Running head: OBSERVED AND UNOBSERVED MEASUREMENT EQUIVALENCE

# Using mixed-measurement item response theory with covariates (MM-IRT-C)

# to ascertain observed and unobserved measurement equivalence

Louis Tay<br>
Department of Psychology<br>
University of Illinois at Urbana-Champaign<br>
603 East Daniel Street<br>
Champaign, IL 61820<br>
Email: sientay@illinois.edu<br>
Phone: 217-721-8587

Daniel A. Newman<br>
Department of Psychology<br>
University of Illinois at Urbana-Champaign<br>
603 East Daniel Street<br>
Champaign, IL 61820<br>
Email: d5n@illinois.edu

Jeroen K. Vermunt<br>
Department of Methodology and Statistics<br>
Faculty of Social and Behavioural Sciences<br>
Tilburg University<br>
PO Box 90153<br>
5000 LE Tilburg, The Netherlands<br>
Email: j.k.vermunt@uvt.nl

## Abstract

Traditional item response theory (IRT) measurement invariance approaches examine measurement equivalence (ME) between observed groups (e.g., race, gender, culture). By contrast, mixed-measurement item response theory (MM-IRT) ascertains ME among unobserved groups (i.e., *latent classes* [LC] of respondents distinguished by differences in scale use). Both approaches can be integrated by using the MM-IRT-C model, in which covariates (i.e., observed characteristics) are modeled in conjunction with LCs, thereby elucidating if ME is attributable to observed and/or unobserved groupings. An advantage of the technique is that it can be used to ascertain ME over multiple observed characteristics (categorical and/or continuous) concomitantly. In general, the MM-IRT-C can serve several purposes: (a) infer underlying latent measurement classes (LCs), (b) determine associations of LC membership with observed characteristics, and (c) determine if observed measurement nonequivalence occurs predominantly within a particular latent measurement class. This method is illustrated using a measure of union citizenship behavior, with years of work experience and gender as covariates. The substantive and methodological contributions of this model for rethinking ME and its use in organizational research are discussed.

# Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence

Observed groupings (e.g., race, gender, culture) have been integral in the analysis of measurement equivalence (ME). Their use is generally driven by theoretical/practical/legal concerns over whether subgroups employ the same frame-of-reference on a measure of interest (Riordan & Vandenberg, 1994; Vandenberg & Lance, 2000) and whether scores on the measure are comparable across groups (Drasgow, 1987; Stark, Chernyshenko, & Drasgow, 2004). However, focusing on differences in scale use across *observed groups* is not particularly informative as to whether latent differences in scale use exist (e.g., response sets/styles, see Eid & Rauber, 2000). In contrast to the *observed ME* approach, mixed-measurement item response theory (MM-IRT) (Mislevy & Verhelst, 1990; Rost, 1990, 1991) focuses on *unobserved ME* by identifying *latent classes* of individuals who use scale items in a distinct manner when responding to psychological measures (e.g., Hernandez, Drasgow, & Gonzalez-Roma, 2004; Zickar, Gibby, & Robie, 2004).

In this paper, we present the use of MM-IRT with covariates (MM-IRT-C) (see also Maij-de Meij, Kelderman, & van der Flier, 2008; Smit, Kelderman, & van der Flier, 1999, 2000) as a method for examining both observed and unobserved ME. By employing a latent class measurement model in which observed groupings are simultaneously modeled as covariates, we advance an integrated framework for assessing both observed and unobserved ME in organizational research. At this juncture, we clarify some terminology: in the IRT literature, measurement nonequivalence is also referred to as differential item functioning (DIF; for further elaboration, see Stark, Chenyshenko, & Drasgow, 2006; Vandenberg & Lance, 2000). The conceptual differences between observed and unobserved DIF are delineated in Table 1. This table not only serves to show how different ME procedures detect observed or unobserved DIF, it also conveys the key notion that differences in scale use may be a function of both observed and unobserved individual characteristics (e.g., Cohen & Bolt, 2005; De Ayala et al., 2002; Maij-de Meij et al., 2008).

-- insert Table 1 about here --

### *Conceptual Presentation of the MM-IRT-C Model and its Precursor Models*

The mixed-measurement item response theory with covariates (MM-IRT-C) model extends IRT-DIF approaches commonly used by organizational researchers in several ways, as described in Table 1. A restricted form of the MM-IRT-C model can be used to model multiple covariates, enabling testing of uniform DIF (item difficulty/location) and non-uniform DIF (item discrimination) on *multiple observed characteristics/groupings simultaneously*. Also, because continuous covariates can be used (e.g., work experience), partitioning individuals into dichotomous groups (e.g., less work experience vs. more work experience) for the purposes of testing observed ME is unnecessary. Further, the use of the MM-IRT-C model in general has several additional advantages over conventional IRT-DIF approaches: (1) unlike traditional tests for observed group ME, it is not assumed here that the same measurement model necessarily holds for *all* individuals within each observed group; (2) not only can unobserved differences in scale use be ascertained, we can also determine how the latent trait standing *within* each latent class is related to observed characteristics of individuals; (3) we can assess if observed DIF occurs within a LC of individuals after taking into account latent measurement differences; thus determining not only *if* DIF occurs, but also *for whom* DIF occurs. Table 2 presents a summary of these pertinent issues and their applicability to exemplar organizational topics such as research on aging, cross-cultural comparisons, and diversity.

-- insert Table 2 about here --

We note that the MM-IRT-C model is not a new model (see Maij-de Meij, Kelderman, & van der Flier, 2008; Smit, Kelderman, & van der Flier, 1999, 2000). In this paper, however, we not only review the utility of this model, but also extend past applications of it. In our current presentation of the MM-IRT-C model, we provide the following methodological extensions: (a) we show how to test for *class-specific* covariate effects within each latent class; that is, testing for whether the *latent traits in each class* are related to external covariates (e.g., gender and work experience), (b) we propose steps for testing *whether DIF occurs between latent classes* (i.e., testing whether latent classes have full measurement nonequivalence, or partial measurement nonequivalence by testing for both uniform DIF [differences in item locations] and non-uniform DIF [differences in item discriminations]), (c) we show how to test whether observed DIF occurs in only a subset of individuals (i.e., examining if observed DIF occurs in only one latent class [LC] and not another LC), and (d) given at least partial measurement invariance between LCs, we can test whether latent scores differ between LCs.

The structure of the paper is as follows. First, we present the statistical underpinnings of the MM-IRT-C model, incorporating our proposed extensions. Second, we offer an empirical illustration of the MM-IRT-C model on a measure of union citizenship behavior with the observed characteristics years of work experience and gender as covariates, using the software Latent GOLD 4.5 (Vermunt & Magidson, 2008). Practical issues in correctly specifying the MM-IRT-C model are discussed. Finally, we propose how MM-IRT-C can foment new avenues for theoretical and methodological research within organizational science.

### *Mathematical Presentation of the MM-IRT-C Model and its Precursor Models*

In this section, we clarify the primary differences among observed, unobserved, and overall DIF, explaining the limitations of traditional IRT DIF methods used by organizational researchers and how MM-IRT-C can expand the conceptualization and testing of DIF.

*IRT and Observed DIF.* IRT expresses the mathematical relationship between the latent trait level $\theta_j$ and the probability of item endorsement (see Hulin, Drasgow, & Parsons, 1983). In this paper, a 2-parameter logistic (2PL) model following the parameterization in Latent GOLD

4.5 (Vermunt & Magidson, 2008) is utilized. Let $y_{ji}$ denote the response of individual *j* on questionnaire item *i*; the probability of item endorsement is then

$$P(y_{ji} \mid \theta_j) = \frac{1}{1 + \exp(-[a_i\theta_j + b_i])}, \tag{1}$$

where $a_i$ and $b_i$ represent the item discrimination and item location respectively. Observed DIF occurs when the expected score given the same latent score $\theta_j$ is different by virtue of observed group membership (*z*; see Table 1). If so, the measurement model for the item differs between observed groups and each observed group has its own unique item discrimination $a_{iz}$ (representing *non-uniform DIF*) and/or item location $b_{iz}$ (representing *uniform DIF*). In traditional IRT-DIF procedures, these differences in item parameters (i.e., discriminations and locations) are tested simultaneously (Lord, 1980), or indirectly examined via differences in observed group item response functions (Raju, 1988). However, the limitation of these approaches is that (a) multiple observed characteristics cannot be tested for DIF simultaneously, (b) continuous observed characteristics cannot be directly utilized, and (c) testing of DIF in multiple groups (>2) is often engaged in a "piece-meal" fashion (i.e., multiple two-group comparisons are needed) (see Hambleton & Kanjee, 1995). We note that in view of this, recent research has worked toward developing tests of equivalence across multiple groups in the IRT framework (e.g., Kim, Cohen, & Park, 1995; Penfield, 2001).

In contrast to traditional DIF detection strategies, we propose a *single class/restricted MM-IRT-C* approach (IRT-C)[1], which is a procedure akin to the logistic regression (LR) method for testing observed DIF (Swaminathan & Rogers, 1990), but uses the latent trait score $\theta_j$ (as shown in equation 1) rather than the observed total score $X_j$. This is an IRT counterpart of the Multiple-Indicator-Multiple-Cause (MIMIC) model used for detecting DIF items within the factor analytic (FA) or structural equations modeling (SEM) literature (Woods, Oltmanns,

& Turkheimer, 2009). In the single class/restricted MM-IRT-C (IRT-C) approach, we can examine both uniform and non-uniform DIF, whereas the MIMIC model only allows testing of uniform DIF (see Woods, 2009).

This model is graphically depicted in Figure 1A, where differences in item discrimination and item location can be tested via paths 2 and 3 respectively; thus we can determine if there are significant differences in the item discriminations (corresponding to non-uniform DIF) and item locations (corresponding to uniform DIF) among observed groups (Swaminathan & Rogers, 1990). Further, because a vector of observed characteristics $z_j$ (either continuous or nominal) can be used, it overcomes the three specific limitations of traditional IRT-DIF procedures described above. In the current case, the depicted associations (paths 2 and 3 in Figure 1A) demonstrate differences in the probability of responding by virtue of being in a different observed group (e.g., gender) or having different levels on a continuous observed characteristic (e.g., work experience), which is a standard definition for observed DIF (Drasgow, 1987). Statistically, the conditional probability of the item response is extended from $P(y_{ji} | \theta_j)$ to $P(y_{ji} | \theta_j, z_j)$. The latter is modeled using a logistic function but with an additional term $c_i z_j$ for uniform DIF and $d_i z_j \theta_j$ for non-uniform DIF.

-- insert Figure 1 about here --

Path 1 in Figure 1A shows that differences in latent trait levels among observed groups/characteristics are modeled in the testing of DIF within IRT-C. Thus, metrics of the observed groups are implicitly taken into account and it is not necessary to undertake IRT linking procedures among the different observed groups. This is in contrast to the IRT-DIF techniques that require linking: Lord's $\chi^2$ (1980) test and differential functioning of items and tests (DFIT) methodology (Raju, 1988) -- two methods that are commonly used by organizational researchers (e.g., Collins, Raju, & Edwards, 2000; Raju, Laffitte, & Byrne, 2002;

Stark, Chernyshenko, Chan, Lee, & Drasgow, 2001; Stark et al., 2004). That is, the IRT-C model avoids the conventional multi-step approach where item parameters are first calibrated separately between groups, and then a second step of equating is undertaken before DIF is tested.

To elaborate, observed score differences may be alternatively attributable to either DIF (i.e., *bias*; denoted by Paths 2 and 3) or to actual latent trait differences between observed groups (i.e., *impact*; denoted by Path 1). IRT-C models the contributions of both bias and impact on observed scores. Nevertheless, observed groupings may not be the only source of bias or impact. For instance, it may not be known a priori which subpopulation of individuals uses the scale in a distinct manner, or has a different mean level on the latent trait. Because such unseen differences may not map tidily onto observed characteristics (e.g., gender, race, country membership), one can alternatively infer unobserved groups via the MM-IRT model.

*MM-IRT and Unobserved DIF.* The mixed measurement item response theory model (MM-IRT; see Figure 1B) may be viewed as an extension of the IRT model, where latent classes (*k*) of individuals underlie the set of observed responses. Thus, the conditional probabilities of responding to each item become $P(y_{ji} \mid k, \theta_j)$, which depends not only on the latent trait standing ($\theta_j$) but also on the latent class (*k*),

$$P(y_{ji} \mid k, \theta_j) = \frac{1}{1 + \exp(-[a_{ik}\theta_j + b_{ik}])}. \tag{2}$$

Let $\underset{\sim}{y}_j$ be the vector containing all *I* item responses (*i*= 1,…, *I*); the MM-IRT model is then

$$P(\underset{\sim}{y}_j) = \sum_{k=1}^{K} \pi_k \int \prod_{i=1}^{I} P(y_{ji} \mid k, \theta_j) f(\theta_j) d\theta_j, \tag{3}$$

where the unconditional class membership probabilities $\pi_k$ serve as weights and sum to one, $\sum_{k=1}^{K} \pi_k = 1$; $f(\theta_j)$ is taken as the standard normal density and $d\theta_j$ represents the latent trait over which the integration is performed. Similar to Figure 1A, item parameters may have distinct item discriminations or locations across the unobserved groups (or LCs), indicating unobserved DIF. Unobserved DIF can be tested across LCs in the MM-IRT framework via paths 5 and 6 in Figure 1B; significant effects indicate significantly different item discriminations and item locations respectively. As defined in Table 1, if unobserved DIF occurs, differences in expected observed scores occur for the same latent trait standing by virtue of latent group membership.

*MM-IRT-C and Overall DIF.* Instead of using a two-step procedure where MM-IRT LCs are first obtained and then associated with other external variables (e.g., Eid & Rauber, 2000; Hernandez et al., 2004; Zickar et al., 2004), we can model the associations of LCs with external observed characteristics within a single, integrated model. Figure 1C presents the graphical depiction of the MM-IRT with a covariate, or observed characteristic, *z*. From the model, we see that the association between inferred latent classes and an external covariate (*z*) is modeled by path 7; that is, observed group membership may predict latent class membership. For multiple covariates (*p*=1,…,*P*), the MM-IRT-C model may be written as

$$P(\underset{\sim}{y}_j \mid \underset{\sim}{z}_j) = \sum_{k=1}^{K} \pi_{k|\underset{\sim}{z}_j} \int \prod_{i=1}^{I} P(y_{ji} \mid k, \theta_j) f(\theta_j \mid \underset{\sim}{z}_j) d\theta_j \, , \tag{4}$$

where $\underset{\sim}{z}_j$ is the covariate vector (nominal or continuous) for the $j^{\text{th}}$ person. As can be seen, the class membership probabilities are assumed to be affected by the covariates, which is typically modeled by specifying a logistic regression model for $\pi_{k|z_j}$. That is,

$$\pi_{k|z_j} = \frac{\exp(\alpha_k + \sum_{p=1}^{P} \beta_{pk} z_{jp})}{\sum_{k'=1}^{K} \exp(\alpha_{k'} + \sum_{p=1}^{P} \beta_{pk'} z_{jp})}, \tag{5}$$

where $\alpha_k$ and $\beta_{pk}$ are the intercept and slope coefficients, respectively, for LC *k*. Based on the coefficients $\beta_{pk}$, the statistical significance of the covariate $z_{jp}$ predicting the LC proportions can be examined.

Further, one can look at the covariate distributions *within classes* by aggregating and rescaling the posterior class membership probabilities (for more details see Vermunt & Magidson, 2005, p. 70), which makes it possible to compare proportions for dichotomous covariates (e.g., gender) and compare means for continuous covariates (e.g., work experience) across the LCs. As in the IRT-C model, the latent trait $\theta_j$ is regressed onto the covariates using a linear regression model and is reflected in the $f(\theta_j \mid z_j)$ term (Figure 1C path 1; see also Maij-de Meij et al., 2008). Here we go beyond past applications and show how to ascertain class-specific effects; that is, testing whether the latent trait *within each LC (* $\theta_{jk}$ *)* is associated with the covariates. It is possible that there is a relationship between the covariate and the latent trait in some unobserved subgroups, but not in others.

Additionally, we show how the MM-IRT-C model can be used to examine *overall DIF* – both observed and unobserved. Unlike IRT and MM-IRT approaches, which independently investigate observed and unobserved DIF respectively, the MM-IRT-C model can be used to analyze both types of DIF synchronously. There are several important reasons for doing so. First, we can test if the occurrence of *observed DIF may be attributable to more nuanced, unobserved DIF*. As an example, one may obtain observed DIF on gender, but these DIF effects may be accounted for by three LCs with unobserved DIF on the item of interest. For instance, a majority of individuals (70%) consisting of equal proportions of males and females may in fact share the same measurement model. However, 20% of individuals, primarily female (80%), may exhibit a distinct frame-of-reference; while the remaining 10% of individuals, primarily male (85%), use the scale in yet another way. In effect, we can show that the two smaller LCs may account the observed group DIF, even though most males and females may share the same measurement model. Simply applying traditional observed ME approaches may lead to stereotypical views that *all* males and *all* females use the scale differently (e.g., Cohen & Bolt,

2005) and we elaborate on this more fully in the discussion section. A second reason for using MM-IRT-C is that DIF on multiple observed characteristics may be accounted for by unobserved measurement groups. Third, because not all observed DIF can be accounted for by unobserved measurement groups, it is necessary to examine residual observed DIF beyond that of unobserved DIF. Conceptually, unobserved differences (i.e., latent class differences) may not fully demarcate how individuals differentially use a scale; such differences may be attributable to observed group membership (see Figure 1C, paths 2 and 3). In this case, the conditional probability $P(y_{ji} \mid k, \theta_j)$ in equation 2 becomes $P(y_{ji} \mid k, \theta_j, \underset{\sim}{z}_j)$. We show how to examine if residual observed DIF occurs in specific latent classes of individuals. In effect, we present a broad framework for examining various types of DIF effects with the MM-IRT-C model.

*Empirical Example of MM-IRT-C: Union Citizenship Measure with Years of Work Experience and Gender as Covariates*

Using an empirical example, we next show how the MM-IRT-C framework can be used to examine overall DIF. First, we demonstrate the restricted MM-IRT-C model, or the IRT-C model, which can be used to identify both uniform and non-uniform observed DIF across age and gender simultaneously. This procedure is compared to standard IRT DIF approaches, both Lord's $\chi^2$ (1980) and DFIT methodology (Raju, van der Linden & Fleer, 1995), where DIF on years of work experience and gender are examined separately. The purpose is to show convergence of IRT-C with conventional IRT-DIF procedures used in organizational science. Subsequently, we demonstrate how by specifying additional LCs in the MM-IRT-C model — unobserved DIF (i.e., DIF among latent classes) — we can account for a portion of observed DIF on the covariates. The aim is to provide not only procedures for examining different forms of DIF, but also to detail the practical decisions involved in specifying the correct model using the software Latent GOLD 4.5 (Vermunt & Magidson, 2008). Finally, we compare differences found when focusing only on observed DIF (via the single class/restricted MM-IRT-C model) versus examining overall DIF (via the full MM-IRT-C model).

## Method

Data were collected as part of a large-scale study of union involvement among public school employees (school teachers, librarians, counselors, and nurses) who belonged to a national education association in a northeastern state. Members of 446 school districts were surveyed by mail at their home addresses. Of the 4,000 surveys distributed to union members, 1,436 were returned, and 1,380 of these individuals provided usable data on the union citizenship scale (effective response rate = 35%). The average years of work experience of these respondents was 16.52 (SD = 10.83) and 66.1% were female. These two observed characteristics – years of work experience and gender – were used as continuous and nominal covariates, respectively, in our IRT analyses. The focal 8-item union citizenship/participation scale used for the current study was part of a longer survey, that also included scales assessing union and job attitudes (for more detail on the other survey measures, see Landis, Beal, and Tesluk, 2000). The union citizenship scale was derived from McShane (1986). It comprised items that asked, "In the last two years, have you:," and then listed seven activities designed to assess union citizenship, including running for union office, attending a union meeting, serving on a union committee, filing a grievance with the union, and participating in community related work for the union. The eighth item on this scale simply asked, "Have you been, or are you now, an elected officer in the local Association?" Data were collected in a dichotomous ('1' = Yes; '0' = No) response format. Table 3 presents the items, means and standard deviations for the union citizenship scale, for which Cronbach's alpha reliability was 0.75. Scale unidimensionality was ascertained via confirmatory factor analysis in the Mplus software (Muthén & Muthén, 2007), by specifying a one-factor model with categorical indicators and robust weighted least squares estimation. Fit of the unidimensional model was judged to be adequate ($\chi^2_{(df=16)}$ = 104.86; CFI = .98; TLI = .98; RMSEA = .064).

*Analytic Strategy*

The computer program Latent GOLD 4.5 (Vermunt & Magidson, 2008) was used to estimate the single class/restricted MM-IRT-C model (e.g., Figure 1A) and the full MM-IRT-C model (e.g., Figure 1C). The 8 union citizenship items were entered as observed indicators while work experience and gender were respectively entered as continuous and nominal covariates. It was necessary for us to specify the type of coding for the 8 dichotomous items/indicators; specifically, dummy coding was chosen ('0' is the lowest category) to produce item parameter estimates that are in line with the 2PL parameterization. Additionally, work experience was mean-centered to enhance interpretability of the coefficients. Latent GOLD 4.5 uses an expectation-maximization (EM) algorithm for maximum likelihood (ML) estimation. To avoid local minima for the log-likelihood, 10 random starts were used; and to increase the accuracy of the latent trait estimates, we set the number of quadrature points to 20 instead of the default of 10.

The focus of the current paper is to conceptualize and describe the MM-IRT-C model depicted in Figure 1C. This model simultaneously incorporates observed groupings and unobserved/latent groupings in the assessment of ME. Before we estimate the focal model (Figure 1C), however, we first estimate the restricted/single class MM-IRT-C (IRT-C) model (see Figure 1A). This restricted/single class IRT-C model is estimated for the purpose of showing how the current framework can incorporate testing of *observed* DIF; this initial submodel (Figure 1A) therefore accomplishes the same objective as traditional IRT-DIF methods used in organizational research.

*Single Class/Restricted MM-IRT-C (IRT-C).* To examine only observed DIF in the MM-IRT-C framework, we estimated an initial model where no DIF was specified (i.e., we estimated path 1 in Figure 1A, but not paths 2 or 3). This is akin to a fully constrained baseline approach, which has been the customary approach in IRT-DIF analyses (cf. Stark, Chernyshenko, & Drasgow, 2006). The possible presence of DIF on both years of experience and gender was determined by examining the bivariate residual statistic (BVR) between each covariate and each indicator. The BVR is analogous to a modification index (MI) for pairs of variables in factor analysis. See also Glas (1998) for the use of Lagrange Multiplier tests, a type of MI, to assess DIF. In the past, BVR values much larger than 1 or 2 have been proposed to indicate local misfit (cf. Vermunt & Magidson, 2000), and could indicate DIF may be present. An iterative stepwise process was used to identify the presence uniform or non-uniform DIF:

(Step a) We inspected the covariates × indicators BVR matrix, to identify the largest BVR value.

(Step b) For the largest BVR value from a covariate to an indicator, uniform DIF was specified (i.e., allowing for differences in item locations across levels of the observed covariate; path 3 in Figure 1A was freely estimated to attempt to account for the large BVR), and the statistical significance was determined using the Wald statistic.

(Step c) Non-uniform DIF on the same covariate-indicator pair was then added to the model (different item discriminations; path 2 in Figure 1A) and examined for statistical significance.

(Step d) The restricted MM-IRT-C model was re-estimated keeping only significant observed DIF effects. We note that if non-uniform DIF was found (significant path 2 in Figure 1A), modeling differences in item locations was necessary (path 3 in Figure 1A). This is analogous to testing interactions (differences in item discriminations) where main effects (differences in item locations) have to be kept in the model. Steps (a) through (c) were repeated until all the remaining BVR values between the covariates and indicators were sufficiently small.

The most parsimonious restricted MM-IRT-C (IRT-C) model was selected using log-likelihood information criteria, including the Bayesian information criterion (BIC) (Schwarz,

1978) and the consistent Akaike information criterion (CAIC; Bozdogan, 1987). Taking into account sample size and number of parameters, lower information values indicate better fit. Because penalty terms differ between the fit indices, we considered them both in deciding on the presence of uniform and non-uniform DIF.

*IRT-DIF analysis.* To examine the validity of our claim that the restricted MM-IRT-C model can identify observed group DIF, we compared it with standard IRT procedures. For the union citizenship scale we estimated a traditional 2PL IRT model (e.g., Reise & Waller, 1990) separately calibrated to years of work experience (junior vs. senior employees) and gender (male vs. female) groups. A median split was used to obtain two work experience groups of equal sizes. Iterative linking (Candell & Drasgow, 1988) was used to put items on a common metric, and Lord's $\chi^2$ values (1980) were computed using the software ITERLINK (Stark, 2002). After a Bonferroni correction, significant $\chi^2$ values would indicate DIF. Similarly, the DFIT program (Raju, 1999) was used to determine DIF, and the metrics between the groups were linked with the software EQUATE 2.1 (Baker, 1995); non-compensatory DIF (NCDIF) values larger than .006 indicated DIF (see technical manual by Raju, 1999).

*The MM-IRT-C model.* Unlike the single class/restricted MM-IRT-C model (Figure 1A), this model further broadens the conceptualization of DIF, by allowing the concurrent estimation of observed and unobserved measurement equivalence (Figure 1C). An initial unconstrained MM-IRT-C model was specified, allowing item discriminations and locations to be freely estimated across the LCs. As shown in Figure 1C, this is depicted by estimating both paths 5 and 6 across all the union citizenship indicators. Estimation of path 4 (theta mean difference across classes) requires that at least 1 item be invariant across LCs, otherwise such a model is not identified. Therefore, path 4 was not estimated here because the initial model does not have any invariant items across the LCs. In contrast, path 1 could be estimated because items are invariant with respect to the covariates.

Following similar procedures used previously for mixed-measurement models (Lubke & Muthén, 2005; Magidson & Vermunt, 2004; von Davier, 1997; Zickar et al., 2004), the number of latent measurement classes was determined based on the aforementioned log-likelihood information criteria. A recent simulation study comparing information criteria has shown that with sample sizes of 600 and 1200, the correct numbers of LCs were usually recovered with the BIC for a mixed-measurement model with a 1-, 2- or 3-parameter logistic IRT response function (see Li, Cohen, Kim, & Cho, 2009). Thus, we relied more on the BIC in making decisions regarding the numbers of LCs. We fit incremental numbers of latent classes for the union citizenship scale and stopped when the BIC criterion increased. If the information criteria point to more than one LC, it would indicate that there are distinct latent measurement classes not fully captured by a single measurement model.

After determining the appropriate number of LCs, the MM-IRT-C model was further pruned to increase parsimony and to determine the significance of covariate and DIF effects using the following steps:

(Step a) *Relationship of covariates to the theta distribution within classes (Path 1 in Figure 1C).* We determined if the covariates years of experience and/or gender related to the latent trait of union citizenship within each LC by allowing class-specific effects. Any non-significant effects were constrained to zero.

(Step b) *Relationship of covariates to the latent class proportions (Path 7 in Figure 1C).* work experience and/or gender were allowed to predict class membership. For this analysis, non-significant effects between covariates and latent class proportions were set to zero.

(Step c) *Unobserved non-uniform DIF; equality of item discriminations across latent classes (Path 5 in Figure 1C).* Items that did not have significantly different item discriminations across LCs were constrained to equality. Further, class-specific item discriminations that were not significantly different from zero were constrained to zero. Such items did not discriminate between individuals within the LC. At this juncture, it is important to note that path 6, representing uniform DIF, would still be freely estimated for the item as we specified an initial unconstrained model. If item discriminations differ between LCs, item locations may or may not differ between LCs, and we can examine the significance of these differences in locations as well.

*(Step d) Unobserved uniform DIF; equality of item locations across latent classes (Path 6 in Figure 1C).* For items with equal item discriminations across the LCs, we ascertained if uniform DIF was present. If the item locations were not significantly different across the LCs, we constrained the item locations to equality, noting that these items were measurement invariant across unobserved groups (or latent classes).

*(Step e) Observed group DIF within each LC; equality of item locations and/or discriminations across observed characteristics (Paths 2 and 3 in Figure 1C).* Because unobserved measurement groupings may not fully account for differences in scale use, we determined if there was observed DIF (on work experience and gender) within each LC. This was accomplished by examining large BVR values in the covariate × indicator BVR matrix, which may be indicative of residual DIF. To assess if residual observed DIF occurs within a LC, and to also assess the type of DIF (uniform or non-uniform), several submodels were compared: (1) class-specific uniform residual observed DIF effects; non-significant effects within a class can be constrained to equality across the covariate of interest; (2) class-specific non-uniform residual observed DIF; again, non-significant effects within a class can be constrained to equality across the covariate of interest. Steps (1) and (2) were repeated for each large covariate × indicator BVR in turn, until there was little evidence of residual observed DIF.

*(Step f)* If invariant items were found among the LCs, path 4 could be estimated to ascertain if the latent trait levels differed among the LCs.

Across steps (a) to (f), the BIC was used to determine if further restrictions would increase model parsimony. Additionally, aside from relative global fit determined from information criteria, local misfit was evaluated using the BVR. If the average BVR values are fairly small, it indicates good model-data fit.

*Measurement Equivalence and Possible Capitalization on Chance*. At this point, before we present the results, we should address a potential problem that is inherent in the majority of measurement equivalence research—the possibility of capitalization on chance. That is, in most studies of DIF, it is common for researchers to test for item nonequivalence without making *a priori* predictions about exactly which items are likely to differ in location and discrimination parameters (see Stark et al., 2006; Vandenberg & Lance, 2000). In that sense, item-level ME tests are often carried out in an exploratory/inductive fashion. Whereas some researchers have suggested applying Bonferroni corrections in ME tests to control family-wise error rates, this practice has been shown to result in low statistical power (less than .50) under conditions of small uniform DIF (Stark et al., 2006). As such, the current study follows the majority of the ME literature by proceeding in a stepwise/exploratory fashion. Our goal is to illustrate the use of a novel technique (Figure 1C) for assessing both observed and unobserved group DIF. As with any such analysis, greater confidence in the mixture model results can eventually be obtained through replication using a validation sample (e.g., Wang, 2007).

## Results

*Single Class/Restricted MM-IRT-C (IRT-C) and Observed DIF.* As seen from Table 4, the BIC and CAIC values showed that Model 6 (M6) was the most parsimonious model and the Latent GOLD syntax for specifying the final model is shown in the Appendix. According to this model, observed DIF was obtained for work experience on items Union8 (non-uniform DIF), Union7 (uniform DIF) and Union1 (uniform DIF). These results were compared to standard IRT-DIF procedures Lord's $\chi^2$ and DFIT. Because the implementation of Lord's $\chi^2$ in the ITERLINK method uses a Bonferroni correction, less DIF was detected as compared to the DFIT procedure. DIF was found for work experience on Union8 and for gender on Union8 using Lord's $\chi^2$. However, DIF was detected on more items with DFIT: work experience on Union1, Union2, Union7, and Union8; and gender on Union8. The direction of DIF corresponded across all three methodologies; a plot of the item response functions (IRFs) in Figure 2 shows graphically that the direction of DIF detected in the IRT-C procedure was the same as that yielded by the Lord's $\chi^2$ and DFIT procedures[2].

-- insert Figure 2 about here --

Thus, the restricted IRT-C model (Figure 1A; which examines observed DIF but treats the observed grouping variable as a covariate) and the traditional observed group DIF procedures produced similar results, with the primary exception being that the DFIT procedure signaled gender DIF on item 8. To better interpret this gender result, we also note here that the restricted MM-IRT-C approach models both covariates (work experience and gender) simultaneously, while traditional DIF procedures model only one observed grouping variable at a time. Because the point biserial correlation between work experience and gender is .25 (men have more work experience), the weaker residual DIF effects on gender after accounting for DIF on work experience may imply that work experience can be a *mechanism* by which gender DIF occurs on item 8 (e.g., Baron & Kenny, 1986). Traditional DIF methods, which model the two observed grouping variables separately, cannot assess such phenomena (i.e., residual gender DIF on item 8 after controlling for work experience DIF on item 8). This is a conceptual strength of the restricted MM-IRT-C model for DIF detection. By incorporating multiple observed covariates simultaneously, we can begin to test explanatory variables (such as work experience) that might start to answer *why* DIF is observed between gender groups.

*MM-IRT-C and Overall DIF.* The single class/restricted MM-IRT-C (IRT-C) procedure showed that observed DIF occurred on several items for individuals with varying levels of work experience. In this subsequent analysis, we determine if the observed DIF can be accounted for by unobserved subpopulations (latent classes) who use the scale in distinct manners. The results in Table 5 show that a two latent class (LC) solution fit the data well, as indicated by the smallest BIC and CAIC values[3]. We note that this two-LC solution (M2) had much lower information criteria than all the *IRT-C* models shown in Table 4. Hence, positing unobserved subpopulations – latent classes with differences in scale use – resulted in relatively better model-data fit. After the pruning strategy employed in steps (a) through (f) as described above, we found that Model 11 (M11) was the best fitting model, with the lowest BIC and CAIC values among the models compared. Further, the average BVR values were very low, indicating good absolute model-data fit. All the BVR values between the covariates and indicators were smaller than 2 with the largest being 1.21 between work experience and Union1. Given that non-invariant items were present, we could test for differences in latent trait levels between LC1 and LC2 (path 4) as depicted in Model 14 (M14); however, this difference was non-significant. Thus, both LCs had on average similar levels of latent trait scores. Also, additionally freeing the slightly larger BVR between Union4 and Union5 in the final model Model 11 (M11) did not improve relative model-data fit. The Latent GOLD syntax for specifying the final model (M11) is shown in the Appendix.

In the final model, we found that there was no DIF on 4 union citizenship items (Union3 to Union6) across both observed (work experience and gender) and unobserved (LCs) groups. Thus, there were no unobserved differences in scale use on these four items (e.g., response sets), and individuals did not differ in their expected response because of their observed characteristics on work experience and gender. On the other hand, unobserved DIF between LCs 1 and 2 was found on items Union1, Union2, Union7 and Union8, and residual observed DIF was found on Union8 across gender groupings within LC2. The form of DIF on items will be explicated more as we describe the LCs. For now, it is sufficient to note that there was a close correspondence between the results of the MM-IRT-C procedure and the two preceding procedures: the single class/restricted MM-IRT-C (IRT-C) model (when only observed group DIF was examined), and other standard IRT-DIF procedures. That is, the same items are flagged for DIF. However, the attributed source of DIF – observed versus unobserved – differs when one transitions from an IRT-C model to a latent class MM-IRT-C model.

*Description of LCs.* The latent class proportions for LC1 and LC2 were .68 and .32 respectively. Item response functions for the two LCs (Figure 3) ultimately showed four major differences between LC1 and LC2 (as seen in items 1, 2, 7, and 8). Most notably, individuals likely to be in LC2 exhibited *no discrimination* on item 1 ("Run for an elected local Association office") nor item 8 ("Have you been, or are you now, an elected officer in the local Association?"). In other words, whereas individuals in LC1 were likely to manifest their latent levels of union citizenship by attempting to participate in elected offices, individuals in LC2 did not manifest their latent union citizenship in this way. In contrast, LC2 individuals displayed greater discrimination on item 7 ("Participated in community related work for the local Association"), suggesting that within this latent class union citizenship levels were more likely to be manifested through community service, not through seeking political office. As such, we labeled LC1 the "politico" latent class, and labeled LC2 the "non-politico" latent class, because of the differences in how these two latent classes manifest their underlying levels of union citizenship.

The covariate effects on latent trait and latent class membership (Figure 1C, paths 1 and 7) revealed several differences between the two LCs. Foremost, being male and having more work experience were related to a higher latent trait standing on union citizenship in LC1 (politico class), but not in LC2 (non-politico class), as tested in step (a). Also, having more work experience was related to a higher probability of being in LC2 (posterior M= 28.49, SD = 5.56) as compared to LC1 (posterior M= 11.90, SD = 8.42); on the other hand, gender was not related to LC membership, as tested in step (b). Thus, the effect of observed characteristics relates to scale use in a nuanced manner: having more work experience was related to having a higher probability of being in LC2 (non-politico). However, given that a participant was a member of LC1 (politico), work experience and gender were related to quantitative differences on union citizenship. In contrast, the covariates were not related to differences on union citizenship within LC2.

A comparison of the model-predicted marginal endorsements of union citizenship items showed that there was a large degree of similarity across both LCs, as illustrated in Figure 3. In general, there were slightly higher endorsements of citizenship items in LC2 than in LC1. A plot of the item response functions (IRFs) between the LCs showed that half of the union citizenship items were ME (Union3 to Union6) as tested in steps (c) and (d). Uniform DIF was present on Union2, “Held a local Association position?” in that this item was less likely endorsed by individuals within LC2 (versus LC1), given the same latent trait standing. In other words, non-politicos (LC2) required a higher level of union citizenship in order to participate in any local positions (including non-elected local positions; i.e., item 2). Non-uniform DIF was present between the LCs on the remaining items. Union7, “Participated in community related work for the local Association?” were more discriminating for LC2. A cross-over effect was observed on Union7: among individuals with a lower union citizenship latent standing, those in LC1 had a higher probability of endorsing the item than did those in LC2, while the converse was true for individuals with higher latent trait standings (see Figure 3). It is important to note that these trends matched observed DIF effects shown in Figure 2, as LC2 (non-politico) individuals generally have more work experience than LC1 (politico) individuals. Although these trends are similar, we stress that the conventional observed DIF procedure stereotypically assumes that individuals with more work experience (> median work experience) share one common measurement model, as compared to individuals with less work experience (< median work experience). In contrast, the MM-IRT-C approach allows for a more detailed account of the association between work experience and scale use. Using a median split to capture work experience differences in scale use fails to consider the imperfect overlap of work experience with unseen LC differences in scale use (see Figure 4).

Additionally, as mentioned earlier, we found that Union1 “Run for an elected local Association office?” and Union8 “Have you been, or are you now, an elected officer in the local Association?” did not discriminate individuals within LC2 as reflected in the flat IRFs, but these items were highly discriminating in LC1 (see Figure 3). These differences indicate the degree to which individuals were attempting to manifest their latent union citizenship by seeking elected office in the union. Males who had more work experience in LC1 (politicos) had a higher probability of endorsing such items because these characteristics were related to a higher union citizenship latent trait standing. Unlike items Union2 and Union7, however, the unobserved DIF effects here did not match the observed DIF effects. Whereas LC1 corresponded to lower work experience and LC2 corresponded to more work experience, simply comparing ME on the observed characteristic of work experience (more vs. less) did not signal underlying differences in scale use, which were captured by the LCs. This illustrates the advantage of the MM-IRT-C procedure: it simultaneously considers the similarity of item response vectors across latent classes *and* individual characteristics (work experience and gender), and incorporates all this information to appropriately describe how individuals differently use the measurement scale.

We found residual observed uniform DIF on Union8 for gender within LC2, as tested in step (e). That is, there was gender DIF on item 8 among non-politicos (LC2). Within LC2, males had a greater probability of endorsing the item across the latent union citizenship continuum, which corresponded to the observed DIF effects in Figure 2. This finding is not surprising because gender did not distinguish LC differences in scale use. Because responses by males and females were equally likely to describe the LCs, marginal comparisons across gender would yield primarily *observed group* differences in scale use. What is also important to note however, is that a large proportion of individuals in the sample, as indicated by LC1 (politicos, 68%), shared a common measurement model and no gender DIF was present among these individuals. Hence, observed DIF on gender was primarily attributable to a smaller number of males and females (LC2, non-politicos) who used the scale in a distinct manner.

*Summary of findings.* The single class/restricted MM-IRT-C (IRT-C) procedure introduced here showed a moderate number of DIF items (Union1, Union7 and Union8); slightly more than identified through Lord's $\chi^2$ implemented in the ITERLINK (Union8) procedure, and about the same as found via DFIT methodology (Union1, Union2, Union7 and Union8). Hence, observed DIF was effectively detected with the MM-IRT-C (IRT-C) model. A comparison of observed DIF procedures and overall DIF procedures using MM-IRT-C showed striking similarities in that DIF was found on largely the same set of items. However, by applying the MM-IRT-C model, it appears that much of the observed group DIF found in the former procedures was alternatively attributable to unseen differences in scale use, in the form of two LCs characterized by partial measurement equivalence on the union citizenship scale. The MM-IRT-C model fit the data much better than the single-class/restricted MM-IRT-C

(IRT-C), indicating that unseen differences in scale use were present. There was a nuanced relationship between observed characteristics (work experience and gender) and unseen differences in scale use. More work experience was related to LC2 (non-politico) membership, while having more work experience and being male were related to higher union citizenship within LC1 (politico) (but this was not the case for individuals within LC2). It is important to note that we were able to ascertain these differential quantitative effects of demographic covariates on union citizenship response styles after taking into account latent class (unobserved) DIF in the union citizenship scale.

## Discussion

Measurement equivalence (ME) is a key topic in organizational research (Vandenberg and Lance, 2000). Two prime examples of organizational ME/DIF research are cross-cultural comparisons (Riordan & Vandenberg, 1994) and test fairness/personnel selection research (Stark et al., 2004). The current paper described how past organizational research on DIF was limited by methods designed to ascertain DIF across two observed groups at a time. In contrast to past organizational DIF research, many additional questions about DIF might also be asked (see Table 6). As summarized in Table 6, there are at least eight DIF-related questions that future researchers can pursue. Only the first of these eight DIF questions has been the province of traditional DIF methodology used in organizational research. By presenting the MM-IRT-C model, the current paper attempts to extend organizational research to better address DIF questions 2 through 8 in Table 6. By proposing a set of integrated tests to answer new DIF questions, we hope to reveal a new frontier for organizational DIF research, explaining both *why* and *for whom* DIF occurs.

By introducing the theoretical differences between observed and unobserved ME, and by illustrating the corresponding methodological approach for integrating the two, we hope to advance and stimulate ME research among organizational scientists in two important ways. Foremost, we generalized how observed DIF can be tested with the single class/restricted MM-IRT-C (IRT-C) model, enabling a simultaneous examination of DIF on multiple observed characteristics, both categorical and continuous. This allows us to consider and test for the relative importance of multiple observed characteristics leading to DIF/differences in scale use. Second, we broaden the common usage of ME within the organizational literature by presenting a conceptual framework for observed, unobserved, and overall DIF (Table 1). This is conjoined with a presentation of the MM-IRT-C approach for testing overall DIF. We further extend past presentations of MM-IRT (Hernandez et al., 2004; Rost, 1990, 1991; Rost, Carstensen, & von

Davier, 1997; Zickar et al., 2004) and MM-IRT-C (Maij-de Meij et al., 2005, 2008; Smit et al., 1999, 2000) procedures by allowing class-specific covariate effects on the latent traits within LCs, testing for uniform and non-uniform unobserved and observed DIF, and examining if latent trait scores differ between LCs given unobserved partial ME. To our knowledge, these aspects have not been previously proposed. We next discuss these two main methodological contributions and their application to organizational research.

*A General Approach to Observed DIF*

The restricted MM-IRT-C model (IRT-C) is specifiable in the general latent variable modeling approach to ascertain observed-group DIF. It is a generalization of traditional IRT procedures in that both uniform and non-uniform observed DIF can be examined on multiple observed characteristics simultaneously, and not limited to categorical variables. To show its utility, we presented these analyses alongside standard IRT-DIF methods commonly applied by organizational researchers. We found that observed DIF was effectively detected in our dataset. This technique has a variety of important theoretical applications to organizational research, as presented in Table 2.

One obvious application for simultaneous DIF detection is to examine the relative influence of observed characteristics on scale ME, particularly for correlated observed variables. In our IRT-C application on the union citizenship scale, observed DIF was found on work experience but not gender. There was a point biserial correlation of .25 between work experience and gender (male = 1; female = 0), and it is possible that residual DIF effects may be weaker on gender after accounting for DIF on work experience. In other words, after gender DIF was detected using traditional IRT-DIF approaches, we showed that modeling work experience DIF in tandem with gender DIF led to non-significant gender DIF effects (i.e., after controlling for work experience DIF). This result suggests that differences in work experience may be the explanatory mechanism (or mediator variable) by which gender DIF occurs on this scale. Such analyses are not possible when only one observed characteristic/covariate is modeled at a time.

This technique may have major implications for cross-cultural research. Expanding the way DIF is tested can impact how ME between countries or organizations is examined, which is of key interest to organizational scientists (see Riordan & Vandenberg, 1994; Schaffer & Riordan, 2003). In making cross-national or cross-organizational comparisons, DIF is commonly attributed to culture/climate or a comparably abstract construct. Nevertheless, this reasoning begs the question as to what aspects of a culture/climate may have led to DIF. We propose there are multiple manifest markers of culture/climate -- beyond that of country or organizational membership -- on which DIF could be tested simultaneously and relative effects could be ascertained. For instance, Hofstede (1984; 2001) proposed multiple cultural dimensions on which countries vary; similarly, individuals within countries may vary on different aspects of individualism-collectivism (Triandis & Gelfand, 1998). Using the restricted MM-IRT-C (IRT-C) model, we can test for the relative influence of cultural dimensions on scale ME, simultaneously with assessing ME for country-membership. Additionally, there are non-cultural factors that may account for the occurrence of DIF such as response context or familiarity with the language in the survey (Robert, Lee, & Chan, 2006). By applying the restricted MM-IRT-C (IRT-C) model, we can determine if cultural or non-cultural effects contribute more to observed DIF.

Another advantage implicit in our discussion above is that because covariates can be continuous, it is not necessary to partition participants into two subgroups (e.g., high-low) when there are clearly quantitative differences even within the subgroups. A conceptual extension is that continuous variables or constructs (e.g., age or job satisfaction) can be used in the restricted MM-IRT-C (IRT-C) analysis to examine DIF (as contrasted to arbitrary partitioning as used in

IRT-DIF analyses). There may be greater sensitivity and power in this approach, as more information is utilized in contrast to a "median-split" approach.

### *Combining Observed DIF with Unobserved(Latent Class) DIF*

At the outset of this paper, we proposed a framework for observed, unobserved and overall DIF (Table 1), noting that observed characteristics alone may not be sufficient to characterize underlying/unseen differences in scale use (see Table 2). For instance, response sets or response styles may be related to, but not fully demarcated by a single observed characteristic (e.g., gender); however, the use of observed groups is entrenched in ME theory and methodology such that it is not uncommon to presuppose that say, all women would respond to a scale or a selection test in a distinct manner from men. However, in making a case for considering unobserved ME (latent response classes), it has been suggested that the observed ME approach in educational testing "[can] … unfortunately lead to a stereotypical view of an item as being advantageous to all members of the group (males or females) favored by the DIF item, while ignoring the true heterogeneity within each group" (p. 134, Cohen & Bolt, 2005). Indeed, in the context of educational testing, there is now greater recognition that observed characteristics may not correspond with unseen differences in how individuals respond to tests; for instance, there may be latent differences in how individuals employ test strategies (Mislevy & Verhelst, 1990) or respond to test procedures (Bolt, Cohen, & Wollack, 2002).

In the organizational context, it is similarly difficult to ascertain *a priori*, based on observed characteristics, how individuals would respond to test items or organizational scales, because there may be differences in faking (Zickar et al., 2004), scale interpretation (Hernandez et al., 2004), and more generally, response styles or substantive interpretations of the construct (Eid & Rauber, 2000; Eid & Zickar, 2007; Rost et al., 1997) *within* observed groups. This has important implications because simulations have shown that where true (latent) measurement groups fail to match observed groups, these techniques can have little power to identify unobserved DIF (De Ayala, Kim, Stapleton, & Dayton, 2002). If there are indeed unseen differences in scale use (latent response classes), it is necessary to discern who the individuals are that respond differently, and on which items. This is because score comparisons are less meaningful when there is DIF (either observed or unobserved DIF).

Consequently, it is critical to consider the use of MM-IRT to infer these underlying measurement groups (unobserved DIF; Hernandez et al., 2004; Rost, 1990, 1991; Rost et al., 1997; Zickar et al., 2004). However, because observed characteristics may also contain information relevant to how individuals respond to the scale, the combined (observed and unobserved DIF detection) MM-IRT-C approach is recommended in this paper (see also Maij-de Meij et al., 2005, 2008; Smit et al., 1999, 2000). Specifically, we can ascertain if there are unseen differences in scale use in the form of LCs that are not ME, and we can further determine whether multiple observed participant characteristics are related to these LCs. As evident in our analysis of union citizenship, this relationship of observed characteristics to LC membership can be modest (years of work experience) or even non-significant (gender). Thus, these subtle, unobserved/latent class differences in scale use may not be easily detectable via analysis of observed characteristics only (see also Eid & Rauber, 2000).

After inferring the appropriate number of LCs, we proposed a procedure for testing unobserved DIF across these LCs, which to our knowledge has not been proposed in past applications of MM-IRT-C. It is usually assumed that LCs employ distinct measurement models, but further constraints of the measurement model are needed to assess whether LCs have partial measurement equivalence versus complete measurement nonequivalence (cf. Lubke & Muthén, 2005), and to determine the type of unobserved DIF (uniform or non-uniform) that is present. It is hoped that future research applying MM-IRT and MM-IRT-C procedures can use our proposed framework to test for the ME of items among LCs. In our analysis, we found that unobserved DIF occurred for only half of the union citizenship items, reflecting partial ME between LCs. In particular, most of the observed DIF found in observed IRT-DIF procedures (i.e., single-class/restricted MM-IRT-C; Lord's $\chi^2$; DFIT methodology) was accounted for by unseen differences in scale use between the LCs.

Another potential contribution for organizational measurement is to ascertain commonalities in scale use among observed group members, despite the presence of DIF on observed group membership. For instance, in our analysis we found that there was a large proportion of individuals in the politico group (LC1, 68%) who all shared the same measurement model. Both males and females within LC1 used the scale in the same manner. At the same time, DIF occurred between males and females in the non-politico group (LC2) on Union8. This finding suggests that differences in scale use (DIF) between men and women may be evident in one latent class, and absent from another. Thus, identifying latent classes might be crucial to understanding the nature of gender DIF. The traditional IRT analysis signaled gender DIF, but it was driven by the non-politicos only.

We propose that with MM-IRT-C, we can address the issue of whether there exist normative classes of individuals who share the same measurement model despite belonging to different observed groups. This issue may be particularly pertinent in diversity and cross-cultural research. For instance, in cross-cultural research it has been found using LC procedures that there is a large proportion of individuals across countries who shared the same LC membership, but a smaller proportion of individuals who were found in idiosyncratic LCs -- LCs which predominantly consist of *country-specific* members (Eid & Diener, 2001). We suggest that via MM-IRT-C it is possible to test whether DIF is driven by a small group of individuals who use the scale differently, while the majority of individuals use a common measurement model. The focus then is not only on *if* DIF occurs, but *for whom* DIF occurs.

The issue of unseen differences in responding is also important for the topic of predictive validity. A recent study by Maij-de Meij and colleagues (2008) showed that by considering unseen subgroup differences in the use of the question mark "?" on personality scales, it is possible to improve the predictive validity of expert judgments of personality. In our analysis, we go beyond Maij-de Meij and colleagues (2008) by allowing differential effects of covariates on different unobserved subpopulations of individuals. We found that for LC1 (politico), there was a positive relationship of latent union citizenship to years of work experience and gender (being male), but this effect was not significant among the other 32% of individuals in LC2. Without considering possible unobserved subpopulations that use the scale in distinct manners, one would presume that years of work experience and being male would be positively related to union citizenship for all individuals in the sample. Taken together, the above analyses allow us to further examine both the *reasons* and the *boundary conditions* underlying observed DIF, rather than simply concluding that the scale is not comparable across members of observed groups.

### *Limitations and Future Directions*

Our goal in this paper has been to introduce the MM-IRT-C model to organizational researchers, while outlining a sequence of steps for testing the various aspects of DIF. In future work, it will be necessary to rigorously examine the requirements and weaknesses of both the single-class/restricted MM-IRT-C (IRT-C) procedure for detecting observed DIF and the MM-IRT-C. Past research using the MIMIC model, which is very similar to the restricted MM-IRT-C (IRT-C) model except that it uses a factor analytic framework, has compared data requirements for testing DIF in the MIMIC framework against a traditional two-group method (Woods, 2009). It was found that the MIMIC model required smaller sample sizes in the focal group (e.g., 200 and 400) for accurate power to detect DIF and accurate item parameter estimates, as compared to the IRT likelihood-ratio approach (Thissen, Steinberg, & Wainer, 1993). Additionally, a recent Monte Carlo study examined the best information criteria to detect the appropriate number of LCs (Li et al., 2009) across different measurement models (1PL, 2PL and 3PL), and showed that the appropriate numbers of LCs were recovered using 600 individuals. Although these studies are partially informative as to the range of sample size requirements necessary in our proposed MM-IRT-C applications, we propose that Monte Carlo studies specific to the MM-IRT-C model are an important future direction.

### *Conclusion*

Past research applying DIF techniques has frequently focused on observed group comparisons. We propose here that the single-class/restricted MM-IRT-C (IRT-C) procedure can be used to ascertain DIF on multiple observed characteristics, both continuous and categorical, simultaneously. Further, although much of ME research has primarily been limited to observed groups, we presented a broader definition of IRT-DIF, in which unobserved/latent class differences in scale use are also considered. We then proposed that the MM-IRT-C model can be used to flexibly test both unobserved and observed DIF within a single model. Because a broader set of substantive ideas can be tested, we hope that organizational researchers will find this model useful in future measurement applications.

Table 1

*Conceptual Table for Different Types of Differential Item Functioning (DIF)*

| Type of DIF | Definition | Method | Limitation(s) |
|---|---|---|---|
| Observed DIF | The expected observed score at a given latent trait level is dependent upon observed group membership (Drasgow, 1984; Drasgow & Kanfer, 1985). Different measurement models hold across different observed groups. | **IRT-DIF commonly used by organizational researchers**<br>-Lord's $\chi^2$ (1980)<br>-DFIT methodology (Raju, 1988) | -Does not account for unobserved group (i.e., latent class) differences in scale use.<br>-Does not allow testing of DIF on multiple observed characteristics within a single model.<br>-Continuous variables are usually dichotomized to create observed groups when testing DIF.<br>-It is more difficult to test for DIF in multiple groups (>2 groups).<br>-Linking of metrics between groups is done as a separate procedure before DIF can be tested. |
| | | **Single Class/Restricted mixed-measurement-IRT with covariates (See Figure 1A)**<br>-We propose an extension of the logistic-regression approach (Swaminathan & Rogers, 1990) for testing DIF.<br>-To test only for observed DIF, a restricted form of MM-IRT (one latent class; i.e., IRT with covariates) can be used to test observed DIF on multiple observed characteristics (both categorical and continuous). Linking of metrics of observed groupings/characteristics is implicitly taken into account.<br>-We can test for both uniform and non- | -Does not account for unobserved group (i.e., latent class) differences in scale use. |

| Type of DIF | Definition | Method | Limitation(s) |
|---|---|---|---|
| | | uniform DIF. | |
| Unobserved DIF (i.e., latent measurement classes) | The expected observed score at a given latent trait level is dependent upon latent group membership (i.e., latent classes). Different measurement models hold across different unobserved groups (Rost, 1990, 1991). | **Mixed-measurement IRT (See Figure 1B)**<br>-Examine DIF over unobserved groups. | -Does not model linkages between observed characteristics/groupings and unobserved groups.<br>-Effectively assumes observed characteristics are uninformative as to differences in scale use. |
| Overall DIF | The expected observed score at a given latent trait level can be dependent upon both observed characteristics and unobserved (latent class) group membership. Overall DIF constitutes both observed and unobserved DIF. | **Mixed-measurement-IRT with covariates (See Figure 1C)**<br>-Determine if there are unobserved differences in scale use; assess the degree to which latent subgroups display distinct measurement models.<br>-Examine linkages between observed characteristics and unobserved group membership.<br>-Examine relationships between observed characteristics and latent trait standing within each unobserved group. | |

Table 2

*Advantages of the MM-IRT with Covariates Approach over Traditional IRT-DIF Approaches, and Example Organizational Topics*

| **Traditional IRT DIF approaches** | | **MM-IRT with covariates** | **Organizational topics** |
|---|---|---|---|
| Methodological limitations | Corresponding conceptual limitations | Advantages | Exemplar Research Topics/Questions |
| Observed DIF needs to be examined for each observed grouping independently and in turn. For instance, DIF on work experience is examined separately from DIF on gender. While partitioning on several observed groupings is possible, it becomes less feasible with many distinct group memberships. | Occurrence of DIF may be due to a particular participant characteristic/ grouping, more so than others. Multiple grouping variables cannot be studied simultaneously. | We can test for DIF using multiple covariates simultaneously. If the same item exhibits DIF on two or more covariates, we can compare their relative effects across covariates. | *Cross-cultural, diversity research*<br>To what degree does DIF occur on specific cultural dimensions aside from, say, country-membership? Is DIF primarily due to cultural or non-cultural factors? |
| Participant characteristics that are continuous variables (e.g., age) are arbitrarily dichotomized/ polytomized into "homogeneous" subpopulations for the purposes of examining DIF. | Partitioning into groups leads to loss of information and oversimplifies results. | Because covariates can be continuous variables, we do not need to split a continuous variable into categorical subgroups to test for DIF. | *Research on Aging or ME among individuals holding different attitudes and dispositions*<br>- Is age related to scale measurement properties?<br>- Do individuals with higher self-awareness use scales differently (e.g., Kulas & Finkelstein, 2007)?<br>- Do individuals with different attitude levels use a scale differentially? |
| (a) Assumes that a participant can be uniquely assigned to an observed group and that all members within an observed group necessarily share the same measurement model. | (a) There is a degree of heterogeneity within observed groups and it is possible that not all individuals within an observed group share the same frame-of-reference or use the scale in the same manner. It may even be stereotypical to assume all members (e.g., males vs. | (a) By using information from observed group membership (in the form of covariates), and inferring latent classes with distinct measurement models, we can obtain a probabilistic association | *Research on Measurement invariance.*<br>By moving away from the traditional approach of “manifest group = measurement group,” we can not only ascertain on which items did DIF occur, but also *for whom* did DIF occur: Do response sets occur for only a small portion of |

| **Traditional IRT DIF approaches** | | **MM-IRT with covariates** | **Organizational topics** |
|---|---|---|---|
| Methodological limitations | Corresponding conceptual limitations | Advantages | Exemplar Research Topics/Questions |
| | females) would respond to a scale in the same manner (Cohen & Bolt, 2005). Observed groups may not necessarily demarcate how individuals use a scale. | between latent measurement groups and observed groups/characteristics. Specifically, we can determine how observed characteristics are probabilistically related to unobserved DIF. | the observed group? For instance, only a fraction of females (small measurement-class [10%] consisting of primarily females) use extreme responding, but most males and females use the scale as expected (large measurement-class [90%]). |
| (b) Where measurement nonequivalence on a scale occurs, it is often difficult to ascertain the reasons (cf. Vandenberg, 2002). | (b) Differences in scale use could be associated with unseen factors such as response styles (Eid & Rauber, 2000). It is important to explore the subgroups of individuals who exhibit distinct frames-of-reference. This cannot be achieved in the traditional IRT DIF approach.<br>As the predictive validities of observed subgroups may differ due to unobserved classes, it is important to take into account such unseen differences in scale use (Maij-de Meij et al., 2008). | (b) It is possible for distinct latent measurement classes of individuals to be characterized using multiple observed variables.<br>We can ascertain if the observed variables have the same relationship with the latent trait across different latent response classes. | |

Table 3

*Means and standard deviations of Union Citizenship Scale items*

| Items | Item description | Mean | SD |
|---|---|---|---|
| Union1 | Run for an elected local Association office? | 0.09 | 0.29 |
| Union2 | Held a local Association position? | 0.18 | 0.38 |
| Union3 | Served on a local Association committee? | 0.29 | 0.45 |
| Union4 | Gone to a local Association meeting? | 0.86 | 0.35 |
| Union5 | Represented the local Association at a state or regional meeting, or at a convention? | 0.09 | 0.28 |
| Union6 | Filed a grievance through your local Association? | 0.07 | 0.25 |
| Union7 | Participated in community related work for the local Association? | 0.34 | 0.47 |
| Union8 | Have you been, or are you now, an elected officer in the local Association? | 0.18 | 0.39 |

Table 4

*Fitting of restricted MM-IRT model with covariates (IRT-C) years of work experience and gender to the Union Citizenship Scale to identify uniform and non-uniform observed DIF*

| Model | Model Description | Significant DIF? | Log-Likelihoood | BIC | CAIC | Npar | BVR Covariate & Indicator (Largest Value) | Mean (SD) |
|---|---|---|---|---|---|---|---|---|
| M1 | Fully constrained model: no DIF | | -3552.48 | 7233.66 | 7251.66 | 18 | Work experience & Union8 (26.40) | 2.84 (5.11) |
| M2 | M1+ uniform DIF: Work experience & Union8 | Y | -3528.26 | 7192.37 | 7211.37 | 19 | Work experience & Union7 (4.85) | 2.13 (3.77) |
| M3 | M2 + non-uniform DIF: Work experience & Union8 | Y | -3514.90 | 7172.80 | 7192.80 | 20 | Work experience & Union7 (5.60) | 2.10 (3.61) |
| M4 | M3 + uniform DIF: Work experience & Union7 | Y | -3511.07 | 7172.29 | 7193.29 | 21 | Work experience & Union1 (4.31) | 2.01 (3.49) |
| M5 | M4 + non-uniform DIF: Work experience & Union7 | N | -3509.13 | 7175.56 | 7197.56 | 22 | Work experience & Union1 (4.13) | 2.04 (3.55) |
| ***M6*** | ***M4 + uniform DIF: Work experience & Union1*** | ***Y*** | ***-3503.96*** | ***7165.21*** | ***7187.21*** | ***22*** | ***Gender & Union7 (2.35)*** | **1.90 (3.44)** |
| M7 | M6 + non-uniform DIF: Work experience & Union1 | N | -3503.91 | 7172.27 | 7195.27 | 23 | Gender & Union7 (2.37) | 1.90 (3.47) |
| M8 | M6 + uniform DIF: Gender & Union7 | N | -3502.31 | 7169.06 | 7192.06 | 23 | Work experience & Union4 (2.05) | 1.84 (3.43) |
| M9 | M8 + non-uniform DIF: | N | -3501.86 | 7175.33 | 7199.33 | 24 | Work experience & | 1.83 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Gender & Union7 | | | | | | Union4<br>(2.06) | (3.42) |
| M10 | M6 + uniform DIF:<br>Gender & Union8 | Y | -3501.98 | 7168.41 | 7191.41 | 23 | Gender & Union7<br>(1.99) | 1.84<br>(3.49) |
| M11 | M10 + non-uniform DIF:<br>Gender & Union8 | N | -3501.97 | 7175.54 | 7199.54 | 24 | Gender & Union7<br>(1.99) | 1.85<br>(3.50) |

*Note*. Npar denotes the number of model parameters. The final model selected was Model 6 (M6), which is italicized. This model has the lowest BIC and CAIC among all the other models. Further, the largest BVR value among covariates and indicators is fairly low, around 2. This final model showed that DIF was on Work experience & Union8 (non-uniform), Work experience & Union7 (uniform) and Work experience & Union1 (uniform). BVR mean (SD) denotes the average (standard deviation) of BVR values among the covariates and indicators.

Table 5

*Fitting of MM-IRT model with the covariates years of work experience and gender to the Union Citizenship Scale*

| **Aims: To determine…** | **Step** | **Model** | **Model Description** | **Log-Likelihood** | **BIC** | **CAIC** | **Npar** | **Items** | **BVR Mean (SD)** |
|---|---|---|---|---|---|---|---|---|---|
| Numbers of LCs | | M1 | 1 class | -3552.48 | 7233.66 | 7251.66 | 18 | Work experience & Union8 (26.40) | 2.84 (5.11) |
| | | M2 | 2 class | -3441.48 | 7147.52 | 7184.52 | 37 | Gender & Union8 (5.30) | 0.84 (1.26) |
| | | M3 | 3 class | -3418.58 | 7237.56 | 7293.56 | 56 | Union5 & Union7 (3.54) | 0.31 (0.61) |
| Covariate effects on latent trait | (a) | M4 | M2 allowing for class-specific effects of Work experience and Gender on latent trait | -3437.14 | 7153.12 | 7192.12 | 39 | Gender & Union8 (5.92) | 0.80 (1.28) |
| | | M5 | M4 + covariate effects of Work experience and Gender only in LC1 | -3439.10 | 7142.74 | 7179.74 | 37 | Gender & Union8 (5.96) | 0.78 (1.30) |
| Covariate effects on LC proportions | (b) | M6 | M5 + no effect of Gender on LC membership | -3440.30 | 7138.00 | 7174.00 | 36 | Union4 & Union5 (5.24) | 0.74 (1.14) |
| Measurement invariant items across LCs | (c) | M7 | M6 + 5 item discriminations (Union2 to Union6) constrained to equality across LCs | -3442.06 | 7105.76 | 7136.76 | 31 | Union4 & Union5 (5.60) | 0.71 (1.13) |
| | | M8 | M7 + 0 item discriminations for Union1 and Union8 in LC2 | -3444.12 | 7095.58 | 7124.58 | 29 | Union1 & Union4 (4.91) | 0.78 (1.10) |
| | (d) | M9 | M8 + 4 item locations (Union3 to Union5) constrained to | -3446.78 | 7072.30 | 7097.30 | 25 | Union1 & Union4 (5.78) | 0.82 (1.20) |

| Aims: To determine… | Step | Model | Model Description | Log-Likelihood | BIC | CAIC | Npar | Items | BVR Mean (SD) |
|---|---|---|---|---|---|---|---|---|---|
| | | | equality across LCs | | | | | | |
| Residual Observed DIF | (e) | M10 | M9 + class-specific effects of Gender on Union8 (Uniform DIF) | -3442.09 | 7077.23 | 7104.23 | 27 | Union4 & Union5 (5.23) | 0.71 (1.09) |
| | | ***M11*** | ***M10 + class-specific effect of Gender on Union8 only in LC2 (Uniform DIF)*** | ***-3442.11*** | ***7070.11*** | ***7096.11*** | ***26*** | ***Union4 & Union5 (5.28)*** | ***0.71 (1.09)*** |
| | | M12 | M9 + class-specific effects of Gender on Union 8 (Non-uniform DIF) | -3438.41 | 7084.16 | 7133.16 | 29 | Union4 & Union5 (5.37) | 0.68 (1.03) |
| | | M13 | M10 + class-specific effect of Gender on Union 8 only in LC2 (Non-uniform DIF) | -3439.30 | 7071.649 | 7098.649 | 27 | Union4 & Union5 (5.37) | 0.68 (1.03) |
| Differences in latent scores between LCs | (f) | M14 | M11 + effect of LC on latent trait | -3441.75 | 7076.54 | 7103.58 | 27 | Union1 & Union4 5.36) | 0.71 (1.09) |
| If freeing large BVR yields better fit | | M15 | M11 + Union4 & Union5 freely estimated | -3439.86 | 7072.77 | 7099.77 | 27 | Union1 & Union4 (3.72) | 0.54 (0.68) |

*Note.* Npar denotes the number of model parameters. The 2-class solution had the lowest BIC and CAIC values when determining the appropriate number of classes. The restricted 2-class solution showed even better fit. Based on the iterative pruning strategy, the final model chosen was Model11 (M11) as italicized; it has the lowest BIC and CAIC values, indicating parsimony. Further, the mean BVR values were very small, indicating good absolute fit. BVR mean (SD) denotes the average (standard deviation) of BVR values among the covariates and indicators.

Table 6.

*Eight Questions about DIF**

| **DIF Question** | **Analytic Model** | **Corresponding Parameters in Figure 1** |
|---|---|---|
| 1. Is there DIF between two observed groups (e.g., gender categories)? | Traditional IRT-DIF analyses; Restricted MM-IRT-C | Figure 1A, Paths 2 & 3 |
| 2. Is there DIF on a continuous observed variable (e.g., work experience)? | Restricted MM-IRT-C | Figure 1A, Paths 2 & 3 |
| 3. Is there DIF on one observed variable, after controlling for DIF on another observed variable (e.g., gender DIF after controlling for DIF on work experience)? | Restricted MM-IRT-C | Figure 1A (with 2 covariates), Paths 2 & 3 |
| | | |
| 4. Is there DIF between latent classes (unobserved DIF)? | MM-IRT;<br>MM-IRT-C | Figure 1B, Paths 5 & 6 |
| | | |
| 5. Are observed variables related to latent DIF classes (e.g., gender effects on latent class membership)? | MM-IRT-C | Figure 1C, Path 7 |
| 6. Are continuous measures related to latent DIF classes (e.g., is citizenship behavior related to latent class membership)? | MM-IRT-C | Figure 1C, Path 7 |
| 7. Is there observed-variable DIF within a latent DIF class (e.g., class-specific gender DIF)? | MM-IRT-C | Figure 1C, Paths 2 & 3 |
| 8. Are observed variables related to the latent trait within latent DIF classes (e.g., class-specific gender effects on the citizenship behavior construct)? | MM-IRT-C | Figure 1C, Path 1 |

*Note*. * All DIF questions can be applied to uniform DIF (item difficulty/location) and/or to non-uniform DIF (item discrimination). Restricted MM-IRT-C = Single-class IRT-C; MM-IRT-C = mixed measurement item response theory model with observed covariates.

## References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

Baker, F. (1995). EQUATE 2.1: Computer program for equating two metrics in item response theory.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331-348.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. In O. Opitz, B. Lausen & R. Klar (Eds.), *Studies in classification, data analysis, and knowledge organization*. Heidelberg: Springer-Verlag.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*, 133-148.

Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85*, 451-461.

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2*, 243-276.

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

Drasgow, F. (1987). Study of 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.

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology, 81*(5), 869-885.

Eid, M., & Rauber, M. (2000). Detecting measurement invariance in organizational surveys. *European Journal of Psychological Assessment, 16*, 20-30.

Eid, M., & Zickar, M. J. (2007). Detecting response styles and faking in personality and organizational assessments by mixed Rasch models. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models* (pp. 255-270). New York: Springer.

Glas, C. A. W. (1998). Detection of differential item functioning using Lagrange Multiplier tests. *Statistica Sinica, 8*, 647-667.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11*, 147-157.

Hernandez, A., Drasgow, F., & Gonzalez-Roma, V. (2004). Investigating the functioning of a middle category by means of a mixed-measurement model. *Journal of Applied Psychology, 89*(4), 687-699.

Hofstede, G. H. (1984). *Culture's consequences: International differences in work-related values* (abridged ed.). Beverly hills, CA: Sage Publications.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement, 32*, 261-276.

Kulas, J. T., & Finkelstein, L. M. (2007). Content and reliability of discrepancy-defined self-awareness in multisource feedback. *Organizational Research Methods, 10*, 502-522.

Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods, 3*, 186-207.

Li, F., Cohen, A. S., Kim, S.-H., & Cho, S.-J. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement, 33*, 353-373.

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

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Magidson, J., & Vermunt, J. K. (2004). Latent class models. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 175-198). Newbury Park, CA: Sage.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2005). Latent-trait latent-class anlaysis of self-disclosure in the work environment. *Multivariate Behavioral Research, 40*, 435-459.

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement, 32*, 611-631.

McShane, S. L. (1986). The multidimensionality of union participation. *Journal of Occupational Psychology, 59*, 177-187.

Mislevy, R. J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*, 195-215.

Muthén, B., & Muthén, L. K. (2007). *Mplus version 5.2* [Computer Program]. Los Angeles: Muthén & Muthén.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures. *Applied Measurement in Education, 2001*, 235-259.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495-502.

Raju, N. S. (1999). *DFIT5P: A Fortran program for calculating DIF/DTF* [Computer software]. Chicago: Illinois Institute of Technology.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353-368.

Reise, S.P., & Gomel, J.N. (1995). Modeling qualitative variation within latent trait dimensions: Application of mixed-measurement to personality assessment. *Multivariate Behavioral Research, 30*, 341-358.

Reise, S. P., & Waller, N. G. (1990). Fitting the two-parameter model to personality data. *Applied Psychological Measurement, 14*, 45-58.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*, 643-671.

Robert, C., Lee, W. C., & Chan, K.-Y. (2006). An empirical analysis of measurement equivalence with the INDCOL measure of individualism and collectivism: Implications for valid cross-cultural inference. *Personnel Psychology, 59*, 65-99.

Rost, J. (1990). Rasch model in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*, 271-282.

Rost, J. (1991). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44*, 75-92.

Rost, J., Carstensen, C., & von Davier, M. (1997). Applying the mixed Rasch model to personality questionnaires. In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences*. Munster, Germany: Waxman.

Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practices approach. *Organizational Research Methods, 6*, 169-215.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.

Smit, J. A., Kelderman, H., & van der Flier, H. (1999). Collateral information and mixed Rasch models. *Methods of Psychological Research, 4*, 19-32.

Smit, J. A., Kelderman, H., & Van der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *Methods of Psychological Research, 5*, 1-13.

Stark, S. (2002). *ITERLINK: Iterative linking and pairwise DIF detection for the 3PL model using Lord's chi-square* [computer program]: Department of Psychology, University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., Chan, K.-Y., Lee, W. C., & Drasgow, F. (2001). Effects of the testing situation on item responding: Cause for concern. *Journal of Applied Psychology, 86*, 943-953.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item (functioning and differential) test functioning on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89*, 497-508.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

Swaminathan, H., & Rogers, J. H. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). New Jersey: Hillsdale.

Triandis, H., & Gelfand, M. J. (1998). Converging measurement of horizontal and vertical individualism and collectivism. *Journal of Personality and Social Psychology, 74*, 118-128.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70.

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD 4.0 User Manual*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont Massachusetts: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2008). *Latent GOLD 4.5* [computer program]. Belmont, MA: Statistical Innovations Inc.

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data. Results on a Monte Carlo study. *Methods of Psychological Research Online, 2*, 29-48.

Wang, M. (2007). Profiling retirees in the retirement transition and adjustment process: Examining the longitudinal change patterns of retirees' psychological well-being. *Journal of Applied Psychology, 92*, 455-474.

Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research, 44*, 1-27.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*, 320-330.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*, 168-190.

## Appendix

Latent GOLD 4.5 syntax for specifying:

*The final single-class/restricted MM-IRT (IRT-C) Model 6 (M6):*

```
(1) theta ;
theta <- Z1 + Z2;                 /*Z1 and Z2 represent the covariates Work experience and Gender respectively
Y1 <- 1 + theta + Z1;             /*Uniform DIF of Work experience on Union1
Y2 <- 1 + theta;
Y3 <- 1 + theta;
Y4 <- 1 + theta;
Y5 <- 1 + theta;
Y6 <- 1 + theta;
Y7 <- 1 + theta + Z1;             /*Uniform DIF of Work experience on Union7
Y8 <- 1 + theta + Z1 + Z1 theta;  /*Non-uniform DIF of Work experience on Union7
```

*The final MM-IRT-C Model 11 (M11):*

```
(1) theta ;
theta <- (g1) Z1 | Class + (g2) Z2 | Class;  /*class-specific Z1 and Z2 effects on latent trait
Class <- 1 + Z1;                             /* Z1 predicting latent class proportions
Y1 <- 1 | Class + (b1) theta | Class;        /*Non-uniform unobserved DIF on Union1; zero item discrimination
                                             /*in LC2
Y2 <- 1 | Class + theta;                     /*Uniform unobserved DIF
Y3 <- 1 + theta;                             /*Measurement equivalence across Union4 to Union6
Y4 <- 1 + theta;
Y5 <- 1 + theta;
Y6 <- 1 + theta;
Y7 <- 1 | Class + theta | Class;             /*Non-uniform unobserved DIF on Union7
Y8 <- 1 | Class + (b8) theta | Class + (c8) Z2 | Class; /*Non-uniform unobserved DIF on Union8 ; zero item
                                                        /*discrimination in LC2; residual observed DIF of Gender
                                                        /*on Union8
g1[2] = 0;                                   /*Specify constraints on model parameters
g2[2] = 0;
b8[2] = 0;
b1[2] = 0;
c8[1] = 0;
```

## Endnotes

[1] By restricting the MM-IRT-C model to only one latent class, it becomes an IRT with covariates model as seen in Figure 1A.

[2] The plots of item response functions are based on estimates from the software EQUATE 2.1 (Baker, 1995), as the largest number of DIF items was detected with the DFIT methodology.

[3] A reviewer raised a concern that the scale may be multidimensional and the two-class solution may be a result of multidimensionality (Reise & Gomel, 1995). Although we found unidimensionality in our CFA procedure, we tested the reviewer's hypothesis by fitting a two-dimensional IRT-C model in which latent trait variances were set to 1 and items freely loaded onto both orthogonal traits. The latent traits were regressed onto both the covariates Gender and Work Experience. Results from the information criteria showed that the multidimensional IRT-C model (BIC = 7202.16; CAIC = 7230.16) did not fit as well as the one dimensional IRT-C model with DIF (BIC = 7165.21; CAIC = 7187.21), or the unconstrained two-class solution (BIC = 7147.52; CAIC = 7184.52).