically, spurious interactions and inflated Type I error rates (at the observed score level) can arise when (1) a logistic IRT model fits the data, and (2) item

difficulty levels are poorly matched to sample ability levels. Fortunately, as described below, these inflated Type I error rates can be controlled by performing the regression analyses on the estimated latent trait scores rather than the observed scores. Before illustrating this point, the next section briefly reviews the controversial role of measurement in the application of parametric statistics.

Measurement Levels, Spurious Interaction Effects, and IRT

Measurement Levels and the Application of Parametric Statistics

Psychologists, physicists, and philosophers of science have long debated the role of measurement in the use of parametric statistics (see Michell, 1999, for a scholarly review of this literature; see also Cliff, 1992). Within psychology, this debate became prominent when Stevens (1946) outlined his famous hierarchy of measurement scales in the prestigious journal *Science*. Few psychology students are unaware of Steven's four measurement levels (nominal, ordinal, interval, and ratio); fewer still are aware that this taxonomy is highly controversial (e.g., Gaito, 1980, 1986; Michell, 1986, 1999; Stine, 1989).

Frederick Lord (1953) captured the essence of this controversy in an oft-cited quip: "The numbers don't remember where they came from" (p. 751). Velleman and Wilkinson (1993) espouse a more recent version of this position. Cliff (1992), on the other hand, has developed a test theory for ordinal true scores in line with Stevens's (1946) taxonomy. Others have taken yet a different approach that considers the robustness of parametric statistics with ordinal data.

Nonlinear Transformations and Spurious Interaction Effects

The most common robustness study of this genre has considered the two-sample t test (Baker, Hardyck, & Petrinovich, 1966; Davison & Sharma, 1988; Maxwell & Delaney, 1985; Townsend & Ashby, 1984). In an elegant treatment of this issue, Davison and Sharma (1988) showed that, when standard assumptions of the t test hold (e.g., normal distributions with equal variances), then a mean difference on observed scores is indicative of a mean difference on latent scores. In a later study, they demonstrated that under similar conditions, researchers can be misled in more complicated designs. Specifically, if the relationship between the observed and latent trait is nonlinear, then the test for interaction in a 2×2 design may have an inflated Type I error rate. In other words, Davison and Sharma (1990) showed that it is possible to detect interaction effects at the observed score level when such effects are absent at the latent score level.

Several years earlier, Busemeyer and Jones (1983) demonstrated similar findings in the context of multiple regression. This work is particularly relevant to the present study because the authors considered linear and nonlinear (monotonic) transformations on both criterion *and* predictor variables. Their article is rich in ideas, and to give it full coverage is beyond the scope of this study. Nevertheless, several conclusions from Busemeyer and Jones's article deserve mention. First, the authors found that when linear transformations were applied to predictor and/or criterion scores, the risk of spurious interactions was small. That finding was not surprising. However, when nonlinear (monotonic) transformations were applied to predictor variables, quadratic effects were frequently identified as multiplicative effects (i.e., as interactions; cf. Lubinski & Humphreys, 1990). This misleading result occurred most often when both the criterion and predictor variables were transformed by monotonic nonlinear functions.

To better understand this issue, consider a simple three-variable regression model (adapted from Busemeyer & Jones, 1983, p. 554). Let θ_1, θ_2, and θ_3 denote three latent variables. Moreover, let X_{θ_1}, X_{θ_2}, and Y_{θ_3} represent observed variables that have been created from the latent variables

with a single, nonlinear function. The relationship between the latent and observed variables can be expressed as follows:

$$X_{\theta_1} = 1 + e^{-\theta_1}, \tag{4}$$

$$X_{\theta_2} = 1 + e^{-\theta_2}, \tag{5}$$

$$Y_{\theta_3} = 1 + e^{-\theta_3}. \tag{6}$$

Assume that the relationship between θ_1, θ_2, and θ_3 is linear:

$$\theta_3 = \theta_1 + \theta_2. \tag{7}$$

Solving equations (4) through (6) in terms of θ_1, θ_2, and θ_3 produces the following results:

$$\theta_1 = -\ln(X_{\theta_1} - 1), \tag{8}$$

$$\theta_2 = -\ln(X_{\theta_2} - 1), \tag{9}$$

and

$$\theta_3 = -\ln(Y_{\theta_3} - 1). \tag{10}$$

These relations imply that

$$\begin{aligned} -\ln(Y_{\theta_3} - 1) &= -\ln(X_{\theta_1} - 1) - \ln(X_{\theta_2} - 1), \\ \ln(Y_{\theta_3} - 1) &= \ln(X_{\theta_1} - 1) + \ln(X_{\theta_2} - 1), \\ \ln(Y_{\theta_3} - 1) &= \ln[(X_{\theta_1} - 1)(X_{\theta_2} - 1)], \\ (Y_{\theta_3} - 1) &= (X_{\theta_1} - 1)(X_{\theta_2} - 1), \end{aligned}$$

and

$$Y_{\theta_3} = 2 - X_{\theta_1} - X_{\theta_2} + X_{\theta_1} X_{\theta_2}. \tag{11}$$

Equation (11) merits attention because it clearly shows that a linear model can be transformed into a multiplicative model. This result is well known in statistics, and several researchers have developed procedures to remove scaling-induced nonadditivity (i.e., interactions) through score transformations (Anderson, 1970; Breiman & Friedman, 1985; Busemeyer, 1980; Krantz & Tversky, 1971; Tibshirani, 1988).

Controlling Spurious Interaction Effects With Item Response Theory

The previous discussion raises at least two important questions. Namely, (1) when should users of multiple regression be cautious about scaling-induced interactions, and (2) what steps, if any, can be taken to minimize these effects? Building on previous results (Embretson, 1996), the following sections demonstrate that modern measurement theory provides potential answers for both questions when the data are well characterized by an IRT model.

In the first study to explicitly consider spurious interaction effects from an IRT perspective, Embretson (1996) demonstrated that spurious interactions in ANOVA models were associated with so-called test inappropriateness. Working within the Rasch model (i.e., the one-parameter logistic IRT model), Embretson studied 2 × 2 ANOVA designs and assumed that the ANOVA

models were run on total scores of summed, binary item responses. In her study design, three factors were manipulated: average item difficulty, model main effect sizes, and test length. Tests of spurious interactions were calculated by deriving asymptotic expectations for the total score distributions using the Rasch model parameters. An important finding from her study was that spurious interactions were associated with tests that were either too easy (a condition associated with ceiling effects) or too hard (a condition associated with floor effects) for the simulated score groups. In other words, spurious interactions occurred most frequently when average item difficulties were poorly matched to average trait levels.

The present study was designed to extend Embretson's (1996) work in two important ways. First, Embretson worked with simulated parameters (item difficulties) rather than parameter estimates. Because researchers will never know the true parameter values for either items or individuals, it is important to replicate Embretson's design using estimated item parameters and trait scores. Second, the current study uses a two-parameter IRT model, rather than a one-parameter Rasch model, to better understand the role of item discrimination on the genesis of spurious interactions.

Method

Two Monte Carlo studies were conducted to explore the conditions under which scales composed of binary items produce spurious interactions in moderated multiple regression. In both studies, several factors were systematically varied in a cross-factor design: item difficulty, item discrimination, test length, and regression weight strength. Regression models were conducted on latent trait scores, observed scores, and estimated latent trait scores from IRT analyses.

A goal of both studies was to compare the relative performance of three metrics in moderated multiple regression: (1) error-free measurement using latent trait scores, (2) observed scores that were composed of summed binary item responses, and (3) estimated latent trait scores from IRT analyses. Equation (12) shows the regression model that was used to simulate the data. This equation represents a simple additive regression model for three latent trait scores (denoted by θ_j, where $j = 1, 2, 3$).

$$\theta_3 = \beta_1\theta_1 + \beta_2\theta_2 + \varepsilon. \tag{12}$$

For each data set, six models (shown in equations (13a)-(15b)) were tested to detect spurious interaction effects.

$$\tilde{\theta}_3 = \beta_1\theta_1 + \beta_2\theta_2, \tag{13a}$$

$$\tilde{\theta}_3 = \beta_1\theta_1 + \beta_2\theta_2 + \beta_3\theta_1\theta_2, \tag{13b}$$

$$\tilde{X}_3 = \beta_1 X_1 + \beta_2 X_2, \tag{14a}$$

$$\tilde{X}_3 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2, \tag{14b}$$

$$\tilde{\hat{\theta}}_3 = \beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2, \tag{15a}$$

$$\tilde{\hat{\theta}}_3 = \beta_1\hat{\theta}_1 + \beta_2\hat{\theta}_2 + \beta_3\hat{\theta}_1\hat{\theta}_2, \tag{15b}$$

where θ_j denotes the vector of latent trait scores for variable j, X_j denotes the vector of observed scores for variable j such that X_j equals the sum of n binary indicators of θ_j, and $\hat{\theta}_j$ represents

estimated trait values for variable j from an IRT analysis of the aforementioned item responses. In all equations, tildes are used to denote predicted scores. The IRT model is explained more fully below.

For each pair of models, spurious interactions were deemed present if the R-squared difference between the interactive and additive models (e.g., $R^2_{\text{Equation 13b}} - R^2_{\text{Equation 13a}}$) was significant at $\alpha = 0.05$.

Study Design

Four design features were manipulated to elucidate the conditions producing spurious interactions in observed score and estimated trait score models: item difficulty, item discrimination, regression weight (i.e., the magnitude of the regression weights in equation (12)), and test length (number of binary items in X_j).

Three levels of average item difficulty were studied. The three levels were varied to simulate so-called easy, moderate, and difficult tests. Test difficulty was defined by the mean difference between the average latent trait score and the average item difficulty (b) for a given scale. The latent trait scores were drawn from a normal distribution with a mean 0.00 and unit standard deviation. In the first Monte Carlo study, the easy, medium, and difficult tests were simulated by selecting item difficulty values (b) from one of three normal distributions: N(–1.50, 1.00) for easy tests, N(0.00, 1.00) for moderate tests, and N(1.50, 1.00) for difficult tests. In the second study, highly peaked tests were simulated—that is, tests with high fidelity but low bandwidth (Stocking, 1987)—by selecting item difficulty values from N(–1.50, 0.50) for easy tests, N(0.00, 0.50) for moderate tests, and N(1.50, 0.50) for difficult tests. Peaked tests are characterized by having a narrow range of item difficulty values (i.e., many item difficulties are similar in value). Peaked tests are common in several testing contexts, such as psychopathology assessment, where an important goal is the discrimination between normal-range and pathological trait levels. The second simulation study was designed to address these situations. Study 2 was similar in design to Study 1 with one important difference: It includes highly peaked tests with test information maximized at the low, middle, or high sections of the trait range.

In both studies, two levels of item discrimination (a) were selected to represent moderate and highly discriminating tests. In the first condition, the discrimination values were sampled from a uniform distribution that ranged from 0.31 to 0.58. In the second condition, the uniform distribution ranged from 0.58 to 1.13. Uniform distributions were chosen because previous work (e.g., Reise & Waller, 2003) suggests that the distribution of item discrimination values from widely used psychology tests (such as the Minnesota Multiphasic Personality Inventory) is well approximated by a uniform distribution. The two ranges of these distributions were carefully selected to represent items with moderate to strong factor loadings.[2]

Two levels of regression weights were studied. In both conditions, the within-model regression weights were equal ($\beta_1 = \beta_2$) and had values of either 0.30 or 0.50. Finally, test length was varied to represent short or moderately long tests by simulating either 20 or 50 items. These choices resulted in a simulation design with 24 levels (in each study): 3 (item difficulty) $\times$ 2 (item discrimination) $\times$ 2 (regression weight) $\times$ 2 (test length).

Procedure

Pseudo-random observations on three latent variables (θ_1, θ_2, θ_3) were simulated in the R programming language (Ihaka & Gentleman, 1996) in the following manner. First, values for

the two independent variables, θ_1 and θ_2, were randomly selected from $N(0.00, 1.00)$ distributions. Next, values for the dependent variable, θ_3, were simulated using the following equation:

$$\theta_3 = \beta_1\theta_1 + \beta_2\theta_2 + \sqrt{1 - (\beta_1^2 + \beta_2^2)} \times e, \tag{16}$$

where $e \sim N(0.00, 1.00)$.

Values for the three observed scores (X_1, X_2, and X_3) were simulated from the previously generated latent trait scores for 250 participants using equation (17), which represents the item response function for the two-parameter logistic IRT model.

$$P(U_{ik} = 1|\theta_k,\ a_i,\ b_i) = \frac{1}{1 + e^{-1.7a_i(\theta_k - b_i)}}, \tag{17}$$

where U_{ik} denotes the ith item response for the kth participant with latent trait score θ_k, a_i is the item discrimination for the ith item, and b_i is the item difficulty for the ith item.

For each item response, the probability that a hypothetical examinee correctly answered the item given his or her latent ability was calculated. Next, this probability was converted to a binary (0/1) score as follows: (1) The probability was compared to a randomly sampled number from a uniform distribution that ranged from 0.00 to 1.00, and (2) if the estimated probability exceeded the random number, the item response was considered a "pass" and was recoded 1. Otherwise, it was recoded 0. This is a common method for generating simulated scores in IRT studies (Embretson, 1994; Stone, 1992) because it generates observed item responses with an appropriately modeled error term. Observed scores were computed by summing the simulated item responses on either 20-item or 50-item tests.

IRT item parameter estimates were obtained with a user-written program that implements marginal maximum likelihood estimation via the EM algorithm. Estimated latent trait scores ($\hat{\theta}_j$) were also calculated via the expected a posteriori estimation method (Bock & Mislevy, 1982). To facilitate the data and file linking from the (IRT) parameter estimation and regression testing phases of the present studies, all IRT analyses were performed with a user-written FORTRAN program that was modeled on code given in Baker (1992). To ensure program accuracy, the estimates from this program were compared to the estimated item parameters and trait scores from BILOG 3.10 on several test data sets. Across all tested samples, the parameter estimates from the two programs correlated at least .99.

In both studies, sample size was set to 250 hypothetical examinees. This choice was not arbitrary. Previous work (Stone, 1992) indicates that a sample size of 250 is sufficient to yield accurate item parameter estimates with marginal maximum likelihood estimation (see also Drasgow 1989; Seong, 1990). Moreover, this sample size is typical of applied individual-differences research using moderated multiple regression. A survey of 20 randomly selected studies published between 1985 and 2003 (selected from the PsychINFO database using several keywords such as *moderated multiple regression*, *interaction effect*, *multiple regression*, *personality*, *social support*, and *stress*) had a mean sample size of 208.4 ($SD = 141.77$). Finally, using a user-written program in *R* (Ihaka & Gentleman, 1996), Type I error rates for 500 replications in each condition were tallied for the three moderated multiple regression models (i.e., the models based on θ_j, X_j, and $\hat{\theta}_j$).

Results

The main findings of Study 1 are presented in Table 1. Although many aspects of this table merit comment, one finding stands out. Namely, when participant characteristics and test characteristics are well matched, the Type I error rates for the interaction terms are well approximated by the

theoretical alpha level. This was true regardless of whether the analyses were carried out on latent traits (θ), estimated latent traits ($\hat{\theta}$), or observed (X) scores.

Notice, however, that when the item difficulty and trait distributions were poorly matched, the empirical Type I error rates for the observed score models were often considerably higher than the expected alpha levels. Specifically, when item discriminations were high ($a \sim U(0.58, 1.13)$) or main effects were strong ($\beta_1, \beta_2 = 0.5$), inflated Type I error rates for the observed score models were often two to three times larger than the expected Type I error rates (range $= .09-.15$). When both conditions prevailed (i.e., high discrimination and strong main effects, such as in Conditions 7, 8, 23, and 24), Type I error rates were five to eight times larger than expected (range $= .27 - .41$). Note that test length is also a salient factor in these results. Although test length explains less variance in Type I error rates than the preceding factors, as a general finding, longer tests produced more Type I errors than shorter tests.

To better understand the conditions that produced spurious interactions, two ANOVAs were performed on the average Type I error rates from Table 1. Table 2 displays the findings from the analysis of the observed score (X_j) results. Note that the model includes all main effects, two-way interactions, and three-way interactions among the aforementioned design factors. Also note that the final column in Table 2 reports effect sizes as measured by omega-squared (ω^2; Hays, 1994, p. 409).

For the observed score models, three factors account for the brunt of the variance in the empirical Type I error rates: (1) average item difficulty, (2) average item discrimination, and (3) regression strength (i.e., as indexed by regression weight magnitude). Specifically, poorly matched tests (i.e., tests that were either too easy or too hard for the examinees) with highly discriminating items were most vulnerable to spurious interaction effects. Moreover, as the regression weights increased, the Type I error rate also increased.

These findings raise an important question for researchers who work with total scores that are produced by summing binary item responses. Namely, can researchers predict when an observed score regression model is most vulnerable to the aforementioned biasing influences? To address this question, five types of indicators were examined that may flag problematic analyses (i.e., scenarios in which spurious interactions are likely to occur): (1) classical item difficulties[3], (2) scale reliability coefficients (KR20; Kuder & Richardson, 1937), (3) the Shapiro-Wilks normality test (SW; Shapiro & Wilks, 1965) of model residuals, (4) test skewness, and (5) test kurtosis. The Shapiro-Wilks test is hypothesized to be particularly informative because it allows one to test an important assumption of the statistical assessment of the model parameters—namely, that the model residuals are normally distributed. If the residuals are not normally distributed, as determined by the Shapiro-Wilks test, then the p value for the interaction term may be biased. When this occurs, researchers could use alternative methods for determining parameter significance, such as bootstrap-generated confidence intervals (Davison & Hinkley, 1997; Efron & Tibshirani, 1993).

The last nine columns in Table 1 summarize the findings from the five indicators discussed above. To conserve space, the classical item difficulties, reliability coefficients, skewness, and kurtosis coefficients were averaged across the three manifest variables (X_1, X_2, and X_3). This does not bias the interpretation of the results because the within-condition statistics for the observed scores were virtually identical.

The results in Table 1 indicate that "spurious interactions" occur with increased frequency with highly reliable tests that have extreme item difficulty values (i.e., with tests that are either too easy or too hard for the trait distribution). This conclusion is qualified by quotes in the previous sentence to remind readers that some of the significance levels in Table 1 may be inaccurate due to violations of the normality assumption, as assessed by the SW test.

Turning our attention to the residual analyses, the entries in Table 1 labeled the "Shapiro-Wilks test" denote the proportion of samples in which this test was *not* significant. In this context, a

Table 1

Type I Error Rates for Spurious Interaction Effects

Number	b	a	$\beta_1,\ \beta_2$	Items	θ	X	$\hat{\theta}$	$\bar{p}$ Value	KR20	SW_θ	SW_X	$SW_{\hat{\theta}}$	SK_X	KR_X	$SK_{\hat{\theta}}$	$KR_{\hat{\theta}}$
1	N(–1.50, 1.00)	U(0.31, 0.58)	0.30	20	0.06	0.06	0.05	0.71 (0.13)	0.65	0.96	0.06	0.50	0.57	0.02	0.30	0.29
2	N(–1.50, 1.00)	U(0.31, 0.58)	0.30	50	0.06	0.07	0.05	0.71 (0.13)	0.83	0.96	0.06	0.91	0.60	0.08	0.15	0.21
3	N(–1.50, 1.00)	U(0.31, 0.58)	0.50	20	0.06	0.09	0.06	0.71 (0.13)	0.65	0.96	0.20	0.71	0.56	0.01	0.30	0.30
4	N(–1.50, 1.00)	U(0.31, 0.58)	0.50	50	0.06	0.13	0.06	0.71 (0.13)	0.83	0.96	0.27	0.95	0.60	0.08	0.15	0.20
5	N(–1.50, 1.00)	U(0.58, 1.13)	0.30	20	0.06	0.10	0.05	0.79 (0.16)	0.81	0.96	0.00	0.27	1.07	0.92	0.38	0.36
6	N(–1.50, 1.00)	U(0.58, 1.13)	0.30	50	0.06	0.11	0.05	0.79 (0.16)	0.92	0.96	0.00	0.60	1.09	0.93	0.16	0.19
7	N(–1.50, 1.00)	U(0.58, 1.13)	0.50	20	0.06	0.27	0.10	0.79 (0.16)	0.81	0.96	0.02	0.78	1.07	0.91	0.38	0.36
8	N(–1.50, 1.00)	U(0.58, 1.13)	0.50	50	0.06	0.39	0.06	0.79 (0.16)	0.92	0.96	0.01	0.70	1.09	0.93	0.15	0.16
9	N(0.00, 1.00)	U(0.31, 0.58)	0.30	20	0.06	0.04	0.04	0.50 (0.15)	0.69	0.96	0.79	0.92	0.00	0.49	0.00	0.35
10	N(0.00, 1.00)	U(0.31, 0.58)	0.30	50	0.06	0.04	0.05	0.50 (0.16)	0.85	0.96	0.84	0.97	0.00	0.51	0.00	0.17
11	N(0.00, 1.00)	U(0.31, 0.58)	0.50	20	0.06	0.04	0.04	0.50 (0.15)	0.69	0.96	0.89	0.95	0.00	0.49	0.00	0.35
12	N(0.00, 1.00)	U(0.31, 0.58)	0.50	50	0.06	0.04	0.05	0.50 (0.16)	0.85	0.96	0.93	0.96	0.00	0.51	0.01	0.17
13	N(0.00, 1.00)	U(0.58, 1.13)	0.30	20	0.06	0.04	0.04	0.50 (0.22)	0.85	0.96	0.36	0.95	0.00	0.79	0.00	0.38
14	N(0.00, 1.00)	U(0.58, 1.13)	0.30	50	0.06	0.04	0.05	0.50 (0.22)	0.94	0.96	0.40	0.97	0.00	0.82	0.00	0.20
15	N(0.00, 1.00)	U(0.58, 1.13)	0.50	20	0.06	0.02	0.03	0.50 (0.22)	0.85	0.96	0.82	0.96	0.00	0.79	0.00	0.38
16	N(0.00, 1.00)	U(0.58, 1.13)	0.50	50	0.06	0.03	0.03	0.50 (0.22)	0.94	0.96	0.88	0.95	0.00	0.82	0.00	0.20

17	$N(1.50, 1.00)$	$U(0.31, 0.58)$	0.30	20	0.06	0.05	0.04	0.29 (0.13)	0.66	0.97	0.06	0.47	0.57	0.06	0.30	0.27
18	$N(1.50, 1.00)$	$U(0.31, 0.58)$	0.30	50	0.06	0.05	0.04	0.29 (0.13)	0.83	0.97	0.04	0.90	0.61	0.11	0.15	0.19
19	$N(1.50, 1.00)$	$U(0.31, 0.58)$	0.50	20	0.06	0.07	0.06	0.29 (0.13)	0.65	0.97	0.17	0.69	0.57	0.05	0.30	0.28
20	$N(1.50, 1.00)$	$U(0.31, 0.58)$	0.50	50	0.06	0.15	0.04	0.29 (0.13)	0.83	0.97	0.24	0.93	0.61	0.11	0.16	0.19
21	$N(1.50, 1.00)$	$U(0.58, 1.13)$	0.30	20	0.06	0.11	0.06	0.21 (0.16)	0.81	0.97	0.00	0.26	1.07	0.90	0.38	0.35
22	$N(1.50, 1.00)$	$U(0.58, 1.13)$	0.30	50	0.06	0.12	0.05	0.21 (0.16)	0.92	0.97	0.00	0.63	1.10	0.95	0.15	0.19
23	$N(1.50, 1.00)$	$U(0.58, 1.13)$	0.50	20	0.06	0.31	0.10	0.21 (0.16)	0.81	0.97	0.01	0.74	1.07	0.92	0.38	0.34
24	$N(1.50, 1.00)$	$U(0.58, 1.13)$	0.50	50	0.06	0.41	0.08	0.21 (0.16)	0.92	0.97	0.01	0.70	1.10	0.97	0.12	0.16

Note. Number of subjects = 250; Number of iterations = 500; b = item difficulty; a = item discrimination; β_1, β_2 = regression weights for main effects; Items = number of test items; θ = Type I error rates of latent trait score; X = Type I error rates of observed (summed total) score; $\hat{\theta}$ = Type I error rates of estimated latent trait score; $\bar{p}$ value = overall mean item difficulty (proportion of examinees who answer the item correctly) across three observed scores, and number in parentheses is the overall mean standard deviation of item difficulty; KR20 = overall mean internal consistency (Kuder-Richardson 20) across three observed scores; SW_θ = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the latent trait score; SW_X = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the observed score; $SW_{\hat{\theta}}$ = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the estimated latent trait score; SK_X = average skewness values of observed score across 500 iterations; KR_X = average kurtosis values of observed score across 500 iterations; $SK_{\hat{\theta}}$ = average skewness values of estimated latent trait score across 500 iterations; $KR_{\hat{\theta}}$ = average kurtosis values of estimated latent trait score across 500 iterations.

Table 2

Factors Influencing Spurious Interaction Results in Observed Score Regression Models

	df	Sum of Squares × 100	Mean Square ×100	F Value	Pr(> F)	$\omega^2 \times 100$
b	2	761.58	380.79	186.51	0.01	25.9
a	1	522.67	522.67	256.00	0.00	17.8
β	1	522.67	522.67	256.00	0.00	17.8
Items	1	60.17	60.17	29.47	0.03	2.0
$b \times a$	2	312.58	156.29	76.55	0.01	10.5
$b \times \beta$	2	308.08	154.04	75.45	0.01	10.4
$b\times$ Items	2	25.58	12.79	6.26	0.14	0.7
$a \times \beta$	1	204.17	204.17	100.00	0.01	6.9
$a\times$ Items	1	6.00	6.00	2.94	0.23	0.1
$\beta\times$ Items	1	42.67	42.67	20.90	0.05	1.4
$b \times a \times \beta$	2	130.08	65.04	31.86	0.03	4.3
$b \times a\times$ Items	2	3.25	1.62	0.80	0.56	0.0
$b \times \beta\times$ Items	2	18.08	9.04	4.43	0.18	0.5
$a \times \beta\times$ Items	1	4.17	4.17	2.04	0.29	0.1
Residuals	2	4.08	2.04			

Note. Items = number of test items; b = average item difficulty; a = average item discrimination; β = regression weight; ω^2 = effect size.

nonsignificant finding indicates that the null hypothesis of normally distributed residuals (from the regression analyses) cannot be rejected. Accordingly, low numbers flag conditions under which the null hypothesis was frequently rejected in the 500 replications of each cell. A quick glance of these results reveals striking differences between the three metrics under study. For instance, the SW test was significant in a high proportion of samples under many conditions in the observed score models. This trend was most pronounced in Condition 8, in which the normality test was rejected in approximately 99% of data sets.

The final four columns in Table 1 report average (across 500 replications) skewness and kurtosis values for the observed scores and estimated latent trait scores. Comparing these values to their standard errors[4] reveals that the score distributions are highly skewed when test inappropriateness is extreme. Within the parameter space, tests with high item discriminations produced kurtotic distributions for the observed scores, whereas tests with elevated skewness and kurtosis values—such as tests in Conditions 5, 6, 7, 8, 21, 22, 23, and 24—were associated with spurious interaction effects in the regression models.

The ANOVA results for the estimated latent trait scores are reported in Table 3. Although no effects in this table exceeded critical significance levels, the pattern of these results is highly similar to that reported in Table 2. This suggests that with larger sample sizes, some tests could yield significant results. Nevertheless, these data provide an important result for applied researchers. Namely, as shown in Table 1, when the analyses were conducted on estimated latent trait scores, rather than observed scores, the Type I error rates for the interaction terms were considerably closer to their nominal alpha levels. Although this general trend is undoubtedly true, readers should recognize that the reported significance levels are likely biased due to violations of the residual normality assumption (e.g., in Conditions 1, 5, 17, and 21, as indicated by the SW test).

Table 3

Factors Influencing Spurious Interaction Rates in Estimated Latent Trait Score Regression Models

	df	Sum of Squares × 100	Mean Square × 100	F Value	Pr(> F)	$\omega^2 \times 100$
b	2	17.58	8.79	11.10	0.08	21.0
a	1	6.00	6.00	7.58	0.11	6.8
β	1	8.17	8.17	10.32	0.08	9.7
Items	1	1.50	1.50	1.89	0.30	0.9
$b \times a$	2	12.25	6.12	7.74	0.11	14.0
$b \times \beta$	2	11.08	5.54	7.00	0.12	12.5
$b\times$ Items	2	4.75	2.37	3.00	0.25	4.2
$a \times \beta$	1	1.50	1.50	1.89	0.30	0.9
$a\times$ Items	1	1.50	1.50	1.89	0.30	0.9
$\beta\times$ Items	1	2.67	2.67	3.37	0.21	2.5
$b \times a \times \beta$	2	4.75	2.37	3.00	0.25	4.2
$b \times a\times$ Items	2	0.75	0.37	0.47	0.68	0.0
$b \times \beta\times$ Items	2	0.58	0.29	0.37	0.73	0.0
$a \times \beta\times$ Items	1	0.67	0.67	0.84	0.46	0.0
Residuals	2	1.58	0.79			

Note. Items = number of test items; b = average item difficulty; a = average item discrimination; β = regression weight; ω^2 = effect size.

Peaked Tests and Spurious Interactions

As noted previously, a so-called peaked test is a test that measures reliably in a limited range of the trait distribution. Peaked tests are common in several assessment domains, such as in the assessment of psychopathology. Thus, it is important to determine whether peaked tests are also associated with spurious interaction effects. To investigate this issue, a second study was performed using item parameters that better mimicked the characteristics of peaked tests. In particular, in Study 2, all design features of Study 1 were left intact except that the standard deviation of the item difficulties was uniformly set to .50 rather than to 1.00.

The major findings from Study 2 are summarized in Table 4. Although they are more extreme, the results in Table 4 are similar to those reported in Table 1. For instance, when test characteristics are well matched to the average trait level of the examinees, the Type I error rates for the interaction terms are close to their nominal alpha levels in the latent trait, estimated latent trait, and observed score models. Note, however, that when test inappropriateness is extreme, the spurious interaction rates are considerably elevated over similar conditions in Table 1. Note also that by using estimated latent trait scores, a researcher can ameliorate these differences, as demonstrated in column 8 of Table 4. ANOVA tables for these findings are available upon request.

Discussion

Two Monte Carlo studies were conducted to explore the Type I error rates in moderated multiple regression of observed scores or estimated latent trait scores from a two-parameter logistic IRT model. The results showed that MMR Type I error rates were substantially higher than the nominal alpha levels when scale scores were composed of summed binary item responses (e.g., true/false, yes/no, disagree/agree items). Performing the regression analyses on estimated trait scores ($\hat{\theta}$) from a two-parameter logistic model improved the error detection rates considerably.

Table 4

Type I Error Rates for Spurious Interaction Effects in Regression Models With Peaked Tests

Number	b	a	β_1, β_2	Items	θ	X	$\hat{\theta}$	$\bar{p}$ Value	KR20	SW_θ	SW_X	$SW_{\hat{\theta}}$	SK_X	KR_X	$SK_{\hat{\theta}}$	$KR_{\hat{\theta}}$
1	N(–1.50, 0.50)	U(0.31, 0.58)	0.30	20	0.06	0.06	0.05	0.73 (0.07)	0.67	0.96	0.01	0.33	0.66	0.14	0.34	0.30
2	N(–1.50, 0.50)	U(0.31, 0.58)	0.30	50	0.06	0.08	0.04	0.73 (0.08)	0.83	0.96	0.01	0.87	0.70	0.22	0.18	0.21
3	N(–1.50, 0.50)	U(0.31, 0.58)	0.50	20	0.06	0.10	0.07	0.73 (0.07)	0.67	0.96	0.09	0.66	0.66	0.12	0.34	0.31
4	N(–1.50, 0.50)	U(0.31, 0.58)	0.50	50	0.06	0.17	0.07	0.73 (0.08)	0.83	0.96	0.12	0.94	0.70	0.22	0.18	0.21
5	N(–1.50, 0.50)	U(0.58, 1.13)	0.30	20	0.06	0.12	0.06	0.82 (0.08)	0.83	0.96	0.00	0.01	1.38	1.69	0.52	0.34
6	N(–1.50, 0.50)	U(0.58, 1.13)	0.30	50	0.06	0.15	0.05	0.82 (0.09)	0.93	0.96	0.00	0.63	1.43	1.81	0.27	0.29
7	N(–1.50, 0.50)	U(0.58, 1.13)	0.50	20	0.06	0.36	0.13	0.82 (0.08)	0.83	0.96	0.00	0.52	1.38	1.66	0.51	0.35
8	N(–1.50, 0.50)	U(0.58, 1.13)	0.50	50	0.06	0.49	0.11	0.82 (0.09)	0.93	0.96	0.00	0.94	1.42	1.81	0.27	0.30
9	N(0.00, 0.50)	U(0.31, 0.58)	0.30	20	0.06	0.03	0.04	0.50 (0.09)	0.71	0.96	0.78	0.94	0.00	0.58	0.00	0.41
10	N(0.00, 0.50)	U(0.31, 0.58)	0.30	50	0.06	0.02	0.04	0.50 (0.09)	0.86	0.96	0.79	0.97	0.00	0.59	0.00	0.20
11	N(0.00, 0.50)	U(0.31, 0.58)	0.50	20	0.06	0.04	0.03	0.50 (0.09)	0.71	0.96	0.89	0.95	0.00	0.58	0.00	0.41
12	N(0.00, 0.50)	U(0.31, 0.58)	0.50	50	0.06	0.03	0.04	0.50 (0.09)	0.86	0.96	0.94	0.96	0.00	0.59	0.00	0.20
13	N(0.00, 0.50)	U(0.58, 1.13)	0.30	20	0.06	0.04	0.04	0.50 (0.12)	0.88	0.96	0.12	0.89	0.00	1.01	0.00	0.53
14	N(0.00, 0.50)	U(0.58, 1.13)	0.30	50	0.06	0.05	0.04	0.50 (0.13)	0.95	0.96	0.14	0.96	0.00	1.04	0.00	0.31
15	N(0.00, 0.50)	U(0.58, 1.13)	0.50	20	0.06	0.03	0.03	0.50 (0.12)	0.88	0.96	0.77	0.97	0.00	1.02	0.00	0.54
16	N(0.00, 0.50)	U(0.58, 1.13)	0.50	50	0.06	0.02	0.03	0.50 (0.13)	0.95	0.96	0.85	0.95	0.00	1.04	0.00	0.31

17	$N(1.50, 0.50)$	$U(0.31, 0.58)$	0.30	20	0.06	0.04	0.04	0.27 (0.08)	0.67	0.97	0.01	0.36	0.67	0.16	0.35	0.28
18	$N(1.50, 0.50)$	$U(0.31, 0.58)$	0.30	50	0.06	0.05	0.04	0.27 (0.08)	0.83	0.97	0.02	0.89	0.71	0.24	0.18	0.20
19	$N(1.50, 0.50)$	$U(0.31, 0.58)$	0.50	20	0.06	0.10	0.06	0.27 (0.08)	0.67	0.97	0.09	0.62	0.66	0.16	0.34	0.29
20	$N(1.50, 0.50)$	$U(0.31, 0.58)$	0.50	50	0.06	0.18	0.05	0.27 (0.08)	0.83	0.97	0.14	0.93	0.71	0.23	0.18	0.21
21	$N(1.50, 0.50)$	$U(0.58, 1.13)$	0.30	20	0.06	0.15	0.07	0.18 (0.08)	0.83	0.97	0.00	0.03	1.39	1.71	0.52	0.32
22	$N(1.50, 0.50)$	$U(0.58, 1.13)$	0.30	50	0.06	0.16	0.06	0.18 (0.09)	0.93	0.97	0.00	0.68	1.43	1.83	0.28	0.28
23	$N(1.50, 0.50)$	$U(0.58, 1.13)$	0.50	20	0.06	0.41	0.17	0.18 (0.08)	0.83	0.97	0.00	0.47	1.39	1.73	0.53	0.31
24	$N(1.50, 0.50)$	$U(0.58, 1.13)$	0.50	50	0.06	0.53	0.09	0.18 (0.09)	0.93	0.97	0.00	0.90	1.43	1.85	0.28	0.27

Note. Number of subjects = 250; Number of iterations = 500; b = item difficulty; a = item discrimination; β_1, β_2 = regression weights for main effects; Items = number of test items; θ = Type I error rates of latent trait score; X = Type I error rates of observed (summed total) score; $\hat{\theta}$ = Type I error rates of estimated latent trait score; $\bar{p}$ value = overall mean item difficulty (proportion of examinees who answer the item correctly) across three observed scores, and number in parentheses is the overall mean standard deviation of item difficulty; KR20 = overall mean internal consistency (Kuder-Richardson 20) across three observed scores; SW_θ = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the latent trait score; SW_X = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the observed score; $SW_{\hat{\theta}}$ = proportion of cases in which the p value of the Shapiro-Wilks test is larger than .05 across 500 iterations for the estimated latent trait score; SK_X = average skewness values of observed score across 500 iterations; KR_X = average kurtosis values of observed score across 500 iterations; $SK_{\hat{\theta}}$ = average skewness values of estimated latent trait score across 500 iterations; $KR_{\hat{\theta}}$ = average kurtosis values of estimated latent trait score across 500 iterations.

These studies extend earlier work in this area (Embretson, 1996) by (1) systematically varying test inappropriateness and (2) using a two-parameter logistic IRT model. Earlier work by Embretson (1996) examined spurious interaction effects in ANOVA models with error-free IRT item parameters. The present studies explored spurious interaction effects in regression models with estimated item parameters and estimated trait scores. By using parameter estimates rather than parameters, the current findings should better reflect the severity of the problem in applied settings.

The most important conclusion from these studies can be stated as follows. When test difficulty is poorly matched to sample characteristics (i.e., average trait levels), spurious interactions can occur with increased frequency in moderated multiple regression. Under extreme conditions—defined as a conjunction of (1) large test inappropriateness, (2) long tests, (3) highly reliable tests, and (4) peaked tests—the spurious interaction rate of observed score models can be an order of magnitude higher than the nominal alpha rate. Importantly, under the conditions studied in this article, the Type I error rates approximated their nominal alpha levels in both the latent trait and estimated latent trait models. This was true despite the fact that in the latter model, the latent traits were calculated using estimated item parameters.

The previous findings raise an interesting question regarding the role of score distributions in the genesis of spurious interactions. As a general trend, spurious interactions were observed most often when the regression models were run on variables with markedly nonnormal distributions. Thus, the interesting question is whether score transformations to normality—such as the Box-Cox power transform—would be as effective as the estimated latent trait scores in controlling the Type I error rate. The results of several modest simulations suggest that the answer to this question is yes when test inappropriateness is moderate and no when test inappropriateness is extreme. Specifically, when Box-Cox transformations were applied to the observed scores in Condition 20 of Table 1, the resulting Type I error rate was .05 to two decimal places. However, when normalizing transformations were applied to the data in Condition 24 in Tables 1 and 4 (the most extreme conditions), the Type I error rates were .19 and .31, respectively. Although these values are less than those for the untransformed data, they are still considerably higher than the associated error rates for the estimated latent trait scores.

The findings from this work have important implications for applied researchers. To wit, when working with scales that are composed of summed binary item responses, the nominal Type I error rate for detecting spurious interactions can underestimate the empirical Type I error rate in moderated multiple regression. This potential bias is most likely to occur in tests in which the average item difficulties are poorly matched to the average trait levels of the examinees, that is, under conditions characterized by test inappropriateness. Running the regression models on estimated latent trait scores from a well-fitting IRT model provides a viable means of ameliorating this potential bias.

Notes

1. At extreme levels, test inappropriateness produces so-called floor and ceiling effects, a condition that attenuates measurement precision for individuals with scores at the tails of the trait distribution.

2. Takane and De Leeuw (1987) have shown that when the latent distribution is normal, two-parameter logistic item discrimination values can be transformed into factor loadings by the following equation: $\frac{a_i(1.7)}{\sqrt{1+\{a_i(1.7)\}^2}} = \lambda_i$. According to this equation, the lower and upper limits of the uniform distribution—0.38, 0.58, and 1.13—correspond to factor loadings of .47, .70, and .89, respectively.

3. Median item difficulty values were also checked. Because the mean and median values were highly similar, only the mean values are reported.

4. Approximate standard errors for skewness and kurtosis can be estimated using formulas reported by Tabachnick and Fidell (2001, p. 73): $s_{\text{skewness}} = \sqrt{\frac{6}{N}}$ and $s_{\text{kurtosis}} = \sqrt{\frac{24}{N}}$, where N denotes sample size. Using these equations and a sample size of 250, two standard errors for the skewness and kurtosis coefficients are .31 and .62, respectively.

References

Aguinis, H., & Stone-Romero, E. F. (1997). Methodological artifacts in moderated multiple regression and their effects on statistical power. *Journal of Applied Psychology*, *82*, 192-206.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Anderson, L. E., Stone-Romero, E. F., & Tisak, J. (1996). A comparison of bias and mean squared error in parameter estimates of interaction effects: Moderated multiple regression versus-error-in-variables regression. *Multivariate Behavioral Research*, *31*, 61-94.

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

Anderson, N. H. (1970). Functional measurement and psychophysical judgment. *Psychological Review*, *77*, 153-170.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, *16*, 291-309.

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

Bissonnette, V., Ickes, W., Bernstein, I., & Knowles, E. (1990). Personality moderating variables: A warning about statistical artifact and a comparison of analytic techniques. *Journal of Personality*, *58*, 567-587.

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

Bogartz, R. S. (1976). On the meaning of statistical interactions. *Journal of Experimental Child Psychology*, *22*, 178-183.

Breiman, L., & Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 580-598.

Busemeyer, J. R. (1980). Importance of measurement theory, error theory, and experimental design for testing the significance of interactions. *Psychological Bulletin*, *88*, 237-244.

Busemeyer, J. R., & Jones, L. E. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin*, *93*, 549-562.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size, and power in moderated regression analysis. *Journal of Occupational Psychology*, *60*, 243-255.

Cliff, N. (1989). Ordinal consistency and ordinal true scores. *Psychometrika*, *54*, 75-91.

Cliff, N. (1992). Abstract measurement theory and the revolution that never happened. *Psychological Science*, *3*, 186-190.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*, 249-253.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analyses recently proposed. *Psychological Bulletin*, *102*, 414-417.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their applications*. New York: Cambridge University Press.

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.

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

Dunlap, W. P., & Kemery, E. R. (1988). Effects of predictor intercorrelations and reliabilities on moderated multiple regression. *Organizational Behavior and Human Decision Processes*, *41*, 248-258.

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

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* (pp. 213-248). Ottawa, Ontario: University of Ottawa.

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

Fisicaro, S. A., & Tisak, J. (1994). A theoretical note on the stochastics of moderated multiple regression. *Educational and Psychological Measurement*, *54*, 32-41.

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

Gaito, J. (1986). Some issues in the measurement-statistics controversy. *Canadian Psychology*, *27*, 63-68.

Ganzach, Y. (1997). Misleading interaction and curvilinear terms. *Psychological Methods*, *2*, 235-247.

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

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt, Brace.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*, 299-314.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression*. Newbury Park, CA: Sage.

Jaccard, J., & Wan, C. K. (1995). Measurement error in the analysis of interaction effects between continuous predictors using multiple regression: Multiple indicator and structural equation approaches. *Psychological Bulletin*, *117*, 348-357.

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *90*, 201-210.

Krantz, D. H., & Tversky, A. (1971). Conjoint-measurement analysis of composition rules in psychology. *Psychological Review*, *78*, 151-169.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Loftus, G. R. (1978). On interpretation of interactions. *Memory and Cognition*, *6*, 312-319.

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

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

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

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effect": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin*, *107*, 385-393.

MacCallum, R. C., & Mar, C. M. (1995). Distinguishing between moderator and quadratic effects in multiple regression. *Psychological Bulletin*, *118*, 405-421.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*, 19-40.

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

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*, 181-190.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, *114*, 376-390.

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

Michell, J. (1999). *Measurement in psychology: Critical history of a methodological concept*. Cambridge, UK: Cambridge University Press.

Paunonen, S. V., & Jackson, D. N. (1988). Type I error rates for moderated multiple regression analysis. *Journal of Applied Psychology*, *73*, 569-573.

Perline, R., Wright, B., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, *3*, 237-255.

Peters, C. C., & Van Voorhis, W. R. (1940). *Statistical procedures and their mathematical bases*. New York: McGraw-Hill.

Ping, R. A. (1996). Latent variable regression: A technique for estimating interaction and quadratic coefficients. *Multivariate Behavioral Research*, *31*, 95-120.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods*, *8*(2), 164-184.

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.

Seong, T. J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*, 299-311.

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

Shepperd, J. A. (1991). Cautions in assessing spurious "moderator effects." *Psychological Bulletin*, *100*, 315-317.

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

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

Stocking, M. (1987). Two simulated feasibility studies in computerized adaptive testing. *Applied Psychology: An International Review*, *36*, 263-277.

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*, 1-6.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston: Allyn & Bacon.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 392-408.

Tellegen, A., Kamp, J., & Watson, D. (1982). Recognizing individual differences in predictive structure. *Psychological Review*, *89*, 95-l05.

Tibshirani, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association*, *83*, 394-405.

Tisak, J. (1994). Determination of the regression coefficients and their associated standard errors in hierarchical regression analysis. *Multivariate Behavioral Research*, *29*, 185-201.

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

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

Acknowledgement

The authors thank Keith Widaman, Scott E. Maxwell, and anonymous reviewers for their helpful comments on an earlier version of this paper. Portions of this research were presented at the 108th annual meeting of the American Psychological Association, Washington D.C., August 2000.

Author's Address

Authorship is listed in alphabetical order. Address correspondence to Sun-Mee Kang, Department of Psychology, California State University, Northridge, CA 91330; e-mail: sunmee.kang@csun.edu or Niels G. Waller, Department of Psychology and Human Development, Vanderbilt University, Nashville TN, 37203; e-mail: niels.waller@vanderbilt.edu.