Measurement Equivalence: A Comparison of Methods Based on Confirmatory Factor Analysis and Item Response Theory

Nambury S. Raju
Illinois Institute of Technology

Larry J. Laffitte
Psychological Associates, Inc.

Barbara M. Byrne
University of Ottawa

Current interest in the assessment of measurement equivalence emphasizes 2 major methods of analysis. The authors offer a comparison of a linear method (confirmatory factor analysis) and a nonlinear method (differential item and test functioning using item response theory) with an emphasis on their methodological similarities and differences. The 2 approaches test for the equality of true scores (or expected raw scores) across 2 populations when the latent (or factor) score is held constant. Both approaches can provide information about when measurement nonequivalence exists and the extent to which it is a problem. An empirical example is used to illustrate the 2 approaches.

There is currently a great deal of interest in the assessment of measurement equivalence. According to Drasgow and Kanfer (1985), a test or a subscale is said to have measurement equivalence across groups or populations if persons with identical scores on the underlying/latent construct have the same expected raw score or true score at the item level, the subscale total score level, or both. Without measurement equivalence, it is difficult to interpret observed mean score differences meaningfully. That is, observed mean score differences may reflect the true mean difference between the groups as well as a difference in the relationship between the latent variable and the observed score that is not identical across groups. When measurement equivalence is present, the relationship between the latent variable and the observed variable remains invariant across populations. In this case, the observed mean difference may be viewed as reflecting only the true difference between the populations.

Currently, there are two popular methods for establishing measurement equivalence. One method is based on structural equation modeling (or, more specifically, confirmatory factor analysis), and the other is based on item response theory. Given the current popularity of these methods, the purpose of this research is to offer a comparison of these two methods with a special emphasis on their methodological similarities and differences.

Confirmaory Factor Analysis

Although several computer programs are now available for structural equation modeling (SEM), three of the most popular are AMOS (Arbuckle, 1999), EQS (Bentler, 1995), and LISREL (Jöreskog & Sörbom, 1996). For convenience, the LISREL notation is used in this document to explain measurement equivalence within the context of SEM.

Following Jöreskog and Sörbom (1996), the linear confirmatory factor analysis (CFA) model may be represented as

\[ x = \Lambda_{x} \xi + \delta, \]  \hspace{1cm} (1)

where \( x \) represents a vector of \( q \times 1 \) observed or measured variables, \( \xi \) is a vector of \( n \times 1 \) latent variables or underlying factors, \( \Lambda_{x} \) is a \( q \times n \) regression coefficient or factor loading matrix that relates \( n \) factors to \( q \) observed variables, and \( \delta \) is a \( q \times 1 \) vector of measurement errors/residuals in \( x \). Equation 1 is commonly referred to as the measurement model for the exogenous variables in SEM. Typically, the vector \( x \) represents items that serve as the indicator variables (i.e., the observed variables generated by their underlying latent constructs); different items serve as indicator variables for different latent constructs (\( \xi \)) in a CFA. As a consequence, the regression paths or lambdas linking the items to their underlying latent constructs are of primary interest.

Assuming that measurement errors have zero means, the expectation \( (E) \) of \( x \) in Equation 1 can be written as

\[ E(x) = \Lambda_{x} \xi. \]  \hspace{1cm} (2)

Further, assuming that measurement errors are uncorrelated with each other and with the underlying factors, the variance-covariance matrix among the observed variables \( (\Sigma_{x}) \) can be expressed as

\[ \Sigma_{x} = \Lambda_{x} \Phi \Lambda^{'x} + \Theta_{x}, \]  \hspace{1cm} (3)

where \( \Lambda^{'x} \) is the transpose of \( \Lambda_{x} \), \( \Phi \) is a variance-covariance matrix of factor variables (\( \xi \)), and \( \Theta_{x} \) is a diagonal matrix of measurement errors.
error variances. The left-hand side of Equation 3 represents the variance-covariance matrix for a given population. In a typical application of CFA, several restrictions are placed on elements of the factor loading matrix \((A_x)\). Without these restrictions on \(A_x\), Equation 3 may be viewed as the statistical model for exploratory factor analysis. When there are several populations, there could potentially be a separate \(\Sigma_x\) for each population. There could also be separate \(\Lambda_x\), \(\Phi\), and \(\Theta_\delta\) matrices for each population.

A Common Factor Model Across Populations

In assessing the existence of a common factor model across populations, several invariance tests are typically performed. Some of these tests and their underlying rationale are described next.

Equality or invariance of \(\Sigma_x\) matrices across populations. Jöreskog and Sörbom (1989) recommended that this test of the equality of the observed variance-covariance matrices be performed before any of the other invariance tests. When the assumption of equal \(\Sigma_x\) is not met, other invariance tests may be performed to pinpoint the source of inequality among \(\Sigma_x\), with increasingly restrictive hypotheses (Jöreskog & Sörbom, 1989, p. 265). Alternatively, when this assumption is met, the populations are treated as equivalent and, therefore, subsequent invariance tests are considered unjustified; data from different groups should be pooled and all subsequent analyses should be based on the pooled variance-covariance matrix. However, there does not appear to be a consensus on the test for the equality of \(\Sigma_x\) matrices, especially in terms of what to do when the equality assumption is satisfied. According to Byrne (1998, pp. 260–261).

Although this omnibus test appears reasonable and is fairly straightforward, it often leads to contradictory findings. For example, sometimes the null hypothesis is found tenable, yet subsequent tests of hypotheses related to the invariance of particular measurement or structural parameters must be rejected (see, e.g., Jöreskog, 1971). Alternatively, the global null hypothesis may be rejected, yet tests for the invariance of measurement and structural invariance hold (see, e.g., Byrne, 1988). Such inconsistencies in the global test for invariance stem from the fact that there is no baseline model for the test of invariant variance-covariance matrices, thereby making it substantively more stringent than is the case for tests of invariance related to sets of parameters in the model (B. O. Muthén, personal communication, October, 1988). Indeed, Muthén contended that the omnibus test provides little guidance in testing for equality across groups and, thus, should not be regarded as a necessary prerequisite to the testing of more specific hypotheses related to group invariance.

Equality of the number of factors and factor pattern matrices. The question of whether or not any of the three matrices \((\Lambda_x, \Phi, \text{ and } \Theta_\delta)\) in Equation 3 is equivalent across populations presupposes that the number of the underlying factors and the factor patterns (where the measured variables are treated as indicator variables for the underlying factors) are the same across populations. Without this requirement, it appears to make little sense to examine the invariance of \(\Lambda_x\), \(\Phi\), and \(\Theta_\delta\) across populations. Importantly, however, Byrne, Shavelson, and Muthén (1989) showed that, given a sufficient number of indicator variables per factor, equivalence of the \(\Phi\) and \(\Theta_\delta\) matrices may still be tested, albeit within the context of partial measurement invariance. Nonetheless, the assumption that the number of factors as well as the factor pattern matrices is similar across populations is considered the baseline model against which the tenability of the other invariance models (described later) is assessed.

Equality of \(\Lambda_x\) matrices across populations (measurement equivalence). One of the central issues in measurement equivalence is whether the extent to which the \(\Lambda_x\) matrices are identical across populations. If they are, Equation 2 will also be identical across populations. That is, two persons from two different populations with an identical vector of factor scores \(s\) will have the same vector of expected raw scores. When there is measurement equivalence, the mean raw scores on the observed variables can be meaningfully interpreted across groups (Drasgow & Kanfer, 1985; Reise, Widaman, & Pugh, 1993). Before providing a detailed account of measurement equivalence in the CFA context with a single underlying factor, we first examine two other types of invariance or equivalence.

Equality of \(\Phi\) matrices across populations (\(\Phi\) equivalence/invariance). The right-hand side of Equation 3 contains two other matrices besides \(\Lambda_x\). As previously noted, \(\Phi\) represents the variance-covariance matrix among the underlying factors. With more than one underlying factor, the equivalence of theoretical structure, as represented by the latent factor correlations, is of interest (Byrne, 1998; Vandenberg & Lance, 2000). Accordingly, both the factor variances and covariances can be tested for their equivalence across groups; however, the latter is of most interest with respect to the equality of theoretical structure. For example, in testing for the validity of a hypothesized four-factor structure of adolescent self-concept, the subscale scores of three different measuring instruments could serve as the indicator variables (i.e., the factor loadings \(s\)) relative to each of the four latent factors. Tests for invariance in this instance, then, focus on the equality of subscale scores (measurement equivalence) as well as on relations among the four self-concept factors (\(\Phi\) or structural equivalence; see, e.g., Byrne & Shavelson, 1987). In substantive research, wherein tests for observed mean differences are of interest, it is important to know that both the item factor loadings and the relations among the latent factors are equivalent across groups. Indeed, as noted by Byrne (1998) and others (Drasgow & Kanfer, 1985; Meredith, 1964), it is quite possible for the items to be equivalent across groups, whereas the relations among the latent factors are not so.

Equality of \(\Theta_\delta\) matrices across populations. The last matrix \((\Theta_\delta)\) in Equation 3 is a diagonal matrix with measurement error variances in the diagonal. The question of whether or not these matrices are identical across populations can provide useful information for practitioners with respect to the reliability of the measuring instruments. More specifically, when the assumption of the equality of \(\Lambda_x\) and \(\Phi\) matrices (strictly speaking, only the equality of \(\Lambda_x\) and factor variances) across populations is satisfied, the equality of \(\Theta_\delta\) matrices would imply the equality of the reliability of measured variables across populations (reliability equivalence; Byrne, 1998). For some investigators, the definition of measurement equivalence may include both the equality of expected raw scores (or the equality of \(\Lambda_x\) as well as the equality of \(\Theta_\delta\) across populations. According to Jöreskog (1971) and Byrne (1998), in multiple indicator CFA models, testing for the equality of reliability is neither necessary nor of particular interest when the observed scale scores are used merely as CFA indicators and not as measures in their own right. As Jöreskog showed, although multiple measures of a latent construct or factor are congeneric, they need not exhibit invariant variances and error/unicovariances.
When assessing for invariance/equivalence across populations, several tests are typically performed following a test for the number of factors and their pattern: Test for the equality of $\Lambda_1$ (Condition C), test for the equality of $\Phi$ (Condition D), and test for the equality of $\Theta_{1s}$ (Condition E). A certain hierarchy is typically observed in conducting these tests (Byrne, 1998; Jöreskog & Sörbom, 1989). For example, a test for the equality of $\Theta_{1s}$ matrices may be done only if the assumption of the equality of $\Lambda_1$ matrices is met (Condition C). Similarly, a test for structural or $\Phi$ equivalence is done when the $\Lambda_2$ matrices are considered comparable across populations (Condition C). For several investigators (Drasgow & Kanfer, 1985; Reise et al., 1993), only Condition C is required for measurement equivalence to hold. This is the definition we adopt in this investigation. (For a more extensive discussion and detailed demonstration of tests for invariance using the CFA approach, readers are referred to Byrne, 1994a, 1994b, 1998, 2001a, 2001b.)

A CFA Model With a Single Underlying Factor

Because the currently popular item response theory (IRT) models assume unidimensionality, a single-factor CFA model is described in detail here to facilitate a comparison of these two methodological approaches (CFA and IRT) with respect to measurement equivalence. Let a scale consist of $n$ items with $m$ response options/categories for each item. Let $\xi$ denote the single underlying or latent construct. As previously noted, within the CFA model, the relationship between the underlying construct and the item score may be expressed as (Bollen, 1989; Byrne, 1994a, 1994b, 1998)

$$x_i = \lambda_i \xi + \delta_i,$$  

$$x_i = \lambda_i \xi + \delta_i,$$  

$$x_n = \lambda_n \xi + \delta_n,$$  

where $x_i$ is the observed score on item $i$, $\lambda_i$ is the factor loading for item $i$, and $\delta_i$ is the residual/error for item $i$. Furthermore,

$$x = x_1 + \cdots + x_n = (\lambda_1 + \cdots + \lambda_n) \xi + (\delta_1 + \cdots + \delta_n),$$

or

$$x = \lambda \xi + \delta,$$

where $x$ represents the total raw score for the subscale, $\lambda = \lambda_1 + \cdots + \lambda_n$, and $\delta = \delta_1 + \cdots + \delta_n$.

**True Score or Expected Raw Score**

According to the CFA model, at the item level,

$$E(x_i) = \lambda_i \xi,$$

and

$$\sigma^2_{x_i} = \lambda_i^2 \sigma^2_\xi + \sigma^2_{\delta_i}.$$

In addition, at the subscale total score level,

$$E(x) = \lambda \xi$$

and

$$\sigma^2_x = \lambda^2 \sigma^2_\xi + \sum \sigma^2_{\delta_i}.$$  

The major assumptions underlying these equations are that (a) the relationship between $x_i$ and $\xi$ is linear and (b) the error ($\delta_i$) has zero expectation and is uncorrelated with $\xi$. The same two assumptions are also needed for $x$, $\xi$, and $\delta$ at the subscale total score level. In view of the second assumption, the residual variance at the subscale total score level is equal to the sum of the item level residual variances or

$$\sigma^2_{\delta} = \sigma^2_{\delta_1} + \cdots + \sigma^2_{\delta_n}.$$  

Several important features are to be noted about the single-factor CFA model. First, the item and the subscale total scores ($x_i$ and $x$) depend only on one latent construct. Second, at the item level, the quantity $\lambda \xi$ is the expected item raw score or item true score. If two persons have the same score on the latent variable, then they will have the same expected raw score or true score on item $i$ (Equation 10) or at the subscale total score level (Equation 12). It should also be noted that the true score at the subscale total score level is simply the sum of item true scores (Equations 6–9). Third, the observed score variance of an item is equal to its true score variance plus the residual or error score variance at the item level (Equation 11) and at the subscale total score level (Equation 13). This additive relationship between the true score variance and the error score variance is identical to a relationship in classical test theory, which states that the observed score variance is equal to the true score variance plus the error score variance.

**Measurement Equivalence Within the Single-Factor CFA Context**

An item or a subscale is said to display measurement equivalence (Condition C) if the item factor loadings are invariant across the two populations (i.e., $\lambda_i = \lambda'_i$, with the prime denoting the second population). This means that two persons, one from each of the two populations, with the same score on the latent variable, will have identical true scores (or expected raw scores) at the item level for all items. In view of Equation 8, if there is true score invariance at the item level for all items, then there must also be true score invariance at the subscale total score level. The converse, however, is not necessarily true. It is theoretically possible for the subscale total factor loadings to be equal ($\lambda = \lambda'$) without identical item level factor loadings across the two populations. Because of their additive nature, it is possible for the item level factor loading differences to cancel each other out at the subscale total score level.

A somewhat stricter definition of measurement equivalence (Conditions C and E) may require that the item-level error variances also be equal across the two populations (i.e., $\sigma^2_{\delta_i} = \sigma^2_{\delta'_i}$ for all $i$s). When this happens and if the factor score variances are also equal, the ratio of item true score variance to the sum of item true score variance and error score variance will be the same for a given item across the two populations. According to classical test theory (Lord & Novick, 1968), this means that item reliabilities will be the same across the two populations. Furthermore, in view of Equations 13 and 14, this stricter definition of measurement equivalence along with the equality of factor score variances requires that the error/residual variances as well as true score variances at
the subscale total score level be equal, thus making the subscale reliabilities the same in the two populations.

**Measurement Equivalence and the Metrics for Measured and Latent Variables**

Thus far, discussion of tests for invariance, within the framework of CFA models, has focused solely on the analysis of covariance structures. As such, all observed scores (i.e., the data) are assumed to represent deviations from their means and, thus, their intercepts are of no interest. Thus, the focus of these analyses centers on only the covariance matrices. However, some authors (e.g., Chan, 2000; Little, 1997) have argued that, in testing for invariance across some groups (e.g., cross-cultural groups), it might be more appropriate to test also for the equivalence of the intercepts. In this instance, then, the observed score means are of interest, and testing for invariance entails the analysis of mean structures as well as covariance structures (typically termed MACS analysis.) Examining equations related to the MACS procedures, note that an equation for item \( i \), although similar to Equations 4 and 5, does not require the assumption of zero means and may be expressed as

\[
x_i = \tau_i + \lambda_i \xi + \delta_i.
\]

where \( \tau_i \) is the intercept for item \( i \). The expectation and variance of \( x_i \) can be expressed as

\[
E(x_i) = \tau_i + \lambda_i \xi
\]

and

\[
\sigma^2_{x_i} = \lambda_i^2 \sigma^2_{\xi} + \sigma^2_{\delta_i}.
\]

It should be noted that Equation 17, as expected, is identical to Equation 11.

The intercept in Equation 16 depends on the mean of \( \xi \) and \( \lambda_i \); in fact, \( \tau_i \) may be expressed as

\[
\tau_i = \mu_{\xi} - \lambda_i \mu_{\lambda_i}.
\]

In this equation, \( \tau_i \) reduces to zero when the means of \( x_i \) and \( \xi \) are equal to zero. An intercept for the same item in a second population (denoted by a prime) may be expressed as

\[
\tau'_i = \mu'_{\xi} - \lambda_i \mu'_{\lambda_i}.
\]

With respect to intercepts, measurement equivalence is defined as the equality of \( \tau_i \) and \( \tau'_i \); that is, the intercepts for the two populations must be identical. The equality of either the factor loadings (\( \lambda_i = \lambda'_i \)) or the equality of means (\( \mu_{\xi} = \mu'_{\xi} \) and \( \mu_{\lambda_i} = \mu'_{\lambda_i} \)) alone will not guarantee the equality of intercepts except when all four means are equal to zero.

The fact that the equality of factor loadings (\( \lambda_i = \lambda'_i \)) does not necessarily imply the equality of intercepts suggests that the item or factor means may be different for the two populations. However, test for the equivalence of latent factor means asks a different question and, thus, is not relevant to the current article. The nonequality of factor means is commonly referred to as impact in the differential item functioning (DIF) literature. For example, Dorans and Holland (1993, pp. 36–37) noted that

It is important to make a distinction between DIF and impact. Impact refers to a difference in performance between two intact groups. Impact is everywhere in test and item data because individuals differ with respect to the developed abilities measured by items and tests, and intact groups, such as those defined by ethnicity and gender, differ with respect to the distribution of developed ability among their members. . . In contrast to impact, which often can be explained by stable consistent differences in examinee ability distributions across groups, DIF refers to differences in item functioning after groups have been matched with respect to the ability or attribute that the item purportedly measures. Unlike impact, where differences in item performance reflect differences in overall ability distributions, DIF is an unexpected difference among groups of examinees who are supposed to be comparable with respect to the attribute measured by the item and test on which it appears.

The pertinent question for us is whether a statistically significant difference in item intercepts across two populations reflects DIF or impact. On the one hand, Equations 18 and 19 depend on item and factor means; therefore, the difference between item intercepts probably reflects impact. On the other hand, the right-hand sides of Equations 18 and 19 seem to imply that the observed item means are adjusted for differences in the factor means for the two populations; therefore, a statistically significant difference in intercepts may reflect DIF/measurement inequivalence or a lack of fit to the hypothesized measurement model. Furthermore, interpretation of a significant intercept difference (main effect) in the presence of a significant slope difference (interaction) may be problematic. Needless to say, there is a definite need for research in this area before concluding whether the difference in intercepts reflects DIF or impact. Given this current uncertainty, only the equality of slopes or factor loadings is used as a condition for measurement equivalence in this presentation, while noting that it can easily be expanded to include the equality of intercepts as some have done (i.e., Chan, 2000). Readers interested in testing for differences in latent factor means are referred to Bollen (1989), Byrne (1998), McDonald (1999), Sörbom (1974), and Vandenberg and Lance (2000) for additional discussion and illustrated applications.

As previously noted, computer programs such as AMOS, EQS, and LISREL can be programmed to test for measurement equivalence across two or more subpopulations simultaneously. (For comparative program approaches, see Byrne, 2001b). Readers interested in a more extensive discussion of this CFA approach are referred to Bollen (1989), Byrne (1994a, 1994b, 1998, 2001a, 2001b), Byrne et al. (1989), Cheung and Rensvold (1999), Drasgow and Kanfer (1985), McDonald (1999), Reise et al. (1993), Reise and Widaman (1999), Rensvold and Cheung (1998), and Vandenberg and Lance (2000).

**IRT Perspective**

Measurement equivalence of items and subscales or tests across subpopulations can also be assessed using the IRT-based techniques developed for studying DIF or item bias. The IRT perspective posits a nonlinear relationship between the underlying/latent construct and the observed score at the item/subscale level. Whereas the CFA perspective, described previously, assumes a linear relationship at the item/subscale level. Before offering an account of the differences and similarities between the CFA and IRT perspectives, the IRT perspective is described. Although it is true that the dichotomous IRT models are currently more popular than the polytomous IRT models, our description of an IRT model is based on Samejima’s (polytomous) graded-response model.
(Samejima, 1969). The reason for our choice is that the polytomous IRT models are increasingly used in practice and that the one-parameter and two-parameter dichotomous IRT models (Hambleton, Swaminathan, & Rogers, 1991) can be viewed as special cases of Samejima’s graded response model. Therefore, the IRT perspective to be presented can be easily specialized to the dichotomous IRT models, especially within the DIF framework.

Several researchers (Bock, 1972; Masters, 1982; Muraki, 1990; Samejima, 1969, 1972) have extended the dichotomous IRT models to the polytomous case. Having such models that incorporate multiple response categories is useful for analyzing Likert-type data. According to Samejima (1969), IRT models are used to determine the parameters of an item based on the responses of individuals to that item. In the case of free-response items, each item could potentially lead to an infinite number of different responses. However, in analyzing such data, it may be more practical to categorize these responses into a limited number of categories that can be ranked ordered in order of attainment or intensity. The most typical way of looking at these multiple categories is in terms of increasing levels of an ability. However, another example is a Likert-type attitude scale in which the response categories can be direct indicators of satisfaction. These response categories can be direct indicators of the participant’s position on the latent construct continuum (Samejima, 1972).

**Boundary Response Functions**

Samejima (1969) defined graded response items as those items that have \( m \) ordered response categories (i.e., for item \( i \) with five response categories, \( m_i = 5 \)). To handle this type of data, Samejima proposed a probability function called the boundary response function (BRF). For an item with \( m_i \) response categories, there will be \( m_i - 1 \) BRFs or cumulative dichotomies, and each BRF is characterized by a discrimination parameter \( (a_i) \) and several location parameters \( (b_{ik}) \). In Samejima’s graded response model, each item will have one discrimination parameter and \( m_i - 1 \) location parameters. According to Samejima, a BRF can be defined as

\[
P_{ik}^* = P_{ik}(\theta) = \frac{1}{1 + e^{-a_i(\theta - b_{ik})}},
\]

where \( P_{ik}^*(\theta) \) is the probability that a randomly chosen examinee with a level of satisfaction \( \theta \) will receive a rating score to item \( i \) with a category score greater than the \( k \)th response category. For an item with \( m_i \) response categories, the \( (m_i - 1) \) BRFs may be expressed as follows:

\[
P_{i1}^* = \frac{1}{1 + e^{-a_i(\theta - b_{i1})}},
\]

\[
\cdots
\]

\[
P_{i(m_i-1)}^* = \frac{1}{1 + e^{-a_i(\theta - b_{i(m_i-1)})}}.
\]

In the equations just given, \( \theta \) refers to the underlying construct such as mathematical ability or job satisfaction. It should be noted that a BRF represents a two-parameter, logistic IRT model for the dichotomous case. If \( a = 1 \), then the BRF becomes the one-parameter or Rasch IRT model (Hambleton et al., 1991).

**Category Response IRT Models**

Samejima (1969) defined graded response items as those items with a level of satisfaction \( \theta \) is the probability that a randomly chosen examinee (with a given \( \theta \)) may be expressed as

\[
t_i(\theta) = (1)P_{i1} + (2)P_{i2} + \cdots + (m_i - 1)P_{im_i} + (m_i)P_{im_i},
\]

which, in view of Equations 23–26, reduces to

\[
t_i(\theta) = 1 - \sum_{k=2}^{m_i} P_{ik},
\]

It should be noted that in Equation 27 the category scores or weights are defined as 1, 2, \ldots, \( m_i - 1 \), and \( m_i \). Equation 28 is only valid for these weights.

**Probability of Responding Above a Category**

![Figure 1. Boundary response functions for an item with five response categories: a = 1.00, b1 = −0.75, b2 = 0.00, b3 = 0.30, and b4 = 1.00.](image)
Equations 33 and 35 are referred to as the item response function (IRF) and test response function, respectively. An example of an IRF is given in Figure 3. It should be noted that it is the same item whose BRFs and CRFs appear in Figures 1 and 2, respectively. Several computer programs are currently available for estimating item and \( \theta \) parameters in the polytomous case, among them the PARSCALE program by Muraki and Bock (1997) and the MULTILOG by Thissen (1991).

**Measurement Equivalence Within the IRT Framework**

Items are said to have measurement equivalence (or non-DIF) if the item parameters remain invariant across the two populations. That is, at the item level,

\[
a_i = a_i', \quad b_i = b_i', \ldots, \quad b_{im} = b_{im'},
\]

where the prime represents the second population. When the item parameters are equal, the BCFs and CRFs are also equal for the two populations. Furthermore, the item true scores (Equation 33) are equal for two persons with identical scores on the latent (ability/satisfaction) variable \( \theta \).

There are several IRT-based DIF procedures: Lord’s (1980) chi-square; Raju’s (1988, 1990) area measures; Thissen, Steinberg, and Wainer’s (1988) likelihood ratio test; and Raju, van der Linden, and Fleer’s (1995) procedure based on the differential functioning of items and tests (DFIT) framework. Lord’s chi-square and Raju’s area procedures were initially proposed for assessing DIF within the dichotomous IRT models, and they were later extended to include the polytomous IRT models by Cohen, Kim, and Baker (1993). Thissen et al.’s and Raju et al.’s procedures are appropriate for assessing DIF with both dichotomous and polytomous scoring (Flowers, Oshima, & Raju, 1999). In addition, Raju et al.’s procedure is also appropriate for multidimensional IRT models (Oshima, Raju, & Flowers, 1997). Many of these procedures assess the invariance of item parameters across two populations, commonly referred to as the focal group and the reference group in the DIF literature. Within the DFIT framework, however, the emphasis is on the invariance of item-level true scores for persons with identical \( \theta \). That is, the IRFs are compared across the two populations for equality. In view of Equations 27, 28, and 33, the invariance of item parameters will lead to the invariance of IRFs. Conversely, the invariance of IRFs will also imply the invariance of item parameters. In other words, the assessment of DIF can be done either at the item parameter level or at the IRF level. For the purpose of this investigation, IRFs are used in assessing measurement equivalence because they offer a

**Figure 2.** Category response functions for an item with five response categories.

**Figure 3.** Item-response-theory-based item response function with five response categories.
natural link to the CFA framework. Before comparing the two approaches for assessing measurement equivalence, a brief description of the DFIT framework is offered.

**DFIT framework for assessing measurement equivalence.** According to Raju et al. (1995), the central theme of the DFIT framework is as follows. First, given a person’s score on the underlying construct \( \theta \), what is that person’s true score on item \( i \) when he or she \( s \) is viewed as a member of the focal group \( (t_{is}) \)? Second, what is the same person’s true score on the same item when he or she is viewed as member of the reference group \( (t_{ir}) \)? In view of Equation 28, the two true scores are identical whenever the focal group and reference group item parameters are identical. That is,

\[
d_{is} = t_{is} - t_{ir} = 0
\]

for all values of \( \theta \). Similarly, the difference in true scores at the sub-scale level for a person \( s \) may be defined as:

\[
D_s = (T_s - T_{ir}) = d_{is} + \cdots + d_{ir}
\]

Measurement equivalence at the sub-scale total score level implies that \( D_s = 0 \) for all values of \( \theta \) or for all persons. Therefore, measurement equivalence within the DFIT framework means that the true score differences are equal to zero at the sub-scale and item levels. Measurement equivalence is always guaranteed whenever the item parameters are equal across the two sub-populations. In practice, the assessment of measurement equivalence within the DFIT framework revolves around the degree to which \( d \) and \( D \) are significantly different from zero. Using Equations 38 and 39, Raju et al. (1995) defined two indices:

\[
NCDIF = E(d^2) = \mu_{\delta} = \sigma^{2}_{\delta} + \mu^{2}_{\theta}
\]

and

\[
DTF = E(D^2) = \mu_{D} = \sigma^{2}_{D} + \mu^{2}_{\theta}
\]

The NCDIF (noncompensatory DIF) index is defined at the item level, whereas the DTF (differential test functioning) index is defined at the sub-scale total score level. According to Equation 35, the NCDIF index reflects the average squared difference between the focal group and reference group item-level true scores. Similarly, according to Equation 36, the DTF index is the average squared difference in true scores at the sub-scale level. For example, in Figure 3, if item true scores are plotted as a function of \( \theta \) separately for the two populations (or for the focal and reference groups), the IRFs must be identical if the NCDIF index is zero. The extent to which the two IRFs differ from each other reflects the degree of DIF at the item level. Similar interpretations are valid for the DTF index. It should be noted that \( \sqrt{NCDIF} \) and \( \sqrt{DTF} \) approximate the average absolute true score difference at the item level and at the sub-scale level, respectively. Additional information about the DFIT framework may be found in Raju et al. (1995).

**Similarities and Differences Between the CFA and IRT Perspectives of Measurement Equivalence**

Several similarities and differences exist between the CFA and IRT perspectives in testing for invariance.

**Similarities**

1. Both perspectives examine the relationship between an underlying construct and a set of measured variables (items/subscale scores to which it is theoretically linked).

2. Both approaches (CFA and IRT) examine the degree to which item/subscale level true scores are similar for persons in the two different populations with the same level of satisfaction/attitude/ability score on the latent construct. This is an important similarity between the two frameworks; it is conceptually akin to a definition of parallelism (equality of true scores) in classical test theory. In classical test theory, one requirement for parallelism calls for the equality of true scores on two tests, whereas here it refers to the equality of true scores across two populations when the latent score is held constant.

3. Both the CFA and IRT definitions of measurement equivalence do not imply that the distributions of scores on the underlying constructs in the two populations of interest are identical. In fact, the latent distributions could be and typically are different. This mean difference between these two distributions is commonly referred to as the impact. The definition of measurement equivalence simply means that persons with identical latent scores will have identical true scores at the item/subscale level, irrespective of the two populations they may happen to represent.

4. When there is measurement nonequivalence, both approaches can be used to identify the extent and the source of the problem. Within the IRT context, individual items are specifically assessed for DIF; this is typical of the IRT-based DIF procedures. However, within the CFA context, the proposed model is tested for its goodness of fit to the data for each group separately before looking for the source of nonequivalence. Presented with findings of nonequivalence, one should examine the item content to possibly determine why it was differentially interpreted by the two populations.

5. In examining measurement nonequivalence, the IRFs can be an invaluable source of information in both the IRT (Equation 28) and CFA (Equation 10) contexts. These graphs are extremely useful for practitioners to identify the extent and the location of measurement nonequivalence for a given item or subscale. If measurement nonequivalence is located at either of the extreme levels of an underlying construct, it may typically call for a different type of a practical solution than when it is located in the middle range of an underlying construct.

**Differences**

1. Whereas the actual relationship between the latent construct and the (item/subscale level) true score is linear in the CFA framework (Equations 10 and 12), it is nonlinear in the IRT framework (Equations 22 and 24). This is an important difference between these two approaches. Within the (dichotomous) IRT context, the probability of answering an item correctly \( P \) is expressed as a logistic function, which is by definition nonlinear. However, the logarithm \( \ln \) of an odds ratio—\( \ln(P/(1 - P)) \) is linear even in the IRT case. McDonald (1999) provided additional information about the linear structures underlying the IRT (dichotomous) and CFA models. It should be noted that the linearity embodied in \( \ln(P/(1 - P)) \) for the dichotomous case does not readily generalize to the polytomous IRT models.

2. When measured or indicator variables are dichotomously scored (e.g., multiple-choice items in an achievement or ability test), a logistic regression model is considered more appropriate for expressing the relationship between a continuous underlying construct and a measured variable than a linear regression model. In this context, the IRT-based approach for measurement equiva-
lence is preferable over the CFA approach. However, as the number of possible scores for an item or any measured variable increase, the linear regression model used in the CFA approach will be equally tenable.

3. The CFA methodology is currently well advanced to handle multiple latent constructs and multiple populations simultaneously. On the other hand, much of the IRT-based DIF or measurement equivalence methodology is confined to unidimensional scales. Given the advances by Kim, Cohen, and Park (1995) for multiple group DIF analysis and by Oshima et al. (1997) for multidimensional DIF, it should be possible to expand the IRT-based methodology for simultaneously assessing measurement equivalence in multiple groups across several latent constructs.

4. In a stricter form of measurement equivalence within the CFA framework, item error variances are required to be equal across populations. It should be noted that, whereas Jöreskog initially proposed testing for the invariance of measurement error, the requirement of the equality of the item residual variances is very stringent and not realistic in most practical situations (Byrne, 1994a, 1994b, 1998, 2001a). In the IRT framework, there has not been much explicit discussion of item-level error variance because it is conditional on \( \theta \), and in the dichotomous case, this variance can be expressed as \( P_i(0)(1 - P_i(\theta)) \) for item \( i \), where \( P_i(\theta) \) represents the probability of answering item \( i \) correctly given \( \theta \). The concept that has received a great deal of attention in IRT is the standard error of measurement associated with an estimate of \( \theta \). The fact that it varies as a function of \( \theta \) means that the standard error of measurement could vary from person to person. This is viewed as an important benefit of IRT over classical test theory (Hambleton et al., 1991). It is possible that the average of residual variances across persons within one population may equal the average of residual variances in another population, but this relationship has not been actually explored in the IRT context. The equality of such average residual variances simply means that the group-based standard errors of measurement are equal across populations.

5. Equations 10 and 27, respectively, show the relationship between the underlying construct and the item true score in the CFA and IRT contexts. In addition, within the IRT context, the probability of choosing any one of the categories in an item is also known for a person with a given \( \theta \); Equations 23 to 26 provide this information. Such probabilities are not readily available within the CFA framework.

6. The compensatory nature of DIF, at the subscale level, is clearly anticipated and addressed in the IRT context (Raju et al., 1995). This aspect of measurement nonequivalence does not appear to have received much attention in the CFA context with the possible exception of some work by McDonald (1999) about DTF in the CFA context. Nonetheless, the partial measurement equivalence definition (see Byrne et al., 1989) does allow for both item- and subscale-level factor loadings to be partially different across the two populations. For example, in practice, it is possible for two items to exhibit measurement nonequivalence. One item may exhibit measurement nonequivalence in one direction in the first population, and the other item may exhibit measurement nonequivalence in the opposite direction in the second population such that the sum of the two item-level factor loadings is the same across the two populations. As noted by Raju et al. (1995), such information is useful for practitioners who are constantly faced with the problem of what to do with items that exhibit measurement nonequivalence, especially when the reasons for measurement nonequivalence are not readily apparent.

Example

This section offers an empirical example to illustrate the two approaches. The CFA procedure, with its sequential and up-and-down analysis for assessing the equality of factor loadings, is different from what is typically seen in the literature. Therefore, this example is designed to provide clarity for the various computational steps involved in the CFA procedure, as well as in the IRT procedure, for assessing measurement equivalence.

Collins, Raju, and Edwards (2000) examined DIF using a 10-item scale with five response categories per item, designed to assess satisfaction in work assignment. These items were taken from the 1995 Armed Forces Sexual Harassment Survey (Edwards, Elig, Edwards, & Riemen, 1997). Collins et al. used the polytomous IRT framework to examine DIF between Blacks and Whites and between males and females. Two samples (one Black and one White), similar but not identical to those in the Collins et al. study, are presented here to illustrate both (CFA and IRT) perspectives for assessing measurement equivalence. Each sample consists of 1,000 participants. Additional information about the complete Armed Forces Sexual Harassment Survey and the two samples used in this example can be found in Edwards et al. and Collins et al.

CFA Perspective

To establish a baseline model for purposes of testing equivalence of the 10-item scale across Blacks and Whites, goodness of fit related to a one-factor CFA model was tested simultaneously across groups, with no equality constraints imposed. Essentially, the chi-square statistic related to this simultaneous model represents the sum of the chi-square values resulting from test of model fit for each group separately. Thus, as indicated in Table 1, with the exception of Item 10, all factor loading parameters were freely estimated for Black and Whites. Results bearing on this baseline model are presented in Table 1. Indeed, examination of the goodness-of-fit statistics suggests a well-fitting model for both Blacks and Whites. (For a more extensive and nonmathematical discussion of goodness-of-fit statistics, see Byrne, 1994a, 1994b, 1998, and 2001a.)

Beyond the fit of the model, as a whole, we can also examine the estimated values of the factor loadings, which, of course, are of primary interest. These results are presented in Table 2. With the exception of Items 1 and 2, a review of these estimates reveals all items to be very similar across Blacks and Whites. However, such similarity is no guarantee that the items are being perceived equivalently by both groups. Such equivalence must be tested statistically. Our next task, then, was to test for the equivalence of all items comprising this one-factor scale across Blacks and Whites. Results of these analyses are summarized in Table 1.

In testing for the equivalence of scale items, our first analysis involved comparison of fit between a model in which all factor loadings were constrained to be equal across groups (Model 2) against another model in which they were freely estimated (Model 1). In large samples, the difference in chi-square values is distributed as a chi-square statistic, with degrees of freedom equal to the difference in degrees of freedom between the two models. As
shown in Table 1, this comparison between Models 1 and 2 yielded \( \Delta \chi^2(9, N = 1,000) = 35.35, p < .01 \) (statistically significant). Presented with these findings of noninvariance, the next task was to identify which items in the scale were operating differently across Black and Whites. Thus, following the conventional strategy used in the univariate approach by the LISREL program (and also AMOS) in testing for such invariance (see Byrne, 1998, 2001a), the next model tested was one in which only Item 1 was constrained to be equal across groups. As indicated in Table 1, comparison of this model (Model 3) with Model 1 also yielded a statistically significant difference, \( \Delta \chi^2(1, N = 1,000) = 12.82 \). Likewise, a model in which Item 2 was constrained to be equal across groups (Model 4) was found to be significantly different from Model 1. Thus far, then, we can conclude that Items 1 and 2 were not operating equivalently across Blacks and Whites.

Continuing on in this manner, we tested next for the equivalence of Item 3 across groups. As shown in Table 1, comparison of this model (Model 5) with Model 1 resulted in no statistically significant difference, thereby indicating the equivalence of Item 3 across Blacks and Whites. Given this finding of invariance, the next model tested for the invariance of Item 4 while concomitantly constraining both Items 3 and 4 to be equal across the groups. Likewise, for the remaining tests, presented with evidence of invariant items, their related factor loadings were constrained to be equal, cumulatively, across groups. For example, given findings of invariance related to Items 3 through 8, the specification of Model 11 (which tested for the invariance of Item 9) called for equality constraints to be placed on Items 3 through 9. On the basis of the findings summarized in Table 1, we can conclude that, except for Items 1 and 2, the remaining items comprising this 10-item, one-factor scale were operating equivalently across Blacks and Whites.

Figures 4 and 5 show the relationship between the underlying construct and the true score or expected raw score on Item 1 and Item 2, respectively (Equation 10). The content of Item 1 was “I have been taught valuable skills in the service that I can use later in civilian jobs” and the content of Item 2 was “I am glad that I was assigned to this organization.” Both items had the same five response options: strongly disagree, disagree, neither agree nor disagree, agree, and strongly agree. According to the linear item response function for Item 1, displayed in Figure 4, a Black person with a low factor score is more likely to endorse a higher category (item) response than a White person with the same low factor score. For example, a person with a factor score of \(-2.5\) is more likely to pick Category 3 (“neither agree nor disagree”) if he or she is Black or to select Category 2 (“disagree”) if he or she is White. That is, a Black person at the low end of the satisfaction scale is more likely to conclude that the skills taught in the service may be useful later in civilian jobs.

Even though the factor loadings (\( \lambda_s \)) are also different for Item 2 (see Table 1), the linear item response functions, displayed in Figure 5, do not appear to be substantially different. The measurement nonequivalence for Item 2, although statistically significant,
may be less relevant from a practical point of view. A replication of the current illustration, from a different perspective, may help clarify the status of measurement nonequivalence for Item 2.

**IRT Perspective**

Samejima’s graded response model (Samejima, 1969) and the PARSCALE computer program (Muraki & Bock, 1997) were used for the DIF and DTF analyses. The item parameters for the White group were put on the same scale as the one underlying the Black group, using the EQUATE program (Baker, 1995). The polytomous item parameter estimates (one a and four b parameters) from PARSCALE and the NCDIF indices from the DFIT framework are shown in Table 3. It should be noted that the White item parameter estimates are the equated item parameters. According to the DFIT framework, only Item 2 showed significant DIF; that is, the NCDIF index of .144 was statistically significant at the .01 level and greater than the cutoff of .096 for a five-category item (Raju, 1999). The DTF of .011 was not statistically significant. The finding of one item (Item 2) with significant DIF in the IRT context is at variance with the CFA results in which two items (Item 1 and Item 2) were shown to have measurement nonequivalence. Furthermore, the measurement nonequivalence noted for Item 2 in CFA did not appear to be all that substantial.

The BRFs, CRFs, and IRFs for Item 2 from the Black-White comparison are shown in Figures 6, 7, and 8, respectively. Figure 6

![Figure 6](image_url)

**Figure 6.** Confirmatory factor analysis (CFA)-based item response functions for Item 1 for the Black and White groups.

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b₁</th>
<th>b₂</th>
<th>b₃</th>
<th>b₄</th>
<th>NCDIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>0.537</td>
<td>-3.739</td>
<td>-2.814</td>
<td>-1.907</td>
<td>0.806</td>
<td>0.000</td>
</tr>
<tr>
<td>Item 2</td>
<td>0.557</td>
<td>-3.816</td>
<td>-2.603</td>
<td>-1.769</td>
<td>0.726</td>
<td>0.144</td>
</tr>
<tr>
<td>Item 3</td>
<td>0.671</td>
<td>-1.566</td>
<td>-0.715</td>
<td>0.363</td>
<td>1.825</td>
<td>0.015</td>
</tr>
<tr>
<td>Item 4</td>
<td>0.913</td>
<td>-1.744</td>
<td>-0.978</td>
<td>-0.237</td>
<td>1.159</td>
<td></td>
</tr>
<tr>
<td>Item 5</td>
<td>0.605</td>
<td>-2.508</td>
<td>-1.091</td>
<td>0.935</td>
<td>5.359</td>
<td></td>
</tr>
<tr>
<td>Item 6</td>
<td>0.658</td>
<td>-2.014</td>
<td>-0.883</td>
<td>1.132</td>
<td>5.376</td>
<td></td>
</tr>
<tr>
<td>Item 7</td>
<td>1.021</td>
<td>-1.546</td>
<td>-0.528</td>
<td>0.836</td>
<td>3.922</td>
<td></td>
</tr>
<tr>
<td>Item 8</td>
<td>1.199</td>
<td>-1.482</td>
<td>-0.410</td>
<td>0.967</td>
<td>4.208</td>
<td></td>
</tr>
<tr>
<td>Item 9</td>
<td>0.892</td>
<td>-1.697</td>
<td>-0.707</td>
<td>0.805</td>
<td>4.182</td>
<td></td>
</tr>
<tr>
<td>Item 10</td>
<td>1.020</td>
<td>-1.359</td>
<td>-0.407</td>
<td>1.005</td>
<td>4.308</td>
<td></td>
</tr>
</tbody>
</table>

Table 3

**Estimated Item Parameters and Noncompensatory Differential Item Functioning (NCDIF) Indices**

shows the two sets of four boundary response functions for Item 2 in the Black-White comparison. The fact that the curves for each pair of boundary functions cross is evidence that the a parameters are different for the groups (a₂ = .671 for the Black group and a₂ = .913 for the White group). That is, the slopes of the curves for the two groups are not equal. In addition, differences in the b parameters are also evident. The b parameters for the Black group are b₁w = -1.566, b₂w = -0.715, b₃w = 0.363, and b₄w = 1.825. Whereas the b parameters for the White group are b₁w = -1.744, b₂w = -0.978, b₃w = -0.237, and b₄w = 1.159.

![Figure 7](image_url)

**Figure 7.** Shows the category response functions for Item 2, confirms the trend noted in Figure 6. Figure 8 shows the IRFs for the Black and White groups.

The IRF graphs are particularly informative in delineating the location of differential functioning and its extent across the satisfaction scale. Specifically, the NCDIF index for Item 2 is 0.144, the square root of which is 0.379. As previously noted, 0.379 approximates the average absolute difference in the expected item raw scores for Black and White examinees with the same score on the satisfaction subscale. That is, if the satisfaction level is the same for a Black examinee and a White examinee, then their expected raw scores (or true scores) on Item 2 could, on the average, differ by 0.379 in absolute terms. Because 0.379 is the average difference expressed in absolute terms, the item true score difference and its direction (or sign) could vary as a function of the level of satisfaction. A graphic view of this practically relevant information can be gleaned from the IRFs in Figure 8. For example, a person with a level of satisfaction...
at $\theta = 2.0$ may select either Category 4 (agree) or Category 5 (strongly agree) depending on whether the person is Black or White. If the person is Black, he or she is more likely to give a Category 4 response, whereas a White person is more likely to give a Category 5 response; thus, Item 2 favors White individuals at this level of satisfaction. At the low end of the satisfaction subscale, the opposite appears to be true (favors Blacks) but to a lesser extent.

In addition, the (approximate) average absolute difference ($\sqrt{NC\text{DIF}}$) may be useful in assessing whether a statistically significant NCDIF is also practically significant or meaningful. If a 5-point item has a significant, average absolute difference of 0.25, with most of the difference coming from an extreme end of the satisfaction subscale, should 0.25 be taken seriously from a practical perspective? The answer to this question will depend on, among other things, how the information from this scale is used in practice. The IRFs in Figure 8 can be very useful in such a context.

Figure 9 shows the IRT-based IRFs for Item 1. Figure 9 is presented here to illustrate some of the differences between the IRT-based and CFA-based perspectives. The Black and White IRT-based IRFs are quite similar in Figure 9, whereas the CFA-based IRFs in Figure 4 are substantially different for the Black and White groups. Facteau and Craig (2001), Maurer, Raju, and Collins (1998), and Laffitte, Raju, Scott, and Fasolo (1998) also reported some measurement nonequivalence results that were not consistent across the two perspectives.

Based on the results from the current illustration, it is impossible to state unequivocally which of the two conclusions (measurement nonequivalence for Items 1 and 2 in the CFA context or significant DIF for Item 2 in the IRT context) is on target. The fact that Item 2 showed measurement nonequivalence and that eight other items showed measurement equivalence in both contexts is encouraging, indicating, at least in this example, a high degree of congruence between the two methods. To assess the relative accuracy of these two methodologies, a comprehensive Monte Carlo study, with some of the highlighted differences (e.g., dichotomous vs. polytomous data, sample size, the degree to which the underlying assumptions are met) between the two models as moderators, will be needed. Such an investigation must allow for data generation from both models and an evaluation of each model with the data generated from the other model. We encourage interested researchers to undertake such an investigation.

Because CFA is linear and IRT is nonlinear, it is likely that the CFA-based perspective would result in more instances of measurement nonequivalence in practice than the IRT-based perspective. The current example appears to substantiate this hypothesis. Again, a comprehensive Monte Carlo study may help evaluate this hypothesis. It should be noted that the current example is included here only to illustrate some of the measurement nonequivalence and differential functioning concepts articulated in this article. This example is not intended for assessing the relative merits of the IRT-based and CFA-based methodologies for studying measurement nonequivalence or differential functioning.

In summary, we have tried to offer a comparison of two currently popular approaches for assessing measurement nonequivalence in this research note. Both approaches test for the equality of item level and subscale level true scores for persons with identical latent scores, irrespective of group membership. The CFA approach hypothesizes that the relationship between the underlying construct and true score (at the item and subscale levels) is linear, whereas with the IRT approach, the same relationship is assumed to be nonlinear. In both cases, measurement equivalence simply means that these relationships are invariant across subpopulations. Some other differences and similarities between the two CFA and
IRT approaches are also noted. Important questions such as which method is better and the comparability of results should be examined more comprehensively in future investigations.

References


Received May 12, 2000
Revision received September 11, 2001
Accepted September 13, 2001