Psychological Bulletin
1993, Vol. 114, No. 3, 552-566

0033-2909/93/$3.00

Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance

Steven P. Reise, Keith F. Widaman, and Robin H. Pugh

This study investigated the utility of confirmatory factor analysis (CFA) and item response theory (IRT) models for testing the comparability of psychological measurements. Both procedures were used to investigate whether mood ratings collected in Minnesota and China were comparable. Several issues were addressed. The first issue was that of establishing a common measurement scale across groups, which involves full or partial measurement invariance of trait indicators. It is shown that using CFA or IRT models, test items that function differentially as trait indicators across groups need not interfere with comparing examinees on the same trait dimension. Second, the issue of model fit was addressed. It is proposed that person-fit statistics be used to judge the practical fit of IRT models. Finally, topics for future research are suggested.

Much research and debate has been motivated by the question of how to establish that a test measures the same trait dimension, in the same way, when administered to two or more qualitatively distinct groups (e.g., men and women). The question can also be posed as follows: Are test scores for individuals who belong to different examinee populations comparable on the same measurement scale? The objectives of this study were to review linear confirmatory factor analysis[1] (CFA; Long, 1983) and item response theory (IRT; Lord, 1980) approaches to addressing this important question and to suggest, by way of real-data application, advantages and disadvantages of each approach.

Measurement Invariance

To compare groups of individuals with regard to their level on a trait, or to investigate whether trait-level scores have differential correlates across groups, one must assume that the numerical values under consideration are on the same measurement scale (Drasgow, 1984, 1987). That is, one must assume that the test has "measurement invariance" across groups. If trait scores are not comparable (i.e., on the same measurement scale) across groups, then differences between groups in mean levels or in the pattern of correlations of the test with external variables are potentially artifactual and may be substantively misleading.

Because establishing measurement invariance of a test across distinct groups is critical to progress in many domains of psychology, much discussion has been devoted to this topic (e.g., Byrne & Shavelson, 1987; Drasgow & Kanfer, 1985; Frederiksen, 1987; Hui & Triandis, 1985; Linn & Harnisch, 1981; van der Flier & Drenth, 1980). One central principle, evident throughout this literature, is that psychological measurements are on the same scale (i.e., comparable) when the empirical relations between the trait indicators (e.g., test items) and the trait of interest are invariant across groups.

As Windle, Iwawaki, and Lerner (1988, p. 551) explained, the primary approach to addressing measurement invariance "involves the study of similarities and differences in the covariation patterns of item-factor relations." Prior to the 1970s, various heuristic strategies for checking the invariance between two or more factor structures were proposed (Reynolds & Har-

Steven P. Reise, Keith F. Widaman, and Robin H. Pugh, Department of Psychology, University of California at Riverside.

Steven P. Reise and Keith F. Widaman contributed equally to the article and were simply listed in alphabetical order. The present work was supported in part by intramural grants from the Academic Senate, University of California at Riverside, to Steven P. Reise and Keith F. Widaman; by Grants HD-21056 and HD-22953 from the National Institute of Child Health and Human Development to Keith F. Widaman; and by Grants G0085300208 and H023C80072 from the U.S. Office of Education (Donald MacMillan, principal investigator).

We would like to acknowledge the assistance of Jiayuan Yu, who developed the translated version of the items and collected data on the Chinese subjects. The most helpful comments of Roderick McDonald, Robert MacCallum, Bengt Muthén, Richard Wagner, and the three anonymous reviewers of a previous version of this article are gratefully acknowledged.

Correspondence concerning this article should be addressed to Steven P. Reise, Department of Psychology, University of California, Riverside, California 92521.

[1] We limit our attention to traditional linear factor analysis. Because of this restriction in focus, we do not consider in detail certain work on relations between factor analysis and IRT models. For example, for 30 years McDonald (e.g., 1962, 1967, 1982, in press; Etezadi-Amoli & McDonald, 1983) has discussed the relations between IRT (or latent trait approaches) and nonlinear factor analysis, specifically as alternate parameterizations of each another. Also, Muthén (1984, 1988) introduced IRT-like threshold estimates into confirmatory structural models and Takane and de Leeuw (1987) recently proved the equivalence of IRT and factor analysis of discretized (e.g., dichotomous) variables. However, we chose to compare linear factor analysis and IRT because they are the most commonly encountered techniques in substantive applications and are most widely available as distributed programs. Extending the comparisons among methods to more recent developments, such as those discussed by McDonald, Muthén, and Takane and de Leeuw, is beyond the scope of this article but would be a worthwhile topic for future investigations.

Table 1
Item Means, Standard Deviations, and Intercorrelations for the Minnesota and Nanjing Samples

	Minnesota		Nanjing						
Item	*M*	*SD*	*M*	*SD*	1	2	3	4	5
1. Nervous	2.17	1.12	1.89	0.93	—	.79	.66	.73	.66
2. Worried	2.52	1.22	2.09	1.11	.29	—	.55	.74	.84
3. Jittery	2.01	1.09	1.60	1.01	.30	.25	—	.63	.57
4. Tense	2.35	1.19	2.15	1.04	.39	.31	.20	—	.92
5. Distressed	2.29	1.25	1.93	1.12	.40	.51	.42	.52	—

Note. Correlations among items for the Minnesota sample are listed above the diagonal; correlations among items for the Nanjing sample are listed below the diagonal.

ding, 1983). Although discussions of these techniques still appear in the psychometric literature, interest in them has declined substantially with the advent of CFA (Jöreskog, 1971) and IRT (Lord, 1980) modeling. The theory underlying, and applications of, CFA and IRT have been discussed in the psychometric literature, and these techniques are widely used in large-scale achievement testing programs. Nevertheless, these newer models, especially IRT, have failed to find frequent application in the context of more typical, substantive research problems for which their use might be quite illuminating.

To remedy this situation, at least in part, we demonstrate the application of CFA and IRT models to real data. Our nominal objective was to investigate the measurement invariance of mood adjective ratings gathered from American and Chinese subjects. Our practical goals were (a) to focus attention on similarities and differences between CFA and IRT modeling, (b) to provide conceptual clarity regarding key aspects of the investigation of measurement invariance, and (c) to identify topics requiring further research.

General Method

Subjects

Item response data were collected from two distinct groups. The first sample consisted of 540 undergraduates attending the University of Minnesota. The second sample contained 598 undergraduates attending the University of Nanjing Normal in China.

Measure

As part of a larger project, each subject rated his or her current mood on a five-item measure of negative affect. This measure, called NA5, is the basis of all analyses reported in this article. NA5 consists of the adjectives *nervous, worried, jittery, tense,* and *distressed,* to which responses are obtained on a Likert-format rating scale ranging from *not at all* (1) to *extremely* (5).

Chinese equivalents of the English terms were created by a back-translation method (Brislin, 1970). The NA5 item means, standard deviations, and item intercorrelations for both samples are provided in Table 1. For the raw scale scores, coefficient alphas were .84 and .71 in the Minnesota and Nanjing samples, respectively. The mean raw scores on the NA5 scale were 11.3 (SD = 4.6) and 9.6 (SD = 3.5) for the Minnesota and Nanjing samples, respectively.

From the raw score statistics, it appears that the Minnesota sample is higher and more variable on the negative affect trait dimension relative to the Nanjing sample; such speculations are, however, premature. To compare groups on this psychological dimension, one must be assured that the trait scores are on a common measurement scale. As a result, one must work at the level of the latent variable presumed to underlie and cause variation in the observed item responses. We now turn to this task.

The Linear Confirmatory Factor Analysis Approach

The Linear Confirmatory Factor Model

Application of CFA for testing measurement invariance originated in the early 1970s (Jöreskog, 1971; McGaw & Jöreskog, 1971). In the typical CFA model, each measured variable X_m, where $m = 1, \ldots, n$, is represented as a linear function of one particular latent variable, ξ_p, where $p = 1, \ldots, r$, and a stochastic error term, δ_m. This relationship may be represented as

$$\mathbf{X}_m = \lambda_{mp}\xi_p + \delta_m, \tag{1}$$

where λ_{mp} is the regression coefficient representing the regression of X_m on ξ_p and other terms are as just defined. Assuming the presence of n measured variables and r latent variables and concatenating the parameters in Equation 1 into matrices leads to the following:

$$\mathbf{X} = \Lambda\xi + \delta, \tag{2}$$

where $\mathbf{X}$ is a $(n \times 1)$ column vector of scores of person i on n measured variables, Λ is a $(n \times r)$ matrix of loadings of the n measured variables on the r latent variables, ξ is a $(r \times 1)$ matrix of factor scores of person i on the r latent ξ variables, and δ is a $(n \times 1)$ matrix of measurement residuals. It is possible to show that Equation 2 implies the following equation:

$$\Sigma = \Lambda\Phi\Lambda' + \Psi, \tag{3}$$

where Σ is the $(n \times n)$ population covariance matrix among the measured variables in Equation 2, Φ is a $(r \times r)$ matrix of covariances among the latent variables, Ψ is a $(n \times n)$ matrix of covariances among the measurement residuals or unique factors, and Λ is as just defined.

The model in Equation 3 can be fit to a sample covariance matrix S from a sample of size N, leading to

$$S \simeq \hat{\Lambda}\hat{\Phi}\hat{\Lambda}' + \hat{\Psi} = \hat{\Sigma}, \tag{4}$$

where S is the $(n \times n)$ observed sample covariance matrix among measured variables and the $\hat{\Lambda}$, $\hat{\Phi}$, $\hat{\Psi}$, and $\hat{\Sigma}$ matrices contain sample estimates of the population parameters in the corresponding matrices in Equation 3. As shown in Equation 4, the observed covariances among the n measured variables in S are approximated by the linear CFA solution $\hat{\Lambda}\,\hat{\Phi}\,\hat{\Lambda}' + \hat{\Psi}$; this solution, in turn, produces $\hat{\Sigma}$, which contains estimates of the population covariances among the measured variables, Σ, under the assumption that the stated model is a proper representation of the data and therefore holds in the population.

For multiple-group linear CFA modeling, Equation 4 may be modified to denote group membership as

$$S_g \simeq \hat{\Lambda}_g \hat{\Phi}_g \hat{\Lambda}_g' + \hat{\Psi}_g = \hat{\Sigma}_g, \tag{5}$$

where all matrices are as defined earlier, except for the addition of the g subscript to denote that the matrices were derived from the gth sample.

The models in Equations 1–5 are covariance structure models meant to apply only to analyses of covariance matrices. Among others, Cudeck (1989) discussed the ways in which application of covariance structure models to correlation matrices, even in one-sample analyses, may lead to inaccurate results. In multiple-group modeling, additional important issues arise; as a result, analyses must be performed on within-group covariance matrices rather than within-group correlation matrices. The reason that within-group correlation matrices are inappropriate for multiple-group analyses is that, to investigate the invariance across groups in various model parameters, such as the regression weights (i.e., the λ_{mp} estimates) relating latent variables to measured variables, the scores on the measured variables must be on the same scale across groups. Standardizing the data separately for each of the g groups (e.g., to a z-score metric, with unit variance for each variable within each group) would lead to different rescalings of measured variables within each group, destroying the comparability across groups of the common scale for the measured variables and leading to an inability to compare parameter estimates across groups. Further details on these and other more technical matters are discussed in many standard references, such as Jöreskog (1971) and Jöreskog and Sörbom (1989).

In this study, the number of measured variables (i.e., items) in each group was five, and we assumed that a single common factor was being measured in each group. Hence, for both the Minnesota and Nanjing samples, S was a (5×5) covariance matrix among the items, $\hat{\Lambda}$ was a (5×1) matrix of loadings of the five items on a single negative affect factor, $\hat{\Phi}$ was a (1×1) covariance matrix for the common factors (i.e., the variance of ξ), and $\hat{\Psi}$ was a (5×5) matrix of unique factor covariances. Unique factors were assumed to be uncorrelated; thus, all off-diagonal elements in $\hat{\Psi}$ were fixed at zero.

Equation 4 suggests the principle by which estimated CFA models can be evaluated: Latent factor models imply particular covariance matrices. The statistical acceptability of an estimated CFA model depends on how close the estimated covariance matrix $\hat{\Sigma}$ is to the observed covariance matrix S.

Assessing the Fit of a CFA Model

There are two typical ways of judging the adequacy of an estimated CFA model. First, using certain methods of estimation (e.g., maximum likelihood), CFA programs, such as LISREL, provide a likelihood ratio chi-square statistic to test whether the covariance matrix reproduced from the estimated parameters, $\hat{\Sigma}$, differs significantly from the observed sample covariance matrix S. In the multiple-groups CFA context, a single chi-square value assessing aggregate fit across the $\hat{\Sigma}_g$ and S_g matrices for the multiple groups is obtained.

The likelihood ratio chi-square statistic appears to be overly sensitive to trivial discrepancies between $\hat{\Sigma}_g$ and S_g if the sample size is large (Bentler & Bonett, 1980). Hence, so-called "practical" indices of fit (Bentler & Bonett, 1980; Marsh, Balla, & McDonald, 1988) are often used to evaluate CFA models. Although the relative merits of practical fit indices are much debated, it is safe to follow two principles. First, it is useful to calculate two or more indices of practical fit when evaluating a model. Second, no CFA model should be accepted or rejected on statistical grounds alone; theory, judgment, and persuasive argument should play a key role in defending the adequacy of any estimated CFA model.

To assess fit, we used the likelihood ratio chi-square statistical index and three practical fit indices: (a) the Tucker-Lewis Index (TLI; Tucker & Lewis, 1973), which was also termed the nonnormed fit index (NNFI) by Bentler and Bonett (1980); (b) a noncentrality index (NI) derived independently by Bentler (1990) and by McDonald and Marsh (1990); and (c) the root mean square error of approximation (RMSEA), proposed by Steiger and Lind (1980). The TLI was found by Marsh et al. (1988) to be among the best of the then-available indices of practical fit, and the NI has performed somewhat better than the TLI in recent simulations (Bentler, 1990), especially in small samples. These two practical fit statistics are relative fit indices (cf. Bentler, 1990; McDonald & Marsh, 1990) indicating roughly the proportion of covariation among indicators explained by the model relative to a null model of independence in the indicators. Values near 0.0 indicate poor fit, whereas values near 1.0 indicate good fit; practical fit index values greater than .90 are usually considered satisfactory. In contrast, the RMSEA is an absolute fit measure assessing badness of fit of a model per degree of freedom in the model. The lower bound of the RMSEA is zero, a value obtained only if a model fits a set of data perfectly. Browne (1990) stated that RMSEA values of about .05 indicate close fit of a model to data and that values of about .08 reflect reasonable fit of a model.

Specifying Measurement Invariance

The test of measurement invariance across groups in CFA models is equivalent to the test of whether the factor loading matrix, $\hat{\Lambda}_g$, in Equation 5 is invariant across groups (Alwin & Jackson, 1981; Sörbom, 1974). That is, within the context of this study, the mood adjectives must relate to the single latent trait in the same way for the Minnesota and the Nanjing samples. The hypothesis of full measurement invariance for two groups can be expressed formally as H_0: $\Lambda_1 = \Lambda_2$, assuming that this model holds exactly in the population. No between-group equality restrictions are placed on the diagonal elements of the $\hat{\Phi}_g$ or $\hat{\Psi}_g$ matrix, because groups are likely to differ with respect to their variances on the latent factor and on the unique factors. Among others, MacCallum and Tucker (1991) argued that fac-

tor loadings should, in theory, be invariant over samples from a given population, whereas factor variances and covariances are sample specific. The hypothesis of full measurement invariance can be tested with the multiple-group CFA estimation routines provided in LISREL (Jöreskog & Sörbom, 1989).

The Baseline Model in CFA

The first step in a multiple-group CFA analysis is to compute the covariances among the observed variables, S_g, for each group. Then, because we assumed that a single latent trait may have accounted for the observed item covariances, we freely estimated a one-factor model for each S_g matrix. That is, we freely estimated values in the $\hat{\Lambda}_g$, $\hat{\Phi}_g$, and $\hat{\Psi}_g$ matrices for each S_g matrix simultaneously. This freely estimated model, which may be called the "baseline model," serves as a benchmark against which the fit of more restricted models is compared.

In any CFA model, an indeterminacy exists between the scale of the item parameters (the λ_{mp}s relating the latent variable to the measured variables) and the scale, or variance, of the latent factor, ξ. That is, the values of the factor loadings depend on the scale of the latent factor. If the scale for the item parameters is to be identified, the scale for the latent variable must be specified, or vice versa. This metric identification problem is typically resolved by fixing the value of one factor loading λ_{mp} to a constant (usually 1.0; cf. Jöreskog & Sörbom, 1989) or, less common, by fixing the variance of the latent variable (the diagonal of Φ) to a constant (usually 1.0).

For illustrative purposes, we estimated the baseline model in three distinct ways in this study. In the first version (a model termed Baseline 1), λ_{11}, the factor loading for the item *nervous* was fixed equal to 1.0 in both groups. This specification was used to freely estimate Φ_{11}, the variance of ξ, in each group. In the second version of the baseline model, termed Baseline 2, the factor variance (Φ_{11}) was fixed to 1.0 in both groups. In the Baseline 3 model, the factor variance (Φ_{11}) was fixed at 1.0 in the first group, the corresponding parameter (Φ_{11}) was estimated in the second group, and the first factor loading (λ_{11}) was constrained to equality across groups. In the Baseline 3 model, the constraint on the first factor loading was sufficient to identify all remaining parameter estimates in the second group, given the fixing of the factor variance, Φ_{11}, in the first group.

All three of the baseline models have identical levels of statistical and practical fit because they are simple respecifications of one another. However, the interpretations of model parameters differ across the models. In Baseline 1, the scale of ξ (the latent factor) within each group is defined by the item *nervous*. If this item had previously been shown to provide equivalent measurement across groups, then the remaining λ_{mp} estimates in the Baseline 1 model could be meaningfully compared across groups. In Baseline 2, the scale of ξ is defined within each group by specifying its variance. This does not imply, however, that the scale for ξ is comparable across groups or that the factor loadings, λ_{mp}, are easily or directly comparable across groups. In fact, the factor loading estimates, λ_{mp}, may be invariant across groups in the population; however, if the groups differ markedly in variance on the latent variable, then the factor loading estimates may appear (incorrectly) to vary markedly across groups as a result of the different rescalings of the latent variable within each group. In the Baseline 3 model, the variance of the latent variable is fixed at unity in one sample. The constraint on the first factor loading will then lead to comparability of factor loadings across groups if the item whose factor loading is constrained to invariance across groups has been shown previously to provide equivalent measurement across groups; the factor variance in the second group is estimated relative to the unit variance in the first group.

Parameter estimates and fit statistics for all three versions of the baseline model are shown in Table 2. Because the chi-square value for each model is more than seven times larger than the degrees of freedom, none of the baseline models adequately explains the observed data on statistical grounds ($p < .0001$). In terms of practical fit, however, the freely estimated one-factor baseline models are adequate for the current data (TLI = .915, NI = .958, and RMSEA = .076). The TLI and the NI reflect fit relative to the null model, which had $\chi^2(20, N = 1{,}138) = 1{,}551.74$, $p < .0001$. These fit indices represent an encouraging result because a well-fitting and theoretically viable baseline model should be established before further invariance analyses are conducted.

Confirming our earlier observations, the estimated factor loadings from the Baseline 1 model, shown in the second and third columns of Table 2, appear quite similar across the two groups; the possible exception is λ_{51}, which had values of 1.10 and 1.57 for the Minnesota and Nanjing samples, respectively. The relative comparability of factor loadings occurred in the context of rather large differences in variance on the latent variable (i.e., the Φ_{11} estimates), with latent variable variances of .68 and .29 for the Minnesota and Nanjing samples, respectively. In contrast, when the latent variable variance is fixed at unity in each group in the Baseline 2 model (see columns 4 and 5 of Table 2), all estimated factor loadings appear to differ considerably across the two samples. Finally, in the Baseline 3 model, the first factor loading was constrained to equality across groups. This restored the across-group comparability of the remaining factor loadings, which appear rather similar across groups; the exception is λ_{51}, which differed considerably across groups, as in the Baseline 1 model. The freely estimated variance in the Nanjing sample was .42, which indicates that the variability on the latent variable in this sample was considerably less than that for the Minnesota sample, the variance for which had been fixed at unity to identify the model.

Testing for Full Measurement Invariance in CFA Models

With fit values from the freely estimated one-factor model established, the hypothesis of full measurement invariance, embodied in the test that $\Lambda_1 = \Lambda_2$, could then be evaluated. We tested this hypothesis using a variant of Baseline 3; here, we ran LISREL exactly as in the Baseline 3 model but with the constraint that all five λ_{mp} parameters were invariant across groups. The resultant chi-square value from this restricted model may be compared with the respective value for the Baseline 3 model because the full measurement invariance model is nested within the Baseline 3 model (cf. Bentler & Bonett, 1980, on nested models). The full measurement invariance model is nested within the Baseline 3 model because one may arrive at the full measurement invariance model simply by applying con-

Table 2
Estimated Parameter Matrices and Fit Statistics for the Baseline, Full Measurement Invariance, and Partial Measurement Invariance Confirmatory Factor Analysis Models

	Baseline 1		Baseline 2		Baseline 3		Full invariance		Partial invariance	
Parameter	Minnesota	Nanjing	Minnesota	Nanjing	Minnesota	Nanjing	Minnesota	Nanjing	Minnesota	Nanjing
λ_{11}	1.00	1.00[a]	0.83[a]	0.54	0.83		0.81		0.82	
λ_{21}	1.07	1.06	0.89	0.57	0.89	0.88	0.86		0.88	
λ_{31}	0.83	0.88	0.68	0.47	0.69	0.73	0.67		0.69	
λ_{41}	1.12	1.14	0.92	0.61	0.92	0.94	0.91		0.93	
λ_{51}	1.10	1.57	0.91	0.84	0.91	1.30	0.98		0.91	1.28
ϕ_{11}	0.68	0.29	1.00[a]	1.00[a]	1.00[a]	0.42	1.00[a]	0.52	1.00[a]	0.43
ψ_{11}	0.58	0.58	0.58	0.58	0.58	0.58	0.60	0.55	0.58	0.58
ψ_{22}	0.70	0.91	0.70	0.91	0.70	0.91	0.71	0.90	0.70	0.91
ψ_{33}	0.73	0.81	0.73	0.81	0.73	0.81	0.74	0.81	0.73	0.81
ψ_{44}	0.57	0.72	0.57	0.72	0.57	0.72	0.57	0.69	0.57	0.72
ψ_{55}	0.73	0.54	0.73	0.54	0.73	0.54	0.70	0.65	0.73	0.54
χ^2	74.84		74.84		74.84		90.03		75.15	
df	10		10		10		14		13	
χ^2 change	—		—		—		15.19		0.31	
df change	—		—		—		4		3	
Tucker-Lewis index	.915		.915		.915		.929		.938	
Noncentrality index	.958		.958		.958		.950		.959	
Root mean square error of approximation	.076		.076		.076		.069		.065	

Note. Parameter estimates that are centered between the Minnesota and Nanjing columns (e.g., the 0.83 loading for λ_{11} in the Baseline 3 model) represent parameters constrained to equality across samples.
[a] Parameters fixed at tabled values to identify each model.

straints on parameters in the Baseline 3 model; no new parameter estimates are introduced in the full measurement invariance model that were not present in the Baseline 3 model. The difference in chi-square values for two nested models is itself distributed as a chi-square value with degrees of freedom equal to the difference in degrees of freedom for the two models. If the restricted, nested model results in a nonsignificant increase in chi-square over that for the less restricted model, then the hypothesis of full measurement invariance is tenable.

We chose to modify the Baseline 3 model to form the full measurement invariance model because the resulting estimates would be in an easily interpreted metric. It is clear from Table 2 that one may arrive at the full measurement invariance model by constraining the remaining four factor loadings in the Baseline 3 model, λ_{21} through λ_{51}, to equality across groups. These four additional constraints account for the difference of four degrees of freedom between the baseline and full measurement invariance models. Similar additional across-group constraints on factor loadings λ_{21} through λ_{51} in the Baseline 1 model would also have achieved a version of the full invariance model but would have resulted in latent variable variances in both groups that differed markedly from unity, leading to estimated variances that would be more difficult to interpret. Finally, it is more difficult to demonstrate that the full invariance model is nested within the Baseline 2 model. However, because all three baseline models are respecifications of one another, the full measurement invariance model is nested within the Baseline 2 model as well.

As shown in Table 2, constraining the Λ matrix to invariance across groups led to a statistically significant decrease in model fit. Moving from the Baseline 3 model to the full invariance model resulted in a significant change in the statistical index of fit, $\chi^2(4, N = 1{,}138) = 15.19, p < .001$. The practical indices of fit provided a more mixed message, with the TLI and RMSEA attaining somewhat improved values and the NI a marginally worse value. Given the statistical test results, we rejected the full invariance hypothesis; all items are not related to the trait in the same way across the two groups. If a common set of λ_{mp} parameters were used to estimate factor scores within each group, these estimates might be biased and potentially misleading indicators of individual and group differences.

Partial Measurement Invariance in CFA Models

Many researchers assume that if the full measurement invariance hypothesis of $\Lambda_1 = \Lambda_2$ is rejected, comparison of groups on ξ is not possible with these particular items. This assumption is incorrect, however, because the only requirement to compare groups on a latent variable is that partial measurement invariance be established (see Byrne, Shavelson, & Muthén, 1989). Partial measurement invariance occurs if some, but not all, of the nonfixed values in Λ are invariant across groups and if these invariant loadings define the latent metric. To ensure the nonarbitrariness of the across-group comparisons, a majority of the items on a given latent variable should have loadings that are invariant across groups.

If the full invariance hypothesis is rejected, as it was here, then further analyses are required to identify whether a subset of items is invariant across groups. The search for a subset of invariant items is facilitated by LISREL modification indices (MIs). One MI value is computed for each fixed or constrained parameter in a given LISREL model. MI values indicate how

much the overall chi-square value would change if the constraint were lifted from the parameter, and each MI value is associated with one degree of freedom.

After fitting and rejecting the full invariance model, MI values for the Λ matrix may be examined. If less than half of the items have significant MI values in the Λ matrix, then partial measurement invariance may hold. In this study, the only statistically significant MI values occurred for parameter λ_{51}, which had an MI of 13.97 in each sample. These MI values suggest that the factor loading for the item *distressed* cannot be equated across samples, as it was under the assumption of full measurement invariance.

Testing for partial invariance required performing another LISREL analysis. The specifications were the same as in the full invariance model, except that the factor loading parameters associated with statistically significant MI values were freely estimated for each group. Fit values from this model were then compared, in the usual way, with the baseline model because the partial invariance model is nested within the baseline model. The parameter estimates and fit values for the model with λ_{51} freely estimated within each group are shown in the last two columns of Table 2. Clearly, this model did not differ significantly from the Baseline 3 model, either statistically, $\chi^2(3, N = 1{,}138) = 0.31$, *ns*, or practically, because all three indices of practical fit were improved over comparable values for the Baseline 3 model. All estimated parameters in the partial measurement invariance model were statistically significant ($p < .0001$); standard errors ranged from 0.041 to 0.100, and all associated z values were greater than 9.3. The estimated variance on the latent variable for the Nanjing sample was 0.43 ($SE = 0.053$), more than 10 standard errors from unity (the value for the Minnesota sample), indicating that the Nanjing sample exhibited considerably less variability on the latent variable.

We concluded, therefore, that the items *nervous, worried, jittery,* and *tense* provide equivalent measurement across groups, whereas the item *distressed* does not. Individual differences can then be scaled (i.e., factor scores could be estimated) with common weights for the invariant items but different weights, depending on group membership, for the item *distressed.* The resulting factor scores will be on a common scale and will be comparable. This illustrates that even if an item has a different relationship to the latent variable across populations, individuals can be assessed on a common measurement scale.

Estimating Mean Differences on Latent Variables Within CFA Models

The partially invariant model established earlier can be used to estimate population group mean differences on ξ by performing a mean structures analysis with LISREL, as explained in Byrne et al. (1989), Everitt (1984), and Muthén and Christoffersson (1981). Although we do not elaborate the technical details of this procedure here, we report one finding. Following the Baseline 3 and partial invariance models, we fixed the latent variable variance at unity in the Minnesota sample and fixed the mean for this group at zero; this scaling led to factor scores for the Minnesota sample that were in a z-score metric. Forcing factor loadings for the first four items to be invariant across groups and allowing parameters associated with the fifth item (*distressed*) to vary across groups (as in the partial invariance model), the Nanjing sample mean was estimated as -0.38, and the variance for the Nanjing sample was 0.43. Thus, the Nanjing sample had a mean on the latent variable that was slightly more than one third of a standard deviation lower than that of the Minnesota sample (with the Minnesota sample used to define the σ units), and the Nanjing sample exhibited rather restricted variance on the latent variable ($SD = 0.65$) relative to the Minnesota sample ($SD = 1.00$). As in the two-group covariance structure analysis reported earlier, all parameters in the final partial invariance mean structure analysis were highly significant ($p < .0001$), with z values ranging upward from 6.2.

The magnitude of the differences between the two groups in mean and variance on the latent variable may be evaluated in at least two ways. First, parameter estimates may be contrasted across groups relative to their standard errors. The Minnesota distribution on ξ was fixed at a mean of zero and variance of unity. The Nanjing sample mean was -0.378 ($SE = 0.060$), so the hypothesis that the means for the Minnesota and Nanjing samples were equal may be rejected, $z = -0.378/0.060 = -6.27$, $p < .0001$. In turn, the Nanjing sample variance was 0.431 ($SE = 0.053$); the hypothesis of equal variances across groups on the latent variable may also be rejected, $z = 8.20$, $p < .0001$.

The second, complementary way of evaluating the group differences on the latent variable is to specify restricted forms of the final partial measurement invariance model in which the mean or variance (or both) on ξ for the Nanjing sample is constrained to be equal to that for the Minnesota sample. The null model for the mean structures analysis had $\chi^2(26, N = 1{,}138) = 1{,}661.46$, $p < .0001$; the comparable partial measurement invariance model had a relatively much improved index of statistical fit, $\chi^2(18, N = 1{,}138) = 96.44$, $p < .0001$, and quite acceptable indices of practical fit (TLI = .931, NI = .952, and RMSEA = .062). Constraining the Nanjing mean on ξ to equal that for the Minnesota sample led to a large change in the statistical index of fit, $\chi^2(1, N = 1{,}138) = 40.45$, $p < .0001$, and rather worse levels of practical fit (TLI = .901, NI = .928, and RMSEA = .074). In addition, constraining the Nanjing sample variance on ξ to equal that for the Minnesota sample led to another large change in the statistical index of fit, $\chi^2(1, N = 1{,}138) = 51.34$, $p < .0001$, and even worse levels of practical fit (TLI = .866, NI = .897, and RMSEA = .086). Each of these model comparisons demonstrates that the differences between the Minnesota and Nanjing samples in mean and variance on the latent variable ξ were rather large and significant.

The Item Response Theory Approach

The Item Response Theory Model

Whereas CFA models account for the covariance between test items, IRT models account for examinee item responses. To accomplish this, IRT models stipulate a nonlinear monotonic function (called an item response function, or IRF) to account for the relation between examinee level on a latent variable (denoted by Θ) and the probability of a particular item response (Lord, 1980). The basic assumptions in IRT modeling are that the item responses are unidimensional and locally independent.

Unidimensionality implies that the set of items assesses a single underlying trait dimension; local independence means that if Θ level is held constant statistically, the test items are pairwise uncorrelated.

For any test item, many IRFs could yield a plausible account of the relation between Θ level and item response probability. Standard texts on IRT (e.g., Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980) describe commonly used functions. Applications can be found in Embretson (1985); Hulin, Drasgow, and Komocar (1982); Jensema (1974); Lord (1977); Thissen and Steinberg (1988); and Waller and Reise (1990). Because we have graded items with five response categories in this study, an appropriate model for our data is the graded response model (GRM) described by Samejima (1969). The fundamental equation for the GRM is

$$P(x = k|\theta) = \frac{1}{1 + \exp[-a(\theta - b_{j-1})]} - \frac{1}{1 + \exp[-a(\theta - b_j)]}$$
$$= P^*(j-1) - P^*(j). \quad (6)$$

Equation 6 specifies the conditional probability of a person responding in a particular category (k, where $k = 1, \ldots, 5$). The boundaries between the response categories are represented by $j = 1, \ldots, 4$, and $j = k$. The terms on the right-hand side of Equation 6 are the IRFs (the P^*s) that give the probability of an examinee responding above a particular threshold (j), conditional on his or her Θ level. With five categories, there are four between-category thresholds. By definition, the probability of responding above threshold $j = 0$ is 1.0 [$P^*(0) = 1.0$], and the probability of responding above threshold $j = 5$ is 0.0 [$P^*(5) = 0$].

The GRM model in Equation 6 requires that, for each test item, one a and four b_i parameters be estimated. That is, because there are four thresholds between the five ordered response categories, four IRFs are necessary to describe responses to an item with five categories. The a parameter is called the item discrimination parameter, and its value is proportional to the slope of the IRFs. Item discrimination parameters are constant for each of the IRFs within an item, so the multiple IRFs for a given item do not cross; a parameters can vary, however, between items. The a parameters are analogous to the λ_{mp} parameters in CFA models because they represent the relationship between the latent variable Θ and item responses. That is, the more strongly responses on an item are related to the latent variable Θ, the larger the corresponding a parameters for the IRFs. The b_j parameters are called category difficulties or thresholds, and each is defined as the point on the Θ scale (i.e., the trait level) at which the probability is 50% that the item response is greater than threshold j. Item threshold parameters are not incorporated in the standard linear CFA measurement model (but, see Muthén, 1984, 1988).

Just as estimated CFA models imply particular covariance matrices, estimated IRT models imply particular distributions of item responses conditional on Θ level. For illustration, the IRFs for an item with $a = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 1.0$, and $b_4 = 2.0$ are shown in Figure 1; these IRFs are the P^* values from Equation 6. For example, IRF_1 indicates that there is only a 20% probability that a person with a Θ score of -3.0 will score above Category 1 ($k = 1$) on the item, that this increases to a 50% probability of scoring above Category 1 for a person with a Θ score of -2.0, and that the probability of scoring above Category 1 increases rapidly to more than 80% for a person with a Θ score of -1.0 or higher. In contrast, IRF_3 shows that the probability of scoring above Category 3 is approximately 20% for a person with a Θ score of 0.0, rises to 50% for a person with a Θ score of 1.0, and then increases to 80% or more for a person with a Θ score of 2.0 or higher.

Given the parameters for the IRFs in Figure 1 and any value of Θ, Equation 6 can be used to compute the expected response proportions, or probabilities, in each of the five categories for the hypothetical item. For example, the probability of responding in Category 2 is $P_1^* - P_2^*$ (cf. Figure 1, specifically IRF_1 and IRF_2), which is the probability of responding above threshold $j = 1$ minus the probability of responding above threshold $j = 2$ for each value of Θ. That is, the probability of responding in Category 2 is the difference in expected response represented by IRF_1 and IRF_2 at each value of Θ. These expected proportions are shown in Figure 2. For example, the probability of scoring in Category 1 is relatively high for individuals with low values of Θ (i.e., about 80% for those with Θ values of -3.0) but drops off rapidly as Θ increases. In contrast, the probability of scoring in Category 3 is maximal for people with Θ values of 0.0, and the probability of scoring in Category 3 falls off symmetrically as Θ values deviate from 0.0.

Figure 2 can also be used to estimate the likelihood that a person will obtain a score in each category conditional on his or her value of Θ. That is, consider a person with an estimated Θ value of 0.0. This person would have the following approximate response probabilities: 5% probability of scoring in Category 1, 13% in Category 2, 64% in Category 3, 13% in Category 4, and 5% in Category 5. These response probabilities sum to 100% across all five categories of response at each value of Θ.

Assessing the Fit of IRT Models

Many statistics have been proposed to test the fit of IRT models (see McKinley & Mills, 1985). Typically, fit is assessed at the item level by a statistic that tests the congruence between the proportion of item responses in a particular category predicted from an IRF and the proportion of responses in a particular category observed in the data. We did not use any of these item-fit statistics. Rather, we adopted a model-testing approach (see Thissen, Steinberg, & Gerrard, 1986) to maintain consistency between the IRT and CFA sections of this research.

In a fashion analogous to CFA, the statistical acceptability of an estimated IRT model is contingent on how close the response proportions predicted from the IRFs are to the response proportions observed in the data. To conduct model-fit tests for the estimated IRFs, we used the value "-2 times the log of the likelihood function"; this value leads to the statistic G^2, which is part of the standard output from MULTILOG. Under certain conditions, G^2 is distributed as a chi-square value with degrees of freedom equal to the number of response patterns minus the number of estimates made in the model (see Thissen, 1991). G^2 values reflect the lack of congruence between the frequency of observed response patterns and the frequency of these patterns predicted by the estimated IRFs; the lower the congruence, the higher the value of G^2.

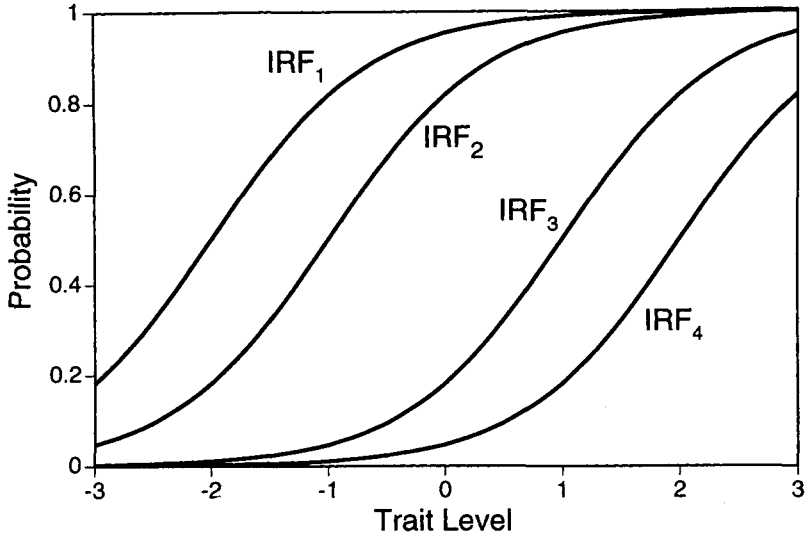

Figure 1. The item response functions (IRFs) for an item with $a = 1.5$, $b_1 = -2.0$, $b_2 = -1.0$, $b_3 = 1.0$, and $b_4 = 2.0$.

With large item sets (e.g., more than five items) or polychotomous item responses, G^2 is not appropriate for judging the fit of baseline models (i.e., models in which all parameters are freely estimated). The reason is that there will be too many unobserved response patterns and the statistic will have no known reference distribution. Nevertheless, G^2 values can be used to compare nested models; the difference between G^2 in Model 1 and G^2 in Model 2 is distributed as a chi-square value with degrees of freedom equal to the difference in degrees of freedom for Models 1 and 2 if the models are nested. The nesting of IRT models is similar to the nesting of CFA models; one IRT model is nested within a second IRT model if one may arrive at the first model only by placing constraints on parameters in the second IRT model and by making no new parameter estimates.

Unlike CFA modeling, in IRT there are few, if any, standard procedures for assessing the "practical" importance of deviations in the overall fit of a model to data.[2] We propose that test score appropriateness (Drasgow, Levine, & Williams, 1985) or, as termed here, person-fit (Reise, 1991) statistics be used to judge practical fit in IRT modeling. Person-fit statistics test the estimated IRFs at the level of the individual (Drasgow, Levine, & McLaughlin, 1991; Drasgow et al., 1985; Levine & Rubin, 1979; Tatsuoka, 1984; Wright, 1977). Lack of person fit indicates that a person's item responses are not congruent with, or predictable from, the estimated IRFs. As a result, the person's Θ estimate, which is derived from the IRFs, is not comparable to other people's Θ estimates.

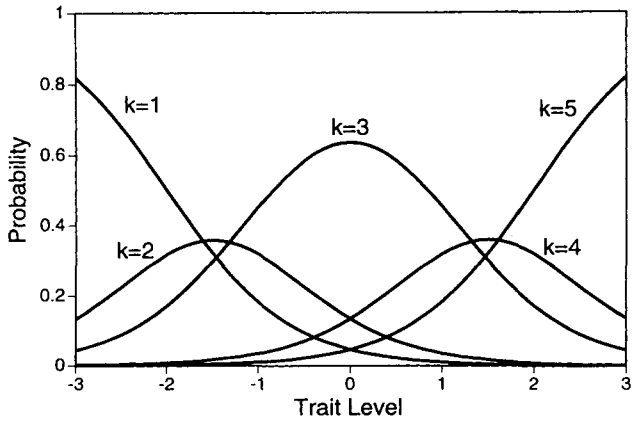

Figure 2. Probability of observing a response in a particular category conditional on Θ level for the hypothetical item in Figure 1.

In this study, we applied a measure of person fit called Z_l (Drasgow et al., 1985) to assist in interpreting the practical importance of lack of model fit. The Z_l statistic is the standardized value of the likelihood of an individual's item response pattern given the IRFs. By assumption, Z_l scores are distributed as standard normal values (i.e., $M = 0.0$, $SD = 1.0$) under the null hypothesis of person fit. Consequently, the reference distribution for evaluating Z_l is the z distribution. Significantly large, negative Z_l values indicate a lack of person fit associated with response patterns that are unlikely given the estimated model.

Specifying Measurement Invariance in IRT Models

One way to place individuals' Θ estimates onto a common scale, even across groups of examinees, is to score all examinees

[2] Some indices of practical fit of IRT models to data have been suggested (e.g., Kingston & Dorans, 1985). However, these indices are not directly analogous to indices of practical fit in CFA. That is, these IRT indices tend to reflect the relative fit of items to the IRT model rather than representing an overall proportional fit of a model to data. Of course, if IRT models are reparameterized as nonlinear factor analysis models (cf. McDonald, in press) or as factor models for discretized variables (Takane & de Leeuw, 1987), then the indices of practical fit developed for CFA models could be extended directly to the reparameterized IRT models.

using the same IRFs. However, the resulting Θ estimates could be biased and substantively meaningless unless measurement invariance has first been established. In IRT terms, test items provide equivalent measurement when the IRFs are the same across groups (Thissen et al., 1986). In CFA, one tests for measurement invariance by establishing the equivalence of factor loadings; in contrast, in IRT analysis, one considers both the item discrimination (a) and item difficulty (b_j) parameters.

In this study, items provided equivalent measurement if the a, b_1, b_2, b_3, and b_4 parameters were the same for the Minnesota and Nanjing samples. Hui and Triandis (1985, p. 138) stated that "an instrument that has similar ICCs [i.e., IRFs] across cultures has, at least in part, demonstrated its item equivalence and scalar equivalence." Stated differently, an item is not biased when examinees at the same Θ level have the same expected probability of response regardless of their group membership. If response probabilities for people at the same Θ level depend on group membership, then an item is said to show differential item functioning, or DIF. Although many researchers suggest discarding items with DIF, the presence of items with DIF need not eventuate in biased measurement, as we show later.

The hypothesis of between-groups equality of IRFs can be tested with the multiple-groups option in MULTILOG VI (Thissen, 1991). This program uses marginal maximum likelihood (Bock & Aitkin, 1981; Thissen, 1982) to estimate IRFs simultaneously in two or more groups. MULTILOG also provides mechanisms to estimate population group mean differences on the latent variable. Interested readers are referred to Thissen (1991) for further details on technical aspects of the program, Koch (1986) for an application of the GRM, and Reise and Yu (1990) for parameter recovery information under the GRM using MULTILOG.

The Baseline Model in IRT

In the CFA section, we noted the identification problem that exists between the scale of the latent variable ξ and the scale of the factor loadings, the λ_{mp}s. In IRT, a similar indeterminacy holds between the scale for the latent variable (Θ) and the scale for the item parameters (a, b_j). In particular, the location of the Θ scale affects the location of the b_j parameters, and the dispersion on the Θ scale influences the size of the a parameters. When data from a single group are calibrated (i.e., the parameters of the model are estimated) with a marginal maximum likelihood algorithm (as used in MULTILOG), the Θ scale is identified by stipulating a population distribution for Θ (Baker, 1990). For example, one can specify that examinees are sampled from a normal population distribution with mean Θ equal to 0.0 and standard deviation equal to 1.0. This specification identifies the scale for Θ, which in turn identifies the scale for the item parameters.

In a multiple-group situation, the ultimate goal is to identify the latent variable between groups. That is, one must make the scale of the latent variable common across groups so that the item parameters can be estimated with respect to this scale and then tested for equivalence. Because the technical details of IRT modeling are infrequently discussed outside the psychometric literature, we devote this section to discussion of identification in the multiple-group situation.

Recall that in the CFA Baseline 2 model (see Table 2), the metric for the latent variable was identified by fixing its variance to 1.0 within each group. Analogously, with IRT, the Θ scale could be identified within each group by conducting independent calibrations, each time stipulating a population distribution for Θ. If this were done, however, the scale for Θ would be identified within each group but would not be common across groups.[3] Consequently, comparison of values of the item parameters across groups would not be legitimate. This exemplifies a fundamental, but not well understood, principle of IRT and CFA; namely, item parameters from independent calibrations are not comparable.

In the CFA Baseline 1 model, we identified the latent variable within each group by stipulating that $\lambda_{11} = 1.0$ for each group. We noted that if Item 1 were invariant, then the λ parameter estimates for the freely estimated items would be comparable across groups. In a similar way, in IRT one can make the Θ scale common across groups by using an anchor test. An anchor test is a set of test items that are constrained to have the same parameters between groups. Calibrating items concurrently with the anchor items results in the Θ scale being identified and on a common metric across the groups.

Here, the goal is to establish a well-fitting baseline IRT model in each group before testing whether item parameters are invariant across groups. An ideal baseline is a model in which all parameters are freely estimated, except for minimal identifying constraints. In CFA, this means that the factor loadings and factor variances are freely estimated in each group, subject to minimal constraints to identify the model. In IRT, this means that the item parameters and the means and variances of the latent variable are freely estimated for each group. Two concerns interfere with this ideal in IRT. First, two parameters must be fixed (e.g., a group mean and standard deviation on Θ) to identify the IRT model. Second, as far as we could determine, the MULTILOG program allows the standard deviations on Θ for each group to be fixed to some value but not to be estimated freely.

With these considerations in mind, the baseline model was established as follows. The input data for MULTILOG consisted of 540 and 598 item response patterns from the Minnesota and Nanjing samples, respectively. We used a form of concurrent item calibration (see Hambleton et al., 1991, p. 135) by setting up the data as follows: First, the data were treated as if 1,138 (540 Minnesota and 598 Nanjing) people had taken a 10-item test. Second, responses to Items 1–5 were coded as missing for the 598 Nanjing response vectors, and item responses to Items 6–10 were coded as missing for the 540 Minnesota response vectors.

This model has 10 items, and the data to be analyzed are

[3] In the case provided earlier, some researchers would suggest placing IRT item parameters from independent calibrations onto a common scale by using a metric linking procedure (Divgi, 1985; Stocking & Lord, 1983; Vale, 1986). However, as pointed out by Lautenschlager and Park (1988), linking is inappropriate when the groups are heterogeneous (i.e., not drawn from the same population) and items potentially contain DIF. More recently, iterative linking and item bias (i.e., DIF) detection procedures have been proposed that may diminish these concerns (Candell & Drasgow, 1988; Park & Lautenschlager, 1990).

1,138 response patterns. As a result, 50 item parameters were estimated in the baseline model (10 a parameters and 40 b_j parameters). The mean and standard deviation on Θ for the Minnesota data were fixed at 0.0 and 1.0 for identification. The mean on Θ was freely estimated for the Nanjing data, and the standard deviation was fixed at 0.65.[4] In this baseline model, the Θ metric is not identified between groups because there are no anchor items. However, the fit of this model serves as a baseline for judging subsequent models in which invariance is investigated across groups of item parameters.

The resulting item parameter estimates for the baseline model are shown in Table 3. Table 3 also lists the means and standard deviations of the observed Z_l (person-fit) values within each group. These values were computed by scoring individuals (i.e., estimating Θ level) in the Minnesota sample with the Minnesota IRFs and then scoring individuals from the Nanjing sample with the Nanjing IRFs. Significance tests for Z_l were conducted at the .05 level (one-tailed).

As mentioned, the G^2 value for the baseline model is not interpretable directly, so this value is not reported. Judgments of fit rely mostly on the distribution of Z_l scores and the number of significant Z_l scores. When response patterns are congruent with the estimated IRFs, Z_l scores are expected to have a mean of 0.0 and a variance of 1.0. In both samples, this appears to be reasonably true. Also, the percentage of significant Z_l scores in each group is close to the nominal Type I error rate. We concluded that, at the person level, a satisfactory baseline model had been established within each group. We would be confident using the Minnesota IRFs to scale Americans and the Nanjing IRFs to scale the Chinese. We emphasize, however, that such scores would not be comparable between groups.

Inspection of the item parameters for the baseline model reveals that the IRFs appear to differ across groups. In particular, the item *distressed* appears to be more discriminating for the Nanjing sample (i.e., the a parameter estimate for the Nanjing sample is much larger than that for the Minnesota sample). Before claiming DIF, however, we must examine whether the apparent DIF is due only to sampling error or chance fluctuation.

Testing for Full Measurement Invariance in IRT Models

To explore whether the NA items exhibited significant levels of DIF, we next tested for full measurement invariance by calibrating the items concurrently with the following stipulations. First, all item parameters were constrained to equality across groups; in effect, this creates a five-item anchor test. We fixed the mean and standard deviation of Θ to 0.0 and 1.0, respectively, for the Minnesota sample. The population mean for the Nanjing sample was an estimated parameter, and the standard deviation for this sample was fixed at 0.65. The G^2 from this full invariance model could then be compared with the G^2 from the baseline model. If G^2 for the full invariance model were significantly greater than that for the baseline, we could reject the hypothesis of full measurement invariance and conclude that at least one item must contain DIF.

As evidenced in Table 3, the change in G^2 between the baseline and full invariance models was statistically significant, $G^2(25, N = 1{,}138) = 118.3$, $p < .0001$. Although the full measurement invariance hypothesis must be rejected statistically, it is important to ask, What are the practical consequences of scoring individuals from both groups with these common item parameters? The means and standard deviations of Z_l scores within groups reported in Table 3 address this question. As far as person fit is concerned, the common IRFs appear to do an adequate job of scaling individuals onto a common scale; 96% of examinees within each group are scalable on the common trait dimension. However, the distributions, especially the standard deviations, of Z_l scores within groups for the full invariance model were rather different from those under the baseline model, suggesting that the assumption of full measurement invariance may have placed overly stringent constraints on the IRT model.

Partial Measurement Invariance in IRT Models

As in CFA, comparing examinees on a common metric does not require that all items be invariant in their measurement. That is, the presence of one or more items exhibiting DIF should not prevent the scaling of individuals onto a common metric. Our basic requirement in IRT is that at least one item be invariant across groups. The invariant item can then be used as an anchor to estimate Θ values for individuals within both groups concurrently on a common scale. Stated differently, certain items can be allowed to have different IRFs across groups (i.e., to exhibit DIF), as long as these IRFs are related to a common scale for Θ across groups. To accomplish this, we used the following method.

Using the fit values from the baseline model, we proceeded to test for measurement invariance on an item-by-item basis. This involved specifying, one item at a time, that the a and four b_j parameters were invariant across groups. For each of these analyses, a (0,1) population Θ distribution was specified for the Minnesota sample, the mean for the Nanjing sample was freely estimated, and the standard deviation was fixed at 0.65. As shown in Table 4, constraining the IRFs for the items *nervous* and *tense* to invariance did not result in a significant (critical $\alpha = .05$) increase in G^2; however, constraining the IRFs for any of the remaining items to invariance across groups led to significant changes in the fit of the model to the data. Hence, we concluded that the items *nervous* and *tense* were invariant across groups.

Although a two-item anchor test might be less than ideal for many practical purposes, the most viable option statistically at this point is to use *nervous* and *tense* as anchors to establish a common metric. Using these items as anchors, we then tested a partial measurement invariance model by constraining item parameters for the anchors to equivalence across groups and allowing all other item parameters to be freely estimated. The item parameters and relevant fit statistics for this partially invariant model are shown in Table 3. The fit of this model is not significantly different from the baseline, $G^2(10, N = 1{,}138) = 15.1$, *ns*. The best estimate of group mean difference on the latent variable is -0.36 Θ units, with the Nanjing sample scoring

[4] The value 0.65 was selected for the Nanjing standard deviation on Θ on the basis of the finding in the CFA analysis section of this research. However, in most multiple-group situations, the optimal standard deviation parameter (i.e., optimal in the sense of leading to the best fit of the IRT model to the data) would have to be estimated by trial and error.

Table 3
Graded Response Model Item Parameter Estimates and Fit Values for the Minnesota and Nanjing Samples

		Baseline		Full invariance		Partial invariance	
Item	Parameter	Minnesota	Nanjing	Minnesota	Nanjing	Minnesota	Nanjing
Nervous	a	2.25	2.32	2.28		2.27	
	b_1	−0.58	−0.64	−0.64		−0.62	
	b_2	0.50	0.36	0.41		0.43	
	b_3	1.26	1.14	1.18		1.21	
	b_4	2.39	1.92	2.18		2.22	
Worried	a	2.10	1.83	2.01		2.16	1.80
	b_1	−0.99	−0.77	−0.86		−1.03	−0.73
	b_2	0.08	0.21	0.12		0.02	0.27
	b_3	0.87	0.83	0.80		0.80	0.89
	b_4	1.91	1.76	1.78		1.82	1.85
Jittery	a	1.56	1.96	1.70		1.59	1.96
	b_1	−0.38	0.05	−0.13		−0.44	0.10
	b_2	0.73	0.70	0.74		0.66	0.75
	b_3	1.81	1.17	1.52		1.73	1.21
	b_4	2.81	1.86	2.39		2.72	1.91
Tense	a	2.37	2.11	2.16		2.14	
	b_1	−0.72	−0.91	−0.85		−0.83	
	b_2	0.27	0.05	0.13		0.15	
	b_3	1.07	0.81	0.93		0.96	
	b_4	2.02	1.77	1.91		1.95	
Distressed	a	2.02	3.34	2.33		2.02	3.41
	b_1	−0.55	−0.48	−0.52		−0.60	−0.44
	b_2	0.36	0.17	0.28		0.30	0.22
	b_3	1.08	0.63	0.89		1.01	0.67
	b_4	2.05	1.12	1.69		1.98	1.15
Mu		0.00[a]	−0.41	0.00[a]	−0.42	0.00[a]	−0.36
Sigma		1.00[a]	0.65[a]	1.00[a]	0.65[a]	1.00[a]	0.65[a]
G^2 change		—	—	118.3		15.1	
df change		—	—	25		10	
$M\ Z_l$		0.35	0.28	0.40	0.24	0.39	0.25
$SD\ Z_l$		0.98	0.80	0.87	0.89	0.95	0.82
No. rejected		27	14	22	22	27	19
% rejected		0.05	0.02	0.04	0.04	0.05	0.03

Note. Parameter estimates that are centered between the Minnesota and Nanjing columns (e.g., the 2.28 estimate of a in the full invariance model) represent parameters constrained to equality across samples. Also, because the G^2 values for each model are not interpretable directly, these values are not listed. Instead, the G^2 values reported are the change in G^2 and the accompanying change in degrees of freedom when comparing a model with the baseline model.
[a] Parameters fixed at tabled values to identify each model.

lower than the Minnesota sample. This final IRT model, with different IRFs across groups for three of the five items, could then be used in applied situations to score examinees in different groups so that the resulting trait level estimates are on a common scale.

Discussion

The primary goals of this article have been to compare CFA and IRT procedures for establishing measurement invariance across populations, to clarify aspects of model fitting with CFA and IRT through the application of both types of procedure to an empirical data set, and to identify topics in CFA and IRT that should be the focus of additional research. Our concern with measurement invariance is an important one, because measurement invariance is a basic requirement or prerequisite for studying group differences with statistical models. Once measurement invariance is established, additional theoretically important questions may be addressed, including questions regarding group differences in means or variances on the latent variables identified.

Comparison of CFA and IRT Modeling Procedures

CFA and IRT procedures may be compared at both theoretical levels and more practical levels. As might be expected, certain differences between the two approaches appear to favor CFA, whereas other differences seem to favor IRT. Consider first more theoretical comparisons between CFA and IRT. Assuming the presence of a single latent variable, the CFA model holds that the latent variable is linearly related to each of its indicators (cf. Equation 1). That is, individual differences on the latent variable (ξ_p) are linearly related to individual differences on a given indicator X_m, and the factor pattern coefficient (λ_{mp}) is the raw score regression coefficient representing this linear relationship. Errors in variables, or variance in indicators linearly un-

Table 4
Fit Values for Partially Invariant IRT Models

Items with invariant parameters	Change in G^2	Change in *df*	*p*	Minnesota		Nanjing	
				M	*SD*	*M*	*SD*
None	—	—	—	0.00	1.00	−0.39	0.65
Nervous	4.4	5	>.50	0.00	1.00	−0.45	0.65
Worried	18.5	5	<.025	0.00	1.00	−0.42	0.65
Jittery	47.3	5	<.001	0.00	1.00	−0.43	0.65
Tense	11.8	5	>.05	0.00	1.00	−0.36	0.65
Distressed	25.3	5	<.01	0.00	1.00	−0.48	0.65
Nervous and tense	15.1	10	>.05	0.00	1.00	−0.36	0.65

Note. The model with no item parameters constrained to invariance across groups is the baseline model.

related to the latent variable, are represented explicitly in the CFA model by the indicator residuals (the δ_ms). A model of this form is analogous to a standard linear regression model, a type of model used throughout areas of psychological research.

Similarly, the latent variable (Θ) in IRT models represents individual differences in response tendency, but individual differences on Θ are presumed to be related only monotonically to responses on each item. In IRT models, the *a* coefficients relate the latent trait (Θ) to item responses, fulfilling the function of the λ_{mp} estimates in CFA models. As with λ_{mp} estimates, the larger the *a* coefficient for an item, the more strongly the item is related to the latent variable. However, the relationship between Θ and the probability of a subject's response is not a simple linear one. Moreover, linear CFA models disregard the category threshold, or difficulty, parameters in IRT (the *b*s) that could be important psychologically. At present, work on nonlinear factor analysis models (e.g., McDonald, 1982, in press; Muthén, 1984, 1988; Takane & de Leeuw, 1987), which contain threshold parameters, is still in relatively early stages of development; few applications of these methods appear outside the technical psychometric literature. However, future research on nonlinear factor analysis approaches should be pursued to allow a more complete comparison of IRT and factor-analytic models for representing psychological data.

One clear and important similarity across CFA and IRT approaches emerged in the partial measurement invariance analyses with regard to the function of invariant and noninvariant items. Specifically, all items, whether invariant across groups or not, were useful in representing individual differences on the latent variable within each group. The invariant items then allowed individual differences on the latent variable within groups to be linked to a common metric for the latent variable across groups. To make this claim more concrete, consider the following: For the IRT analyses, there were five items in each group, each answered on a 1–5 scale. This leads to the possibility of 5^5, or 3,125, potential response patterns that could be exhibited by subjects within each group; across the two groups, this would result in 6,250 potential response patterns. In our sample, we observed 525 response patterns across the 1,138 persons in the two samples. If we had restricted our evaluation of IRT models to only the two items that had invariant parameters across groups, we would have been restricted to consideration of only 5^2, or 25, potential response patterns within each group. Clearly, retaining all five items and basing our analyses on the 525 response patterns obtained allowed a richer representation of individual differences in responding and individual differences on the latent variable, Θ, within each group. Then, with two items exhibiting invariance across groups, we were able to establish a common metric for the latent variable across groups to frame and test questions regarding group differences in mean and variance on the latent variable.

Turning to more practical comparisons, we can identify at least two contrasts between CFA and IRT procedures with regard to the fitting of models to data. These contrasts concern the ease of (a) specifying models for data and (b) evaluating the outcomes of model comparisons.

With regard to specifying a model for a set of data, procedures for CFA seem rather more advanced, simpler, and more user friendly than those developed for IRT. In the present article, we used the LISREL program for our CFA analyses, but other programs (e.g., EQS, Bentler, 1989; CALIS, SAS Institute, 1989) are available, and all are fairly easy to use. Each program for CFA requires the specification of each set of data (e.g., the number of measured variables and the number of subjects), and data from multiple groups are placed one after the other. The user must then perform model comparisons of interest by invoking constraints across groups in the estimation of particular parameters. Once the number of subjects in each group, the number of measured variables per group, and the parameters of the structural model have been specified, CFA programs calculate accurate estimates of the chi-square measure of model fit and its associated degrees of freedom. This statement glosses over many potential difficulties in the fitting of CFA models to data; difficulties involving model misspecification, acceptable parameter estimation, and so forth. However, given successful fitting of a model to a set of data, the chi-square statistical measure of fit and its degrees of freedom are routinely presented in an accurate fashion.

The preceding way of specifying multiple-group CFA differs from the multiple-group model specification in IRT using the MULTILOG program (Thissen, 1991). Although other programs are available for fitting IRT models to data (e.g., BILOG, Mislevy & Bock, 1986), only MULTILOG easily enables fitting the graded-response IRT model (but see the LISCOMP program, Muthén, 1988). In MULTILOG, the specification of the graded-response IRT model with multiple groups must be un-

dertaken as if the two groups are subgroups within a larger group. Then, given n observed items in each group, one must specify that the data consist of $n \times G$ items, where g is an index for group ($g = 1, \ldots, G$), with observed responses by individuals in group g represented on the gth set of items. For example, in the present study, we obtained the responses of all subjects to 5 items. To fit the graded-response IRT model to the two-group data using MULTILOG, we had to specify that there were 10 items (i.e., 5 items $\times$ 2 groups); the data set consisted, then, of 10 items. We recorded the observed responses (i.e., on a 1–5 scale) of each Minnesota subject for the first 5 items and listed a missing value code for each of the second set of 5 items; conversely, each Nanjing subject had a missing value on each of the first 5 items and his or her observed responses on the second 5 items.

More important, the chi-square measure of model fit and its associated degrees of freedom for IRT models using the MULTILOG program are not a simple function of the fit of the model to the data, the number of items, the number of parameters estimated, or the number of subjects. Rather, the chi-square value is a function of observed and expected response proportions, and the degrees of freedom are a function of the number of different response patterns minus the number of parameters estimated. Basically, the MULTILOG program presumes that the program user has a fairly sophisticated understanding of IRT theory and methods. If not, the user may err seriously in the specification of the model and the subsequent evaluation of model fit.

The second practical issue, evaluating the outcomes of model comparisons, also leads to interesting contrasts between CFA and IRT procedures. In CFA modeling, researchers may rely on both statistical and practical indices of fit. The primary statistical measure of fit is the likelihood ratio chi-square test, which provides an overall test of model fit to the data. If nested CFA models are formulated, the chi-square difference test provides a statistical test of the difference in fit of the two models (cf. Bentler & Bonett, 1980). In addition to the statistical measure of fit, a large number of practical indices of fit have been proposed, three of which were used in this study. Given the presence of Monte Carlo evaluations of these indices of fit (e.g., Bentler, 1990; Marsh et al., 1988), evidence is accumulating with regard to the evaluation of these types of indices.

With IRT models, the only standard measure of fit is a likelihood ratio chi-square variate that is evaluated in a fashion similar to that of the statistical measure of fit with CFA. That is, a significant chi-square value provides a statistical basis for rejecting a model, statistical tests of the difference in fit of nested IRT models may be obtained, and the optimal model is one that has a minimal number of parameter estimates but has as small a chi-square value as possible. Unfortunately, practical fit indices analogous to those for CFA models have not yet been widely developed for IRT models. In current applications of CFA, practical fit indices are at least as important as, if not more important than, the statistical index of fit. The same could become true for IRT. If more well-researched practical indices of fit were available for IRT models, researchers could rely on several sorts of information regarding model fit to data, as is now standard in CFA modeling.

Comparisons in Model Fit to Negative Affect Data

A review of the fitting of the CFA and IRT models to the negative affect data revealed notable similarities, as well as certain fairly minor differences, in outcome. Under both models, full measurement invariance was rejected as an adequate description, but partial measurement invariance was accepted. Similarly, under both models, Items 1 and 4 revealed measurement invariance across groups, and Item 5 was not invariant. However, one notable difference between the models involved Items 2 and 3 (*worried* and *jittery*): Under the CFA model, there was no evidence that these two items displayed different characteristics across the two groups, under the IRT model, Items 2 and 3 displayed statistically significant differences in parameter estimates across groups. The primary reason for this difference between the CFA and IRT partial measurement invariance models is that the CFA model ignores certain parameters, namely the difficulty parameters represented by the b_j parameters in the IRT model. As a result, IRT models posit more stringent sets of measurement invariance constraints.

With regard to characterizing the differences between groups on the latent variables defined by the two procedures, quite similar results were found. After the mean of the Minnesota sample was fixed to zero to identify each model, the Nanjing sample mean was approximately one-third σ units below the Minnesota mean; the estimates were -0.38 σ units and -0.36 σ units on the basis of the CFA and IRT models, respectively. Of course, this similarity in estimated mean levels may have resulted from the fixing of the Nanjing sample standard deviation at 65% of the Minnesota sample standard deviation in the IRT analyses. In the IRT analyses with MULTILOG, it appears that standard deviations must be fixed, rather than estimated, for each group. We attempted to use the results of the CFA analyses to inform our IRT modeling; fixing the values of the standard deviations for each group in the IRT analyses on the basis of values from the CFA almost certainly led to rather more similar estimates of mean levels from the CFA and IRT analyses than would have occurred otherwise.

Issues for Future Research

The preceding comparisons between CFA and IRT models point to several issues for future research. The development of practical fit indices for IRT models is one rather pressing topic for future research. This topic was discussed in a preceding section; thus, extended discussion is unnecessary here. In brief, researchers applying CFA procedures to data typically stress practical indices of fit at least as strongly as the likelihood ratio chi-square statistical index of fit. If practical fit indices were available for IRT models, researchers would have a richer and more varied set of indices for evaluating the fit of IRT models to data. Given (a) the rather large sample sizes required to obtain accurate estimates of IRT parameters, (b) the influence of sample size on likelihood ratio chi-square statistics, and (c) the relative independence from sample size of good practical indices of fit, such practical indices would seem most welcome in analyses with IRT models. The most fruitful approach to development of such indices will probably rely on the work of McDonald (in press) and Takane and de Leeuw (1987) on the equivalence of

IRT and nonlinear factor analysis models; if IRT models may be reparameterized as nonlinear factor models, the indices of practical fit developed for CFA models could be extended directly to the evaluation of IRT models.

A second issue that would lead to more direct and informative analyses is the development of modification indices for IRT models. In CFA analyses, modification indices from a full invariance model indicate the likely change in chi-square values that would accompany the freeing of each constrained parameter, thereby moving to a model with partial measurement invariance. Similar modification indices would be a very useful addition to IRT programs. Researchers may currently obtain information similar to that provided by modification indices by brute force specification and testing of all IRT models that differ from a given model by the addition of a single parameter estimate or by the relaxing of a single constraint. In large models, this would be a very inefficient and time-consuming process. Modification indices, even if only moderately accurate, would provide a useful addition to IRT modeling software. Many experts on CFA modeling consider modification indices dangerous, enabling mere "data fitting" or the post hoc modification of models without a priori, theoretical justification. However, given the stipulation that model modifications be replicated across samples to be supported strongly, modification indices provide efficient means for respecifying models and, ultimately, remolding theories.

A third and final topic for future research is the further exploration of the relations between CFA and IRT models and their differential utility for representing data and testing theoretical hypotheses. We have presented an initial exploration into this topic, perhaps the most important of the issues for future research. Both CFA and IRT models provide interesting ways of representing data in the social and behavioral sciences; future investigators should search for ways of deciding which model is best for which purpose. The outcome of such work would be a more coherent framework for the use of current, state-of-the-art methods of psychometric analysis.

References

Alwin, D. F., & Jackson, D. J. (1981). Application of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 249–279). Beverly Hills, CA: Sage.

Baker, F. B. (1990). Some observations on the metric of PC-BILOG results. *Applied Psychological Measurement, 14,* 139–150.

Bentler, P. M. (1989). *EQS structural equation program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika, 46,* 443–459.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1,* 185–216.

Browne, M. W. (1990). *MUTMUM PC: User's guide.* Columbus: Ohio State University, Department of Psychology.

Byrne, B. M., & Shavelson, R. J. (1987). Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal, 24,* 365–385.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456–466.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12,* 253–260.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105,* 317–327.

Divgi, D. R. (1985). A minimum chi-square method for developing common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are central issues. *Psychological Bulletin, 95,* 134–135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70,* 662–680.

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171–191.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67–87.

Embretson, S. (1985). *Test design: Developments in psychology and psychometrics.* San Diego, CA: Academic Press.

Etezadi-Amoli, J., & McDonald, R. P. (1983). A second generation nonlinear factor analysis. *Psychometrika, 48,* 315–342.

Everitt, B. S. (1984). *An introduction to latent variable models.* London: Chapman & Hall.

Frederiksen, N. (1987). How to tell if a test measures the same thing in different cultures. In Y. H. Poortinga (Ed.), *Basic problems in cross cultural psychology* (pp. 14–18). Lisse, The Netherlands: Swets & Zeitlinger.

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

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16,* 131–152.

Hulin, C. L., Drasgow, F., & Komocar, J. (1982). Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67,* 818–825.

Jensema, C. J. (1974). An application of latent trait mental test theory. *British Journal of Mathematical and Statistical Psychology, 27,* 29–48.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd ed.). Chicago: SPSS.

Kingston, N. M., & Dorans, N. J. (1985). The analysis of item-ability regressions: An exploratory IRT model fit tool. *Applied Psychological Measurement, 9,* 281–288.

Koch, W. R. (1986). Likert scaling using the graded response latent trait model. *Applied Psychological Measurement, 7,* 15–32.

Lautenschlager, G. J., & Park, D. G. (1988). IRT item bias detection procedures: Issues of model misspecification, robustness, and parameter linking. *Applied Psychological Measurement, 12,* 365–376.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4,* 269–289.

Linn, R. L., & Harnisch, D. L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18,* 109–118.

Long, J. S. (1983). *Confirmatory factor analysis.* Beverly Hills, CA: Sage.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross cultural psychology* (pp. 19–29). Lisse, The Netherlands: Swets & Zeitlinger.

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

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 501–511.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

McDonald, R. P. (1962). Note on the derivation of the general latent class model. *Psychometrika, 27,* 203–206.

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometrika Monographs, 32*(Suppl. 15).

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (in press). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories of measurement: Problems and issues.* Ottawa, Ontario, Canada: Edumetric Research Group, University of Ottawa.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107,* 247–255.

McGaw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. *British Journal of Mathematical and Statistical Psychology, 24,* 154–168.

McKinley, R. L., & Mills, C. N. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 9,* 49–57.

Mislevy, R. J., & Bock, R. D. (1986). *PC-BILOG: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software.

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

Muthén, B. O. (1988). *LISCOMP: Analysis of linear structural equations with a comprehensive measurement model* (2nd ed.). Mooresville, IN: Scientific Software.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika, 46,* 407–419.

Park, D. G., & Lautenschlager, G. J. (1990). Iterative linking and ability scale purification as means of improving IRT item bias detection. *Applied Psychological Measurement, 14,* 163–174.

Reise, S. P. (1991). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement, 14,* 127–137.

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

Reynolds, C. R., & Harding, R. E. (1983). Outcome in two large sample studies of factorial similarity under six methods of comparison. *Educational and Psychological Measurement, 43,* 723–728.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs, 34*(Suppl. 17).

SAS Institute, Inc. (1989). *SAS/STAT user's guide, Version 6* (4th ed., Vol. 1). Cary, NC: Author.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

Steiger, J. H., & Lind, J. (1980, May). *Statistically based tests for the number of common factors.* Paper presented at the meeting of the Psychometric Society, Iowa City, IA.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49,* 95–110.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. *Psychometrika, 47,* 175–186.

Thissen, D. (1991). *MULTILOG: Multiple, categorical item analysis and test scoring using item response theory* (Version 6). Chicago: Scientific Software.

Thissen, D., & Steinberg, L. (1988). Data analysis using item response theory. *Psychological Bulletin, 104,* 385–395.

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

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38,* 1–10.

Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10,* 333–344.

van der Flier, H., & Drenth, P. J. D. (1980). Fair selection and comparability of test scores. In L. J. T. van der Kamp, W. F. Langerak, & D. N. M. de Gruijter (Eds.), *Psychometrics for educational debates* (pp. 87–101). New York: Wiley.

Waller, N. G., & Reise, S. P. (1990). Computerized adaptive personality assessment. *Journal of Personality and Social Psychology, 57,* 1051–1058.

Windle, M., Iwawaki, S., & Lerner, R. M. (1988). Cross-cultural comparability of temperament among Japanese and American preschool children. *International Journal of Psychology, 23,* 547–567.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97–116.

Received May 8, 1992
Revision received May 1, 1993
Accepted May 3, 1993 ■