Item Response Theory Models and Spurious Interaction Effects in Factorial ANOVA Designs

Susan E. Embretson

University of Kansas

In many psychological experiments, interaction effects in factorial analysis of variance (ANOVA) designs are often estimated using total scores derived from classical test theory. However, interaction effects can be reduced or eliminated by nonlinear monotonic transformations of a dependent variable. Although cross-over interactions cannot be eliminated by transformations, the meaningfulness of other interactions hinges on achieving a measurement scale level for which nonlinear transformations are inappropriate (i.e., at least interval scale level). Classical total test scores do not provide interval level measurement according to contemporary item response theory (IRT). Nevertheless, rarely are IRT models applied to achieve more optimal measurement properties and hence more meaningful interaction effects. This paper provides several conditions under which interaction effects that are estimated from classical total scores, rather than IRT trait scores, can be misleading. Using derived asymptotic expectations from an IRT model, interaction effects of zero on the IRT trait scale were often not estimated as zero from the total score scale. Further, when nonzero interactions were specified on the IRT trait scale, the estimated interaction effects were biased inward when estimated from the total score scale. Test difficulty level determined both the direction and the magnitude of the biased interaction effects. *Index terms: factorial designs, interaction effects, interval measurement, item response theory, level of measurement, measurement scales, statistical inference.*

Interaction effects are quite important in psychological research. A prototypic design for psychological experiments is crossing treatment conditions with a context variable (e.g., Siegler, 1981). The context variable may define groups that differ in either the particulars of the treatment administration (e.g., setting, duration, mode) or some individual status variable (e.g., demographics, prior experience, or standing on some theoretical construct). In many theoretical studies, the context by treatment interaction effect reflects the major research hypothesis. For example, in lifespan psychology, the interaction of training or task presentation conditions with age is interpreted as support for qualitative changes in development (e.g., Burke & Light, 1981; Salthouse & Mitchell, 1989). In applied settings, the relative impact of different programs on populations defined by gender, socioeconomic status, prior instruction, learning style, or learning attitudes often is a primary focus of research (e.g., Snow, 1988).

However, it is well known that the level of measurement for the dependent variable influences interaction effects (e.g., Tukey, 1949). Nonlinear transformations of the dependent variable can eliminate (or create) ordinal interactions in analysis of variance (ANOVA) factorial designs. Therefore, for example, the treatment effect that appears relatively larger in one group than in another, as defined by the context variable, may disappear by rescaling the data. Davison & Sharma (1990) demonstrated how a factorial ANOVA can provide a mixture of appropriate and inappropriate inferences about a latent variable when the statistics are applied to a monotonically related observed variable. In their analysis, the latent variable was assumed to be measurable at the interval scale level, but the observed variable was not. Davison and Sharma proved that interaction effects are directly influenced by the measurement scale for the observed variable, but that main effects are indirectly influenced.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 3, September 1996, pp. 201–212

0146-6216/96/030201-12$1.85

Unfortunately, the effect of measurement scales on ANOVA interactions is largely ignored in psychological research. Total test scores, calculated as unweighted sums across tasks or items, are often used as a dependent variable. Classical test theory (CTT) often is used to justify the scaling of such data. If, in fact, CTT yields interval scales, the results on interaction effects would be trustworthy because nonlinear transformations would be inappropriate.

However, in the view of contemporary test theory, classical total scores provide only ordinal level measurement (e.g., Fischer, 1995). As elaborated below, item response theory (IRT) models provide more theoretically justifiable scaling for several reasons. With the one-parameter logistic IRT model, total score is monotonically, but not linearly, related to latent trait level (θ). In more complex IRT models, total score is not even monotonically related to θ.

Although IRT models are commonplace in large-scale testing programs, rarely are they applied to substantively oriented psychological experiments that involve multiple group comparisons. However, two recent studies (Embretson, 1991, 1994; Maxwell & DeLaney, 1985) suggest the potential importance of IRT scaling in estimating and evaluating group differences. In these studies, the IRT model was assumed to be appropriate but, similar to typical psychological experiments, the statistical comparisons were based on total scores (or their linear transformations).

Maxwell & DeLaney (1985) demonstrated how group comparisons on an observed variable are influenced by test difficulty level. Assuming equal latent variable means in two groups, they showed that unequal group variances and inappropriate test difficulty levels lead to spurious mean differences on the observed variable.

Embretson (1991, 1994) conducted IRT simulation studies that demonstrated how inappropriate inferences result from the total score scale. Although no interaction of context and treatment existed on the θ scale, substantial interactions were estimated from the total score scale. Further, and significantly, the relative appropriateness of the test difficulty level for the various groups influenced both the magnitude of the interaction and the direction of the effect (i.e., for which context the treatments were most effective). However, no formal development of the test difficulty effect was presented in these studies.

Purpose

IRT models are appropriate for many dependent variables in psychological research. What is the inferential cost of not applying them when appropriate? This paper examines how estimates of interaction effects are influenced by using the total score scale, rather than using an appropriate IRT θ scale. A formal development is presented to relate the IRT model parameters to expectations for population parameters that are calculated from the total score scale. That is, from this development, total score means and variances (as required for estimating ANOVA interactions) may be anticipated for various hypothetical states (i.e., patterns of means across the groups) on the IRT θ scale. The expectations then are used to show how varying conditions on θ (i.e., magnitude of context and treatment main effects) and varying methodological conditions (test difficulty level and test length) influence the interaction effect that is estimated from total scores. Because interpreting the findings depends on the theoretical superiority of IRT scaling over the classical total score scale, the measurement properties of each scale are briefly reviewed.

Measurement Properties of CTT Versus IRT

CTT fails to achieve three desirable measurement properties that are achieved with IRT models: (1) item and person parameters that are invariant (within a linear transformation), (2) interval scale trait level measurement, and (3) measurement of items and persons on a common scale. A variety of textbooks show how the first property, invariant item and person parameters, is not met in CTT (see Hambleton, Swaminathan, & Rogers, 1991 for a clear demonstration). The CTT item difficulties and discriminations depend on the population that is tested, but the person's score depends on the properties of the items that are administered.

The second property, interval level measurement, is not achieved by CTT partly because the person parameters are not invariant across items. Item properties influence several aspects of score distributions, including skew, kurtosis, and the constancy of measurement error, as well as means and variances (Gulliksen, 1950, p. 365). Thus, because the relative distances between persons on classical total scores depend on item properties only, ordinal measurement is achieved, at best, in CTT. In contrast, the θ scale in IRT models does not depend on item difficulties because item properties are included within the model to adjust the θ estimates accordingly.

The third property, placing items and persons on a common scale, is not achieved in CTT because the observed score model (of true and error scores) does not include item parameters. Thus, in CTT item difficulties are not directly linked to trait level. In contrast, IRT models do place items and persons on a common scale because item difficulty and θ have additive effects on item response probabilities. That is, increasing the θ level has identical effects as decreasing item difficulty by the same amount.

Additive person and item parameters, particularly in the one-parameter logistic model (Rasch, 1960), are sometimes regarded as reaching the important property of additive conjoint effects in fundamental measurement (see Andrich, 1988). In fact, the fundamental measurement principle of additive decomposition as a numerical basis for score comparisons can be shown to yield all three desirable measurement properties—common scale measurement, interval level measurement, and invariant parameters (Embretson & DeBoeck, 1994).

Thus, several optimal measurement properties are not achieved in CTT. In contrast, IRT models, particularly the Rasch model, do achieve desirable measurement properties, including interval scale measurement. Furthermore, IRT models are more basic than CTT; the CTT indexes for a particular group may be derived from IRT model parameters (e.g., Lord & Novick, 1968). This property is used in the expectations below.

Expectations for Total Score Distributions From an IRT Model

To relate ANOVA group contrasts based on total scores to contrasts based on IRT θs, it is necessary to develop some asymptotic expectations for total score distributions from IRT model specifications. Then, given some specified hypothetical state on θ (i.e., group distributions, including a particular pattern of mean differences) and some methodological conditions, expectations for group contrasts based on total score can be given.

Lord (1980) presented derivations for CTT indexes as expectations from an IRT model. Because the IRT model is probabilistic, a distribution of total scores, $g(X|\theta)$, may be expected for each θ. The distribution of total scores at θ is given by a generalized binomial distribution, such that the probability of success can vary across items. Summing over r for the R possible response patterns that can yield total score X, the distribution of total scores at θ is

$$g(x|\theta) = \sum_{r=1}^{R} \prod_{i} P_i(\theta)^{x_{ir}} \left[1 - P_i(\theta)\right]^{1-x_{ir}}, \tag{1}$$

where $P_i(\theta)$ is the probability that a person at θ correctly answers item i as given from an IRT model and x_{ir} indicates whether item i is correctly or incorrectly answered in response pattern r.

Lord (1980) noted that the mean of Equation 1 for a person with θ_j is the expected total score, $\mathrm{E}(X_j)$. The mean of Equation 1 is given by the sum of the expectation of positive item responses:

$$\mathrm{E}\left(X_j\right) = \sum_{i} P_i\left(\theta_j\right). \tag{2}$$

The variance of total scores for person j with θ_j is

$$\sigma^2_{x|\theta} = \sum_i P_i(\theta)\left[1 - P_i(\theta)\right] . \tag{3}$$

To generalize the expectations to population statistics, as required for factorial ANOVA designs, a distribution for the latent variable $f(\theta)$ must be specified. An expectation for the mean, as in Equation 2, can be accomplished by integrating over θ:

$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} \sum_i P_i(\theta_j)\, f(\theta) d\theta . \tag{4}$$

Similarly, the variance expectation in Equation 3 also may be generalized to a population by integrating over θ:

$$\sigma^2_X = \int_{-\infty}^{\infty} \sum_i P_i(\theta_j)\left[1 - P_i(\theta_j)\right] f(\theta) d\theta + \int_{-\infty}^{\infty} \left[\sum_i P_i(\theta_j) - \mu_x\right]^2 f(\theta) d\theta . \tag{5}$$

The first term in Equation 5 reflects the error variance of total score at θ_j, and the second term reflects the deviations of the expected value of X at θ_j from the population mean.

To relate these expectations to some hypothetical true states for multiple group means on θ, a computer program was written. Given that $P_i(\theta_j)$ is obtained from an IRT model and, for convenience, that $f(\theta)$ is a normal distribution, the integrals in Equations 4 and 5 may be approximated by Gauss-Hermite quadrature. The program MULTIQUAD was written to compute item response probabilities for $P_i(\theta_j)$ and to evaluate the integrals in Equations 4 and 5. Note that Equations 4 and 5 may be evaluated for any distribution; thus, the derivation is not restricted to the normal distribution.

Implications for Interaction Contrasts in Factorial ANOVA Designs

Method

Contrast effects were estimated in a factorial design from the asymptotic total score means and variances calculated from Equations 4 and 5 by MULTIQUAD for various hypothetical states on the IRT θ scale. For all hypothetical states, the Rasch model was specified as the IRT model and θ was the Rasch person parameter. However, the hypothetical states varied in the pattern of group differences that were specified on θ. Furthermore, several levels of test difficulty and test length also were specified for each hypothetical state because the expectations in Equations 4 and 5 depend on the items in the test.

Design. Interaction contrasts were studied in a 2 (context) × 2 (condition) factorial ANOVA for fixed effects. The ANOVA design is shown in Table 1, where g is an indicator for group. Although for convenience only two levels of conditions are given (treatment and control) and two levels of context (low group and high group), comparable effects could be shown for larger designs.

Three orthogonal contrasts were specified for two main effects and an interaction. If λ_{kg} is the coefficient for group g (see values in Table 1) on contrast k, the context contrast is ψ_1 with coefficient vector $\lambda_1 = \{-.5, -.5, .5, .5\}$; the condition contrast is ψ_2 with coefficient vector $\lambda_2 = \{-.5, -.5, .5, .5\}$; and the interaction contrast is ψ_3 with coefficient vector $\lambda_3 = \{1, -1, -1, 1\}$.

Hypothetical state specifications. The hypothetical states for θ consisted of several combinations of

Table 1
Factorial ANOVA Design for Expectations

	Context	
Condition	Low Group	High Group
Control	μ_{11} ($g = 1$)	μ_{12} ($g = 2$)
Treatment	μ_{21} ($g = 3$)	μ_{22} ($g = 4$)

context, treatment, and interaction contrast values on the θ scale. θ was assumed normally distributed with a variance of 1.0 within each group. Four levels of the context contrast were specified on the θ scale: $\psi_1 = \{0.0, .5, 1.0, 1.5\}$. The marginal mean of the control condition for the treatment factor was specified as 0.0 (i.e., $\mu_{.1} = 0.0$). Thus, the pairs of control group means for low and high context, respectively, were (0.0,0.0), (–.25,.25), (–.5,.5), and (–.75,.75).

The treatment contrast values depended on the specified value of the interaction effect on the θ scale. For hypothetical states with no interaction effect, four levels were specified for the treatment contrast: $\psi_2 = \{0.0, .5, 1.0, 1.5\}$. For the hypothetical states with an interaction effect, the value of the treatment contrast depended on the interaction. Four levels were specified for the interaction contrast: $\psi_3 = \{-1, -.5, .5, 1\}$. The difference between treatment and control in the uneffected context was specified as 0.0; thus, the difference in the effected context equals the value of the interaction contrast. For negative values of ψ_3, given the contrast coefficients λ_3 above, the treatment is effective only in the low context. For positive values of ψ_3, the treatment is effective only in the high context.

Procedure. For all comparisons, expectations for total score were computed from Equations 4 and 5 for varying levels of test length and test difficulty with $P(\theta)$ given by the Rasch model (Rasch, 1960, p. 168, Eq. 1.1). Four test lengths were specified (10, 15, 30, and 60 items) to represent the various test lengths that are observed in experimental studies. Four levels of test difficulty were specified (–.75, 0.0, .75, 1.50) for an easy, moderate, difficult, and very difficult test, respectively.

Item difficulties within a test were specified as an inverse normal transformation of uniformly spaced probabilities. That is,

$$b_i = \Phi^{-1}\left[(i - .5)/n\right], \tag{6}$$

where

b_i is item difficulty,
n is the number of items, and
Φ^{-1} is the inverse normal transformation.

This transformation is similar to a conventional test design because an approximately normal test information function and a mean item difficulty of 0 is produced. The four levels of test difficulty were obtained by adding a constant to b_i.

An index of test inappropriateness was calculated to relate the results to test difficulty in these plots. Test inappropriateness was calculated by the difference between mean item difficulty and mean trait level across groups:

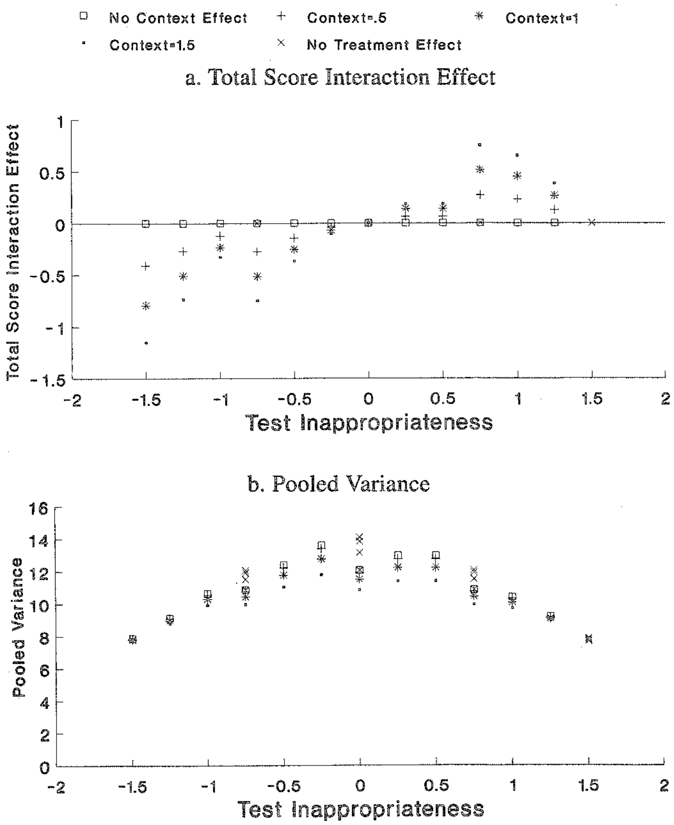
$$\mathrm{TI} = \frac{\sum_i b_i}{n} - \frac{\sum_j \mu_\theta}{G}. \tag{7}$$

Because both bs and θs were normally distributed, the difference between test difficulty level and the grand mean for θ reflects test inappropriateness. A negative value indicates that the test is relatively more appropriate for the low context group; a positive value indicates that the test is relatively more appropriate for the high context group.

Results

Interaction effects of zero. Figure 1 presents the results for hypothetical interactions of 0 for the 15-item test. In Figure 1a, the observed value of the interaction contrast is plotted by the test inappropriateness index. Because the θ interaction effect is 0 in these hypothetical states, any estimated interaction effect is erroneous. Figure 1a plots the observed values of the interaction contrast by test bias for No Context Effect, three levels of Context Effect, and No Treatment Effect. Figure 1a shows that the observed interaction was

Figure 1
Total Score Interaction Effect and Pooled Variance by Test Inappropriateness When IRT Interaction is Zero



either positive or negative, depending on the direction of the test inappropriateness index. That is, the treatment appeared more effective (positive interaction values) in the high context (context = .5, 1.0, or 1.5) if the test inappropriateness index was positive. Similarly, the treatment appeared more effective in the low context (i.e., negative interaction values) if the test inappropriateness index was negative. Only under the special conditions of either No Context Effect or No Treatment Effect, was the observed interaction contrast equal to 0.

Figure 1b shows the impact of test inappropriateness on the pooled variance, which, corrected for degrees of freedom, estimates the mean square error in factorial ANOVAs. Figure 1b shows that the more extreme the test inappropriateness index, the smaller the pooled variance. The results showed that with an inappropriate test, the variance was restricted for one or more groups, consequently lowering the pooled variance estimate.

An estimator of an effect is statistically biased if the expected difference between the parameter and the estimator is not 0. Strictly speaking, the results shown in Figures 1a and 1b are not relevant to bias because the estimators (from total scores) are derived from distributions with different variances than the parameter (the θ scale). To examine bias, standardized interaction contrast values were computed by dividing ψ_3 by the square root of the pooled variance. Because the contrast is standardized for variability, it provides an estimate of the interaction effect that can be compared to the θ scale that is fixed to a variance of 1.

Figure 2 shows the relationship between the standardized interaction contrast and the test inappropriateness index, plotted by level of treatment and context effect. The data in Figure 2 shows that the estimated interaction effect increases with the treatment effect. If the treatment effect was specified as 0 in the hypothetical true state, the observed interaction contrast was estimated as 0, regardless of test inappropriateness as shown in Figure 2a. However, when a treatment effect existed in the specified hypothetical state, increasing bias was observed, with the low groups producing negative bias and the high groups producing positive bias in Figures 2b, 2c, and 2d. The interaction contrasts were appropriately estimated as 0 only in two conditions: (1) the test inappropriateness level was 0, or (2) the hypothetical true context effect was 0.

The analyses were repeated at each test length. Test length had no influence on the relationship of the various design facets to the value of the interaction contrast. However, very slight differences were observed in the magnitudes of the standardized interaction contrast. Longer tests yielded slightly more extreme values for the interaction contrast.

Nonzero interaction effects. Test inappropriateness was calculated somewhat differently for the nonzero interaction effects than for the zero interaction effects. The means from the No Effect Context should not enter into the test inappropriateness index because equal treatment and control distributions were specified in the No Effect Context. However, for the contexts with effects, test difficulty level can influence the differences between treatment and control means. Thus, the test inappropriateness index was computed as test difficulty minus the marginal mean for the context groups with effects. The test inappropriateness index has positive values if the test is too difficult and negative values if the test is too easy for the contexts with effects.

Figure 3 shows that the observed interaction contrast had U-shaped relationships to test inappropriateness for all nonzero levels of the hypothetical interaction effect. The observed interaction contrast was highest when the test was appropriate for the groups with context effects (i.e., the test inappropriateness index equaled 0). Otherwise, the observed interaction contrast became less extreme as the test was more inappropriate.

The impact of test inappropriateness on the pooled variance was also estimated. The test inappropriateness index was defined in the same way as the 0 interaction case because all four groups enter into the pooled variance (and hence, the mean square error). Because the plot was nearly identical to Figure 1b, it is not presented here.

To assess the magnitude of bias, Figure 4 shows the relationship of the estimated interaction contrast (from the standardized interaction) to the hypothetical θ interaction across various hypothetical states. Figures 4a–4d differ in test difficulty level. Within each figure, the interaction effects are plotted by the level of the context effect. For reference, a perfect relationship between the estimated interaction and the hypothetical interaction also is shown. The estimated interaction effect was generally biased inward, yielding less extreme values than the hypothetical interaction effect. The relative level of inward bias depended on the specific test difficulty and the context effect.

For example, on the easy test (Figure 4a), when the treatment was more effective for the low context (i.e., negative hypothetical interactions), the standardized interaction contrast provided the least biased estimate for hypothetical states with larger context effects. When the treatment was effective for the high context (i.e., positive hypothetical interactions), the standardized interaction provided the least biased estimate for the sets with smaller context effects. This reversed impact for context at positive versus negative hypothetical interactions is readily explained by considering the role of test inappropriateness. When the treatment was effective only for the low context group, the easy test was most appropriate for low control group means, which occurred with a large context effect. When the treatment was effective only for the high context, the easy test was least inappropriate for low control means, which occurred with a small context effect.

Similar effects are shown in Figures 4b–4d. Interestingly, in the case of the very difficult test (Figure 4d), the hypothetical interaction was actually overestimated for hypothetical effects that favored the high context group. When a test is very difficult, the pooled variances will be small, especially for the low context. Hence, the standardization will result in overestimation.

Figure 2
Standardized Interaction Effect by Test Inappropriateness and IRT Treatment Effect When IRT Interaction is Zero

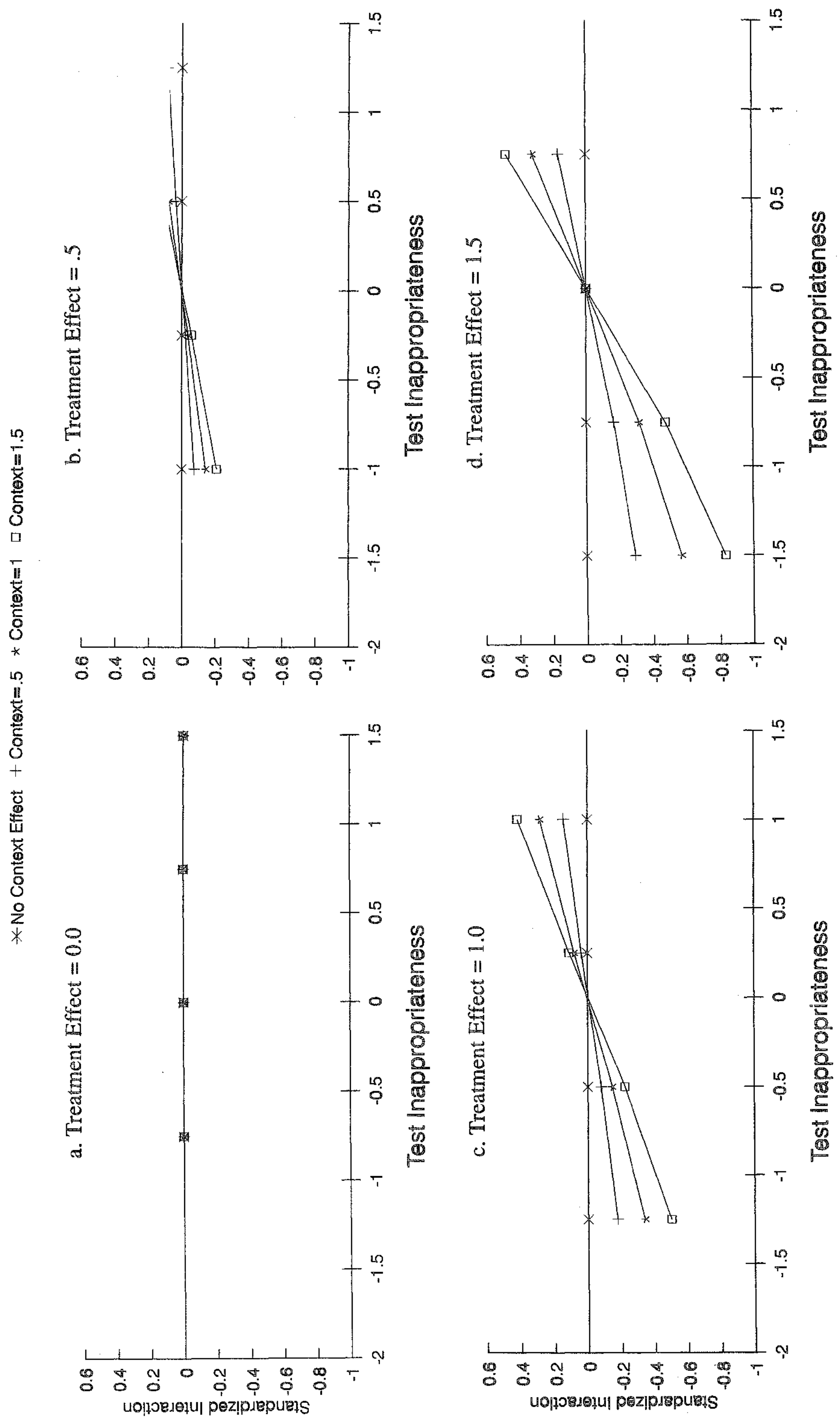
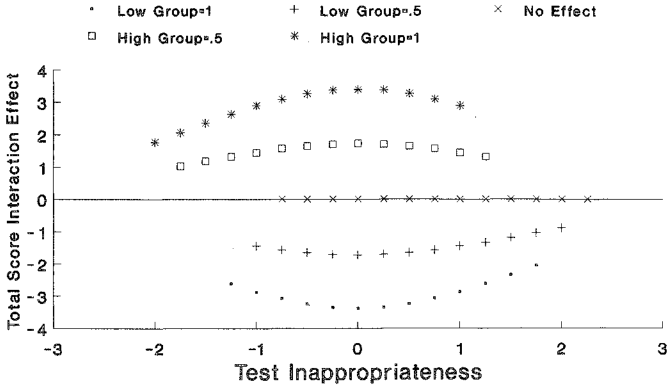


Figure 3
Total Score Interaction Contrast by Test Inappropriateness and Level of IRT Interaction



All analyses for the nonzero interactions were repeated on the 10-item, 30-item, and 60-item tests. The same pattern of results was obtained. To compare across test lengths, the mean value of the standardized interaction contrast was plotted across test lengths for each level of interaction in Figure 5. The results indicated that the standardized interaction contrast did not converge to the hypothetical value with increasing test length.

Discussion

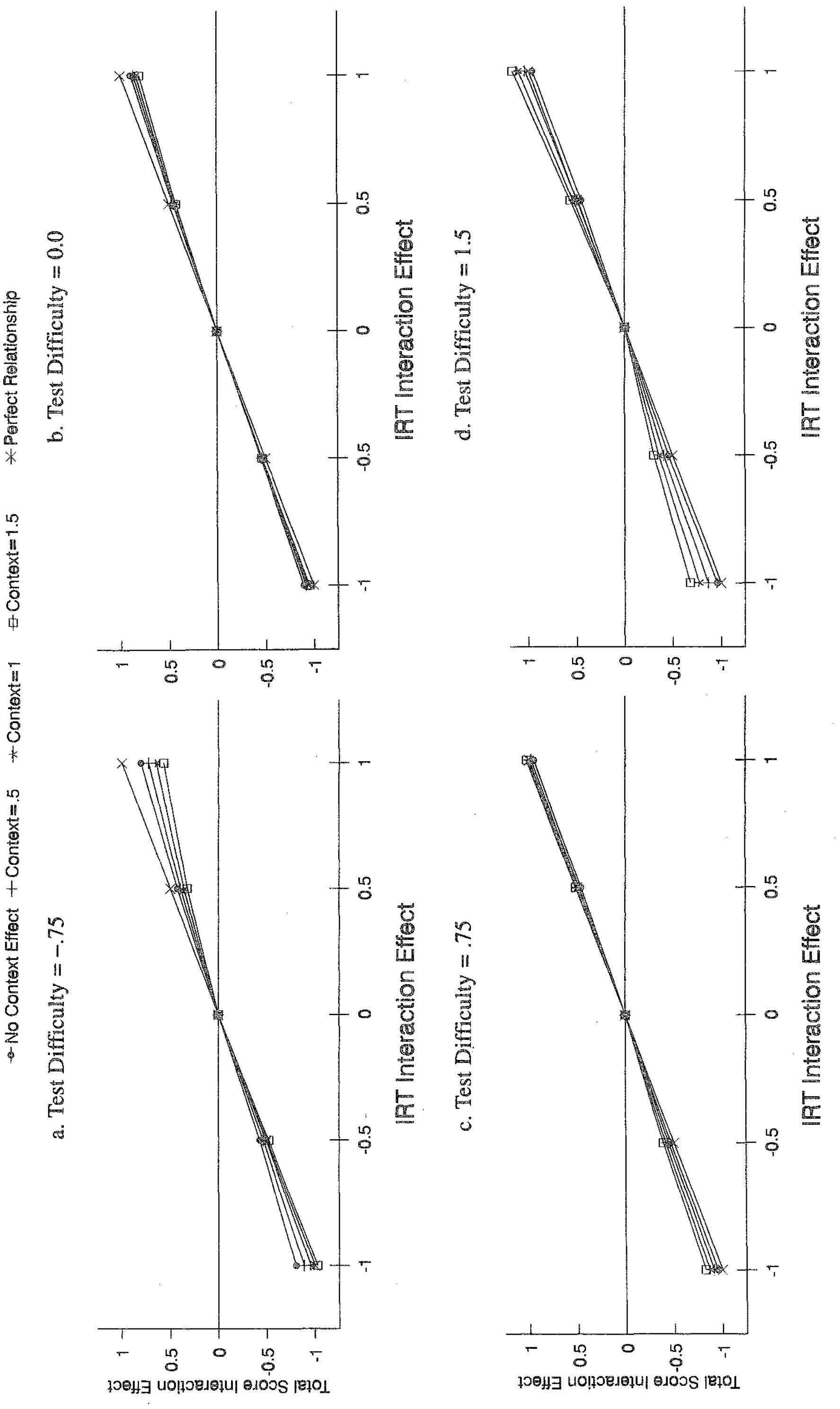
This paper was concerned with the inferential cost of applying a weak measurement model when a strong model is appropriate. In particular, the analyses investigated what happens if interaction contrasts are computed from classical total scores when an IRT model is theoretically and empirically appropriate for the data. The results suggest that when no interaction effects exist on the true latent variable, spurious interaction effects can be readily observed from the total score scale. Further, when interaction effects do exist, they will be underestimated.

In the current study, several hypothetical states were specified in which the latent variable was a parameter in a theoretically appropriate IRT model. When the interaction effect was specified as 0 on the θ scale, interaction contrasts computed from total scores yielded spurious results under most conditions. That is, the interaction contrast was appropriately estimated as 0 from total scores only under special conditions: (1) the context effect was 0 on the θ scale, (2) the treatment effect was 0 on the θ scale, or (3) test difficulty level was equally appropriate for both context groups. If test difficulty level favored either context group, spurious interaction effects were observed. Larger context or treatment effects yielded greater spurious effects. Further, both the direction and the magnitude of bias depended on test inappropriateness. The treatment appeared effective in the context for which the test was most appropriate even though no treatment effect existed on the IRT θ scale. The interaction contrast became increasingly biased as the test became increasingly inappropriate.

These results suggest that spurious interaction effects will be observed from the total score scale under many practical conditions in which a context effect exists. Although selecting optimally appropriate tests would appear to be a solution, usually this is not feasible in practice. Because the magnitude of both the context effect and the treatment effect influence test inappropriateness, it is generally impossible to anticipate the effects sufficiently well to select the most appropriate test difficulty level.

When the interactions were not 0 on the θ scale, test difficulty level influenced estimates of the interaction contrast effect when computed from total scores. The maximum value of the interaction contrast was ob-

Figure 4
Regression of Total Score Interaction Effects on IRT Interaction Effects by Test Difficulty Level



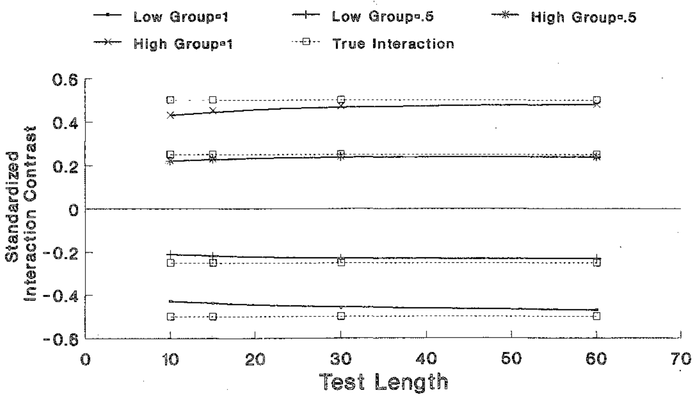
a. Test Difficulty = −.75

b. Test Difficulty = 0.0

c. Test Difficulty = .75

d. Test Difficulty = 1.5

Figure 5
Standardized Interaction Effect by Test Length for Levels of IRT Interaction Effects



tained if test difficulty level was optimal for detecting the effect. Otherwise, the absolute value of the interaction contrast decreased as test difficulty differed from the optimal level. Further, results on the standardized interaction contrast, which may be compared to the latent variable interaction, indicated increasing inward bias as the test difficulty level was increasingly inappropriate. In fact, inward bias was observed even if no context effect was specified.

Test length had little effect for either zero or nonzero interactions. The impact of the specified conditions on the observed interaction effect was nearly identical over different test lengths. Thus, longer tests will not correct for the spurious results that are obtained from the total score scale.

The current results are based only on the Rasch model. It is sometimes maintained that the Rasch model is too restrictive and does not fit real test data sufficiently well. However, even if a more complex IRT model is required to fit the data, the total score scale would not provide a relatively better metric. In fact, if item discrimination parameters are required to obtain fit, total score is not even monotonically related to the IRT θ parameters. The IRT trait score, even for equal total scores, would depend on which items were answered correctly.

Failing to apply an appropriate measurement model can be costly. Although substantive researchers rarely consider the issue of measurement models in interpreting results from factorial designs, IRT models are now widely available, theoretically justifiable, and empirically appropriate for many measurement situations. Perhaps one obstacle to applying IRT models, when they are appropriate, has been that the advantages resulting from the additional effort are unclear. The current results imply that assuring inferences to a latent variable is a clear advantage to applying IRT models to substantive research.

References

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

Burke, D. M., & Light, L. L. (1981). Memory and aging: The role of retrieval processes. *Psychological Bulletin, 90,* 513–546.

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

Embretson, S. E. (1994, August). *Who changes the most? Some non-ignorable artifacts in measuring individual differences.* Presidential address to the Division of Evaluation, Measurement, and Statistics of the American Psychological Association, San Francisco.

Embretson, S. E. (1994). Comparing changes between groups: Some perplexities arising from psychometrics. In D. Laveault & B. Zumbo (Eds.), *Modern theories in measurement: Issues and practice* (pp. 214–248). Ottawa, Canada: Social Science and Humanities Research Council of Canada.

Embretson, S. E., & DeBoeck, P. (1994). Latent trait theory. In R. J. Sternberg (Ed.), *Encyclopedia of intelligence* (pp. 644–647). New York: MacMillan.

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

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park CA: Sage.

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

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

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

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

Salthouse, T. A., & Mitchell, D. R. D. (1989). Structural and operational capacities in integrative spatial ability. *Psychology of Aging, 4,* 18–25.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46,* No. 2.

Snow, R. E. (1988). Cognitive-conative aptitude interactions in learning. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation and methodology: The Minnesota symposium on learning and individual differences* (pp. 385–403). Hillsdale NJ: Erlbaum.

Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics, 5,* 232–242.

Author's Address

Send requests for reprints or further information to Susan Embretson, The University of Kansas, Department of Psychology, 426 Fraser Hall, Lawrence KS 66045-2160, U.S.A.