Running Head: MULTILEVEL MIXED-MEASUREMENT ANALYSIS

Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions around the world

Louis Tay

University of Illinois at Urbana-Champaign

Ed Diener

University of Illinois at Urbana-Champaign and The Gallup Organization

Fritz Drasgow

University of Illinois at Urbana-Champaign

Jeroen K. Vermunt

Tilburg University, The Netherlands

Address for Correspondence:
Louis Tay
Department of Psychology
University of Illinois at Urbana-Champaign
603 East Daniel Street
Champaign, Illinois 61820
Email: sientay@illinois.edu

Abstract

Dimensional approaches assume that all individuals within hierarchical units (e.g., organizations, or countries) share the same measurement model. However, such models are less applicable when researchers are interested in obtaining classes of individuals who share the same measurement model across hierarchical units and to obtain hierarchical latent classes. We present the multilevel mixed-measurement item response theory (MMM-IRT) model as an alternative. This model yields classes of individuals with a common measurement model that span across hierarchical units. Also, hierarchical units are classified together to the extent that they share similar proportions of individual-level classes. We illustrate the MMM-IRT model with data on self-reported emotions from 121,740 individuals across 116 countries where four individual-classes and five country-classes were found. Theoretical and methodological implications concerning cross-cultural, multilevel and measurement equivalence research are discussed.

Multilevel mixed-measurement IRT analysis: An explication and application to self-reported emotions around the world

Organizational science has routinely used dimensional approaches such as factor analysis (FA) (Spearman, 1904) and item response theory (IRT) (Lord & Novick, 1968) for many important purposes such as construct measurement (Cronbach & Meehl, 1955) and the examination of measurement equivalence (ME) between groups (Drasgow, 1984, 1987; Vandenberg & Lance, 2000). Typical measurement approaches rely on observed groupings, assuming that a single measurement model holds for the population. However, methodological advances have led to a synthesis of dimensional and latent class (LC) approaches, enabling one to infer latent groupings that share distinct measurement models (Rost, 1990, 1991; Rost, Carstensen, & von Davier, 1997). In contrast to using a priori groups, a bottom-up approach can be applied to uncover the different measurement classes that exist on the construct(s) of interest. Given that organizational phenomenon is inherently hierarchical (Kozlowski & Klein, 2000; Roberts, Hulin, & Rousseau, 1978), a multilevel approach – known as multilevel mixed-measurement IRT analysis (MMM-IRT) -- can be used for nested data. This yields classes of individuals with a common measurement model that spans across hierarchical units, while taking into account nested dependencies. Further, hierarchical units are classified together to the extent that they share similar proportions of individual-level classes, resulting in latent classes at different levels of conceptualization. This model is new in that it has only been recently proposed (Cho & Cohen, in press), but it falls within the general latent variable modeling framework implementable in latent variable software packages (B. Muthén & L. Muthén, 2007; Vermunt & Magidson, 2000a).

Conceptually, a distinctive feature of the MMM-IRT is the organic derivation of measurement groupings on constructs of interest (e.g., personality, attitudes, climate) that are not bounded by observed hierarchical units. Not only can we uncover how individuals potentially differ in subjective construct definition and/or scale usage (e.g., Eid & Rauber, 2000; Hernandez, Drasgow, & Gonzalez-Roma, 2004; Maij-de Meij, Kelderman, & van der Flier, 2005, 2008; Zickar, Gibby, & Robie, 2004), but we can also explore commonalities among hierarchical units. The MMM-IRT model can not only contribute to organizational research by availing new methodological possibilities to address theoretical issues, but also foment new conceptual and methodological developments. We list some applications to, and generative questions for organizational topics in Table 1, which includes cross-cultural research, multilevel issues, and measurement equivalence procedures. These issues will be elaborated in greater detail in the discussion.

Given these theoretical and methodological implications, it is important for organizational researchers to consider the use of the MMM-IRT model. The article is structured as follows. Foremost, we explicate the MMM-IRT model and its statistical assumptions. Second, we elaborate on issues of observed and unobserved heterogeneity that are relevant to a conceptual understanding of MMM-IRT and model specification. Third, we illustrate the use of MMM-IRT on a large data set that consists of self-reported positive and negative emotions of 121,740 individuals across 116 countries, showcasing its utility in obtaining interpretable individual-level measurement classes and country-level classes. Finally, we discuss how the MMM-IRT model can contribute to organizational research, and the theoretical and methodological issues related to its application.

Multilevel mixed-measurement IRT

In this section, we introduce both IRT and LC models at the onset because they are foundational for understanding MMM-IRT and its assumptions. The IRT model is a dimensional model, assuming a continuous latent trait; in contrast LC models assumes a latent categorical variable. The integration of both approaches leads to the mixed-measurement IRT model, whereby qualitatively distinct dimensions are inferred such that qualitative differences exist between classes (i.e., different measurement models) and quantitative differences hold within classes (cf. Hernandez et al., 2004). For applications to hierarchical data, a multilevel conception of mixed-measurement IRT is used.

Item Response Theory. A dimensional approach commonly used by organizational psychologists is IRT (Hulin, Drasgow, & Parsons, 1983). IRT specifies the probability of endorsing ('1') a dichotomous item y_{ji} given the continuous latent trait θ_j, where *j* and *i* index the persons and items respectively. The two-parameter logistic (2PL) model is

$$P(y_{ji} = 1 \mid \theta_j) = \frac{1}{1 + \exp(-\eta_{ji})}, \qquad (1)$$

where the linear term η_{ji} is equal to $\beta_i + \lambda_i \theta_j$[1]. The coefficients β_i and λ_i represent item easiness (i.e., item location) and discrimination parameter respectively. Note that these two parameters are similar to an item intercept and an item loading in factor analysis, where the main difference is that we use a logistic function here instead of a linear model for the items. For consistency in the paper, we will refer to the parameters β_i and λ_i as item intercepts and loadings respectively. The probability of all *I* responses of person *j* ($\underset{\sim}{y}_j$) is obtained under the assumption that these are independent of one another given θ_j. This assumption is often referred to as local independence; that is, $P(\underset{\sim}{y}_j \mid \theta_j) = \prod_{i=1}^{I} P(y_{ji} \mid \theta_j)$. The marginal probability $P(\underset{\sim}{y}_j)$ is obtained by integrating over all possible values of the unobserved θ_j.

Latent Class Model. On the other hand, the basic LC model is

[1] We note that the linear term can be re-parameterized into the standard 2-parameter logistic IRT model, $P(y_{ji} = 1 \mid \theta_j) = \dfrac{1}{1 + \exp[-a_i(\theta_j - b_i)]}$ where the item discrimination and difficulty are $a_i = \lambda_i$ and $b_i = -\beta_i / \lambda_i$ respectively. However, we choose to present this formulation because in our illustration, we use a multidimensional latent trait model where no simple transformation exists.

$$P(\underset{\sim}{y}_j) = \sum_{k=1}^{K} \pi_k P(\underset{\sim}{y}_j \mid k) . \tag{2}$$

This equation shows that the marginal probability of an individual's responses $\underset{\sim}{y}_j$ to a set of dichotomous items *I* given discrete latent classes k, k =1,…,K, depends on two terms: (a) probability π_k of belonging to latent class *k* and (b) the conditional probability $P(\underset{\sim}{y}_j \mid k)$ of the responses to the set of *I* items given latent class *k*. Note that the latent class probabilities sum to one ($\sum_{k=1}^{K} \pi_k = 1$). Similar to IRT models, the probability $P(\underset{\sim}{y}_j \mid k)$ is restricted by assuming that the *I* responses are locally independent of one another (here, mutually independent given a person's class membership); that is, $P(\underset{\sim}{y}_j \mid k) = \prod_{i=1}^{I} P(y_{ji} \mid k)$.

Mixed-measurement model. In general, the mixed-measurement model describes a class of models that integrates both dimensional and LC approaches (Rost, 1990, 1991). The mixed-measurement model identifies latent classes in which each class of individuals shares the same set of item parameters. Thus, there is ME within classes, but non-equivalence between classes. Specifically, an underlying continuous variable is posited to account for covariation among items within a class. The mixed-measurement model takes the LC form shown in equation (2), however, the conditional probability of endorsing the item depends not only on the latent class, but also on the latent trait θ_j: $P(y_{ji} = 1 \mid k, \theta_j)$. Hence, for a set of items, the probability $P(\underset{\sim}{y}_j \mid k)$ of the response vector for the *j*th person can be rewritten as $\int \prod_{i=1}^{I} P(y_{ji} \mid k, \theta_j) f(\theta_j) d\theta_j$, where $f(\theta_j)$ is the standard normal density. For a 2PL response model, the probability of endorsement becomes,

$$P(y_{ji} = 1 \mid k, \theta_j) = \frac{1}{1 + \exp(-\eta_{jik})}, \tag{3}$$

where linear term $\eta_{jik} = \beta_{ik} + \lambda_{ik}\theta_j$ is allowed to differ across latent classes, which is achieved by allowing the item intercept and loading to be class specific (see index *k*). This is the mixed-measurement IRT (MM-IRT) model proposed by Smit, Kelderman & Van der Flier (2000). For more details on the MM-IRT model, please refer to the article by Tay, Newman & Vermunt (in press) in this issue.

Multilevel mixed-measurement IRT. Multilevel extensions have been developed for IRT models (Fox & Glas, 2001) as well as for latent class models (Rabe-Hesketh, Pickles, & Skrondal, 2001; Vermunt, 2003; Vermunt & Magidson, 2000b). The multilevel extension of the MM-IRT model we propose here is strongly related to these two developments. It fact, it can be seen as an extension of three different models proposed in the literature: (a) MM-IRT is extended by modeling hierarchical classes; (b) the multilevel latent class model (MLC) of Vermunt (2003; 2008a) which posits individual- and hierarchical-level classes is extended by modeling underlying traits at the lower-level; and (c) the multilevel mixture model of Vermunt (2008b), where hierarchical-classes and individual-level traits are estimated, is extended by additionally estimating individual-level classes. Whereas each of these models contains two of the key elements of our model – individual classes, individual traits, and hierarchical classes – the MMM-IRT contains all three. It should be noted that the model we propose here has only been recently proposed by Cho and Cohen (in press) as well. Although the proposal is new, it fits within the general latent variable modeling framework implementable by the software MPlus (B. Muthén & L. Muthén, 2007) and Latent GOLD (Vermunt & Magidson, 2000a, 2005, 2008). Other potential models (not exhaustively) specifiable in the general latent variable modeling framework are presented in Table 2. The type of model one would specify depends in part on the nature of the data and the theoretical issues one wishes to explore. In this case, we have nested data which are dichotomously scored, and would like to explore latent classes of individuals who share the same measurement model across different countries.

We present a graphical representation of the MMM-IRT model in Figure 1 which shows that the nested data structure can be conceptually accounted for by hierarchical and individual latent classes; further, responses are accounted for by an IRT item response model. Some results from our illustration are added to facilitate understanding of the model.

The statistical assumptions of the MMM-IRT model are:

(1) As with mixed-measurement IRT, latent classes for individual-level units have distinct measurement models. As elaborated by Vandenberg and Lance (2000), differences in measurement models reflect differences in the epistemic definition of the construct such that scores are not comparable between latent classes because of qualitative differences in their frames-of-reference; in contrast, within-class comparisons can be made. The formal statistical assumption is that the item responses of an individual are mutually independent conditional on his/her traits and class membership.

(2) Nested dependencies are handled by assuming that hierarchical units influence the probabilities of individual-level LCs, but do not affect the item responses directly. In other words, the effect of belonging to a hierarchical unit on an item response is fully mediated by the individual-level class membership. It is assumed that hierarchical units differ in the *probability* of belonging to individual-level latent classes, which means that the influence from hierarchical units is modeled as probabilistic in the sense that, say, in an organizational setting, departmental characteristics may influence the views of some individuals, but not all. Similarly, it is assumed that in cross-cultural settings, country characteristics do not exhaustively modify individual experiences such that *all* individuals within a country view the psychological construct in a similar manner. For example, it is assumed that cross-group influences (e.g., globalization

forces) and shared biological foundations among individuals can result in common individual-level latent classes that span across hierarchical units.

(3) Hierarchical-level units are assumed to belong to one of *G,* where $g = 1,\ldots,G$, hierarchical-level latent classes. Hierarchical-units are similar to the extent that they share similar proportions of individual-level latent classes, a commonality between hierarchical-units that can be accounted for by the use of a nominal latent variable at the hierarchical level; that is, by assuming that they belong to the same hierarchical latent class. As an example, organizational units belonging to the same organizational-level latent class are similar because they have comparable proportions of individual-level latent classes. To make it more specific, consider that the sales department in an organization may have 80% of individuals described by measurement model 1 on job satisfaction, and 20% of individuals described by measurement model 2. These proportions are similar to the advertising and public relations departments within the organization and will be classified in the same hierarchical-level latent class. However, the accounting department consists of 20% of individuals described by measurement model 1 on job satisfaction, and 80% of individuals described by measurement model 2. Because these proportions are similar to the engineering department, both these organizational units will be classified in the same hierarchical-level latent class, which is distinct from sales, advertising and public relations.

This description should provide sufficient information for the use of MMM-IRT in practice. Readers interested in learning more about the statistical details are referred to the Appendix.

The issue of population heterogeneity

To fully develop the conceptual underpinnings of MMM-IRT, we discuss the issues of observed and unobserved heterogeneity, borrowing in part from Lubke and Muthén (2005).

Distinguishing these two forms of heterogeneity can aid understanding of how heterogeneity is used and represented at multiple levels in MMM-IRT.

A clarification of what the term “heterogeneity” means is important here. Heterogeneity in this paper refers to *distinct subpopulations*, which does not necessarily correspond with score variability. For instance, a researcher may collect data from a heterogeneous population consisting of two subpopulations – male and female. Consequently, on the attribute of gender, these subpopulations are homogeneous. In general, organizational data are sampled from a heterogeneous population, in which two forms of heterogeneity are present: observed and unobserved (see Skrondal & Rabe-Hesketh, 2004). Observed heterogeneity within a population can be demarcated by the use of observed variables that are known a priori. For instance, demographic variables enable one to pre-specify gender or racial subpopulations, where multiple-group comparisons can be made. Based on legal and socio-political contexts (e.g., majority versus minority group), research designs (e.g., experimental versus control group), or theoretic considerations (e.g., country A versus country B), observed heterogeneity is frequently used for describing and discerning subpopulations.

On the other hand, unobserved heterogeneity has to be inferred from the data because it is not known which subpopulation an individual belongs; one may lack informative observed variables, or choose not to predefine the subpopulations with observed variables and let the “data speak for itself” as it were. The latter rationale prioritizes the modeling of underlying phenomenon. Because unobserved heterogeneity is given precedence over observed heterogeneity, it is arguably a more direct approach for *uncovering subpopulations on a construct of interest*. In the case of mixed-measurement modeling (Rost, 1990, 1991), the patterns of item responses are used to infer the underlying subpopulations which are represented as underlying subgroups that share the same measurement model. Not unexpectedly, these underlying subpopulations may, or may not correspond to observed heterogeneity. For instance, there may not be an exact correspondence between subpopulations that share the same attitude measurement class and gender subpopulations (cf. Eid & Rauber, 2000). Indeed, the relation between observed and unobserved heterogeneity is akin to ideas of surface-level diversity (e.g, demography) and deep-level diversity (i.e., psychological constructs like attitudes (e.g., Harrison, Price, & Bell, 1998); subpopulations that are differentiated on measured constructs need not necessarily be distinguished by manifest categories.

While ME procedures only utilize observed sources of heterogeneity and mixed-measurement modeling (i.e., MM-IRT) captures only unobserved sources of heterogeneity, MMM-IRT considers both types of heterogeneity. The observed source of heterogeneity in this case consists of hierarchical units (e.g., teams, organizations, countries); within multilevel research, such groupings have theoretical significance (Klein, Dansereau, & Hall, 1994), and are also commonly associated with between-group variation and nested dependencies (Raudenbush & Bryk, 2002).

Unobserved heterogeneity is conceptualized in MMM-IRT at two levels: the individual-level and the hierarchical-level. At the individual-level, latent subpopulations are not perfectly demarcated by hierarchical groups. For example, consider that even though organizational hierarchies (e.g., teams, departments) serve as a convenient way of partitioning individuals who may share similar perceptions, informal networks and interconnections within the organization could also affect shared perceptions (Newman, Hanges, Duan, & Ramesh, 2008). Thus, MMM-IRT utilizes information from observed hierarchical groupings but simultaneously posits subpopulations of individuals that share the same measurement model.

At the hierarchical-level, unobserved heterogeneity is also assumed. This heterogeneity is represented as distinct hierarchical-level latent classes, and is reckoned to impinge upon the proportions of individual-level subpopulations within each hierarchical unit. In other words, there are unobserved clusters of hierarchical units which are internally homogeneous, but qualitatively distinct externally from other clusters of hierarchical units.

At this juncture, we note that although the uncovering of unobserved heterogeneity is essentially a bottom-up procedure where latent groupings are inferred, the procedure is ultimately based upon a preliminary conceptual model of the psychological/organizational phenomenon presupposed by the researcher. Such conceptual models are a "workable approximation of reality" (Wedel, 2002, p. 182) that need to be validated and compared with other possible conceptual configurations. For instance, a researcher may propose that two individual-level subpopulations exist, but will need to compare this proposition with other alternatives, like models with three or four individual-level subpopulations. This is analogous to exploratory factor analysis where one compares solutions across differing the numbers of factors.

Another issue related to model-specification is that there are two distinct types of conceptualizations for MMM-IRT models in general as shown in Table 2. Instead of a hierarchical-level latent class, it is also possible to conceptualize a hierarchical-level dimension where hierarchical-units are ordered on a continuum. In other words, rather than positing that latent classes undergird hierarchical units, one could assume that hierarchical units lie on a continuous dimension. Where researchers are interested in these competing configurations, model comparisons can be undertaken to compare if either type fits the data better. However, because this special issue focuses on latent class methodologies, we only present the former. Additionally, conceptualizing MMM-IRT with a dimensional hierarchical-construct warrants a separate explication, with different theoretical and methodological implications for organizational research. Interested readers are directed to several excellent sources in marketing research that discuss representing unobserved heterogeneity as discrete (i.e., latent classes) or continuous distributions (i.e., latent dimension) (Allenby & Rossi, 1999; Wedel & Kamakura, 2000; Wedel et al., 1999).

An illustrative application of MMM-IRT

In this section, we illustrate the different issues and judgment calls that arise in the practice of using MMM-IRT. We apply the MMM-IRT model to self-reported emotions around the world because the study of emotions is steadily gaining interest from organizational researchers (e.g., Weiss & Cropanzano, 1996), and there is an increasing theoretical interest in how culture influences emotions (Markus & Kitayama, 1991; Mesquita & Walker, 2003). Recent research has focused on the similarity and differences of emotion structure across countries using factor analytic or multidimensional scaling approaches (e.g., Church, Katigbak, Reyes, & Jensen, 1999; Russell, Lewicka, & Niit, 1989; Yik & Russell, 2003). Following this trend, we explore the structure of self-reported emotions, but instead of using cross-country comparisons, we empirically derive individual-level measurement classes and hierarchical-level country classes; for ease of reference, we term these classes as individual-classes and country-classes respectively.

For individual-classes, we examine the possible structures of self-reported emotions via a multidimensional measurement model. This is because the structure of self-reported emotions is generally thought to be described by two orthogonal dimensions: pleasantness and activation (see the review by Russell & Feldman Barrett, 1999). Indeed, these two dimensions accounted for 73% to 97% of variance (Green, Goldman, & Salovey, 1993) in 4 different models of affect (Larsen & Diener, 1985; Russell, 1980; Thayer, 1978; Watson & Tellegen, 1985). Although the emotion circumplex as seen in Figure 2 is thought to hold for all individuals, we explored whether potentially different individual-classes (i.e., different self-reported emotion structures) can be found. Further, because the application of MMM-IRT also yields homogenous country-classes, we explored whether these country-classes are interpretable, indeed, whether countries appear to share similar socio-cultural roots, economic backgrounds or are connected via geography.

Data. The Gallup World Poll collected representative sampling of the world from 138,666 individuals in 134 countries in 2005-2006 (Gallup, 2009, Aug 28). For the purposes of our analysis, only countries that were administered the variables that tapped into positive and negative feelings were used. This resulted in a sample size of 121,740 individuals from 116 countries. There were at least 500 individuals sampled within each country, with a mean sample of 1049.48 (SD = 311.94) per country.

For each of the affect items, respondents were asked to indicate whether they experienced any of the following feelings on the previous day ("No" = 0; "Yes" = 1). Where respondents refused to answer or did not know the answer to the question, the response was coded as missing. Table 3 shows the means, standard deviations and percentage missing in each of the ten affect variables. Overall, there was a low percentage of missing values. To use all information available in the response patterns for the subsequent analyses, listwise deletion was not undertaken. Instead, missing values were dropped out of the likelihood equation in the estimation process. This procedure is recommended for dealing with responses that are missing at random and where maximum likelihood estimation is used (Rubin, 1976).

Additionally, to help interpret the obtained individual-level measurement classes, we utilized several individual-level external variables. This included a 3-item global life evaluation measure on past, present and future Life Satisfaction (Cantril's Self-Anchoring Striving Scale, 1965), with response options ranging from 0 (worst possible life) to 10 (best possible life) (Cronbach's $\alpha = .73$). Further, prior day experiences of "love" and "physical pain", scored dichotomously ("No"= 0; "Yes"= 1), and demographic indicators age and gender (Male = 1, Female = 0) were included. The purpose was to (a) examine mean level differences among individual-level latent classes on these variables and (b) use an adaptation of property vector fitting (PVF; Kruskal & Wish, 1978) to describe the relationship between the affect structure axes (valence and activation) and the external variables for each latent class. This procedure will be described in greater detail in the results section.

Analysis

Estimation. The parameters in the MMM-IRT model were estimated with the software Latent GOLD 4.0 using maximum likelihood (ML) estimation (Vermunt & Magidson, 2000b). We note that it is also possible to use MPlus software to conduct the analyses with ML estimation as well (L. Muthén & B. Muthén, 1998-2007); interested readers are directed to Chapter 10 of the MPlus user manual.

Unlike procedures recommended by Steenkamp and Ter Hofstede (2002) for obtaining international market segmentation where country-classes are obtained, no re-weighting was implemented because countries were treated as a unit of interest. Weighting by world population sizes would lead to results that are biased toward China and India because these country population sizes far outweigh all other countries combined.

Technical specifications. Latent GOLD 4.0 allows a choice of different information matrices, and we chose to compute the default Hessian matrix. This is necessary to estimate the standard errors of the item parameters and to determine if items are statistically distinguishable among classes. To reduce the likelihood of obtaining local minima, 20 random sets of starting values were used instead of the default of 10.

Determining the number of latent classes. To determine the number of individual-classes, the MMM-IRT model was first fitted with successively incremented numbers of classes using Latent GOLD. Model comparisons were based on the Bayesian information criterion (BIC) (Schwarz, 1978). This index is commonly used in evaluating the relative fit of LC models (Hagenaars & McCutcheon, 2002) and mixed-measurement models (Lubke & Muthén, 2005; Magidson & Vermunt, 2004) and has been found to perform well in recent Monte Carlo simulations pertaining to multilevel latent class models (Lukociene, Varriale, & Vermunt, under review; Lukociene & Vermunt, in press). In particular, when the number of observations per hierarchical unit is more than 15 and the number of hierarchical-units is large (> 50), the BIC performs well. Based on this line of evidence, we relied on the BIC in the present analyses to make decisions about the number of latent classes.

We first set the number of country-classes to one, and obtained the best fitting number of individual-classes based on the lowest BIC value. The number of country-classes was subsequently determined by fixing the best fitting number of individual-classes, and increasing the number of country-classes until the lowest BIC value was reached. This procedure follows that used in other multilevel LC examples (Vermunt, 2003).

After determining the best fitting model on the BIC, local indices of absolute fit were evaluated by examining the bivariate residuals (BVR), which are analogous to modification indices within the framework of structural equation modeling (SEM). The BVR is a Pearson χ^2/df statistic in which observed frequencies in a two-way table are compared to the model-based frequencies (with binary items it is just Pearson χ^2). In general, BVR values much larger than 1 or 2 (Vermunt & Magidson, 2000) would indicate misfit. It has been recommended that local dependencies can be specified until the largest BVR is reasonably small. Note that the BVR statistic is a χ^2 value and is sensitive to sample size. Because the analyzed sample size was large (*n*=121,740), we used an adjusted value BVR$_{\text{adj}}$[2].

Endorsement profiles and the dimensional representation of emotions. Latent GOLD produced the endorsement probabilities for each class which were plotted for a graphical comparison. Researchers may be interested in examining whether endorsement profiles are

[2] The expected value of a non-central chi-square is equal to its *df* plus *n* times its noncentrality parameter δ, $E(\chi^2) = df + n\delta$.Thus, an estimate of the noncentrality parameter is $\hat{\delta} = (\chi^2 - df) / n$. An observed χ^2 can be adjusted to fixed sample size of, say, n = 500, by $\chi_{500}{}^2 = [df + 500(\chi^2 - df)/n]$. In this case, $BVR_{adj} = [df + 500([BVR \times df] - df)/n]/df$, where *n* =121,740 and *df* = 2.

significantly different from one another. Although this approach seems intuitive, the purpose of the MMM-IRT analysis is to examine if the measurement models are significantly different. We recommend that item parameters estimates be compared to determine if the intercepts and loadings are significantly different across the estimated latent classes. In Latent GOLD, we can examine the Wald statistic for each item intercept and loading. Significant differences would indicate that items differ in functional form across classes; that is, item parameters are significantly different across classes. Because the item intercepts are related to ease of endorsements, they are indirectly related to endorsement profiles. By examining the endorsement profiles, we can get a sense of these differences on a probability rather than a logit scale.

To obtain the dimensional representation of emotion terms that are interpretable across the different latent classes, several steps were taken. First, the negative affect variable "sadness" was constrained to have zero loading on the first dimension (i.e., $\lambda_{1ik} = 0$ where i = "sadness"). This type of constraint is necessary for model identification in a two-dimensional model (see Vermunt & Magidson, 2005, p. 82). However, if researchers are positing a one dimensional measurement model, such constraints are not necessary for model identification.

Second, the loadings were rotated in the two-dimensional space so that the variable "sadness" rests on the negative pole of the valence dimension. All other items were rotated in the same manner so that the relative configuration remained unchanged. This procedure should yield the appropriate valence and activation dimensions on the x- and y-axis respectively and was applied to each latent class. Interpreting the activation dimension is more complicated because the configuration may be reflected about the valence dimension without changing the position of "sadness" after rotation. However, whether a reflection is necessary can be resolved by regressing external variables, like Life Satisfaction and love (which should be positively activated), on activation scores. This procedure was undertaken to ensure that the loading plots had the same direction.

Results

As shown in Table 4, the BIC criterion showed that the best fitting model was 4 individual-classes and 5 country-classes. An examination of the largest BVR_{adj}= 1.37 and the average BVR_{adj}=1.04 indicating good local fit and overall absolute fit, respectively. Further, the Wald statistics indicated that all the item intercepts and loadings were significant across the different individual-classes ($p < .01$) signaling that different measurement models, and hence affect structures, underlie the data.

Individual-classes

As seen in Table 4, 4 individual-classes were appropriate for our data. The proportions π_k of the 4 individual-classes were 0.28, 0.28, 0.23 and 0.21. This indicates that there were four distinct types of emotion experience, roughly in equal proportions. Each type is represented by a different measurement model. The endorsement profiles and the dimensional representations of the emotion terms are depicted in Figures 3 and 4 respectively[3].

To understand the differences among these individual-classes, external variables were also used. Foremost, mean values for external variables were obtained to describe each class as embedded in the dimensional plots in Figure 4. We note here that gender (i.e., proportions of males) did not appear to differ much among the individual-classes; hence, we do not focus on this variable. Using an adaptation of PVF, for each individual-class, external variables were regressed on latent trait scores and the resultant standardized coefficients were used to plot

[3] We note that because an item response parameterization following Smit, Kelderman & Van der Flier was used (2000), the differences in item intercepts and loadings may additionally reflect differences in means and variances on the two latent dimensions rather than measurement non-invariance. For comparability across classes, rescaled item parameter estimates were used to take into account differences in the latent means and variances so that item loading plots are comparable. Loadings for each trait were rescaled for classes two and higher such that their mean squares equaled the value for class one and intercepts were rescaled such that their mean equaled the value for class one.

vectors within the item loading plots as seen in Figure 4[4]. Each vector reveals the direction and strength of the relationship between the emotion axes and the external variable. In general, there were substantial commonalities among individual-classes on the external vectors: (a) the experience of physical pain was consistently related to negative valence, or unpleasantness, but hardly at all to activation; (b) life satisfaction and the experience of love were both related to positive valence and activation. In particular, experiencing love had higher activation than life satisfaction; (c) age was related to lower activation, but class differences emerged on its relation to positive or negative valence.

Individual-class 1. From the endorsement profile, this individual-class was relatively low on self-reported positive emotions, particularly on “smiling/laughing” and “enjoyment”. Instead, reports of “anger” and “shame” were above average. On the other hand, this class of individuals reported less “stress” as compared to other individual classes. The dimensional representation showed that all items were positively loaded on activation. Thus, the self-reported emotions sampled in this study were activated in the same direction.

The means of external variables revealed that this class consisted of younger individuals, and their life satisfaction and experience of love was the lowest among all the classes. The external vectors showed that age was related to lower activation and negative valence. Thus, older individuals experience more deactivated negativity. Life satisfaction had the smallest relation to both activation and valence as compared to the other individual-classes, implying that life satisfaction was less related to affect in this individual-class.

Individual-class 2. Unlike individual-class 1, the endorsement profile showed the highest endorsements on both positive and negative emotions. The dimensional representation however, showed similar trends to the first individual-class. Most item loadings were spread in

[4] These regression coefficients were also rescaled to reflect differences in latent trait variances across individual-classes so that the vector directions and lengths are comparable graphically. Specifically, the regression coefficients were divided by the standard deviation of the latent traits.

a similar manner, but there was a slightly larger spread on the activation dimension, in particular because “smile” had higher activation. “Stress” was not activated, unlike individual-class 1.

In comparison to other individual-classes, this group was moderately high on life satisfaction and had the highest endorsements of experiencing love, but also highest endorsements of physical pain. External vectors revealed that age, as in individual-class 1, was related to deactivation and negativity as well. Further, gender differences were apparent in this group; being male in this individual-class was related to positive valence. Conversely then, females were more likely to experience unpleasant emotions. Physical pain had the strongest relation to negative valence when compared with other classes. Thus, feelings of unpleasantness were strongly tied to the experience of physical pain.

Individual-class 3. Individuals within this class were likely to endorse positive emotions in general. Relative to other classes, individuals reported the most “enjoyment”. However, moderate amounts of negative emotions were also encountered. Although relatively high on “stress”, individuals were least likely to experience “depression” or “shame”. Similar to the first two individual classes, the dimensional representation revealed that positive and negative emotions straddled the two ends of the valence dimension. Interestingly, “pride” was closer to the center of the dimensional space, indicating that this emotion term was slightly positive compared to the other terms. Unlike individual-classes 1 and 2, all emotion terms had a substantial spread on activation with terms like “depression” and “respect” deactivated. Another distinguishing feature of individual-class 3 is that “sadness” did not anchor the end of the unpleasantness continuum, with terms like “anger” and “depression” perceived as more unpleasant.

Given the high endorsements of positive emotions, it is not surprising that this individual-class had the highest levels of life satisfaction, high endorsements of love and low endorsements of physical pain. This group of individuals was on average older than the other classes, and the age vector showed that while age was related to deactivation, it was also related to pleasantness rather than unpleasantness. Older individuals in this class would expect to experience more pleasant feelings.

Individual-class 4. For the most part, this class of individuals reported relatively moderate positive and negative emotions. However, the endorsements of “treated with respect”, “pride” and “worry” were the lowest. The dimensional plot showed that a clear trend whereby positive emotion terms were more activated and negative emotion terms were less activated. In particular, negative terms had a very low spread on activation. One explanation may be that this group of individuals confounds activation with pleasantness, or that the individuals have neutral activation for unpleasant emotions in general. “Treated with respect” and “pride” fell close to the middle on the valence dimension, showing that these items were moderately pleasant, but not regarded as extremely pleasant, unlike enjoyment.

Individuals in this class had moderate life satisfaction and love, but relatively few reported any experience of physical pain as compared to the other individual-classes. Similar to individual-class 3, age was related to pleasantness rather than unpleasantness. As compared to other classes, external vectors for life satisfaction and love appeared to be more related to activation than pleasantness.

Country-classes

Using the procedure of model selection outlined earlier, the final model selected was one that consisted of 5 country-classes as seen in the lower half of Table 4. All 116 countries were clustered into these 5 country-classes and a geographical map showing the country-classes is depicted in Figure 5. Interestingly, the major countries that share a common classification appear to have some common historical/cultural roots and share similar geographical regions.

Table 5 shows the distribution of individual-classes across different country-classes. We also identify the countries that were clustered into each of the country-classes.

Country-class 1. A third of the 116 countries fell into this country-class. It consists primarily of Russia, Eastern Europe and some African nations. Notably, many countries within this country class were relatively poorer and had ongoing wars or conflicts. It is not surprising that a large proportion (0.76) of individuals were from individual-class 1, in which there was relatively low endorsements of positive emotions, but higher endorsements of “anger” and “shame”. Also, individual-class 1 had the lowest life satisfaction and experience of love.

*Country-class 2.*This country-class most evidently consisted of Latin America, and the Iberian Peninsula (i.e., Portugal and Spain). It is interesting to note that these countries share common histories and language. A substantial proportion of individuals (0.69) were from individual-class 2. Hence, many individuals in this set of countries endorsed the experience of positive and negative emotions in general, and individuals perceived unpleasantness as strongly related to physical pain.

Country-class 3. Western Europe, North America, Canada, Australia and NZ were notable groups in this country-class. This class of countries is representative of many Western developed countries. This country-class had a large proportion of individuals from individual-class 3. They reported high levels of enjoyment, moderately high stress, but had low depression, anger and shame. Further, individuals from individual-class 3 were generally older and age was moderately related to pleasantness. This may be attributable to higher life expectancy and better health care.

Country-class 4. Seven out of seventeen countries in this class were from Asia, while others were from various regions. Given the varied countries clustered together, it is not surprising that the individual-classes were also distributed between individual classes 4, 3 and 2. This class had the most diverse individuals in it, which might reflect either heterogeneity of regions within nation, or ethnic groups, or might reflect rapid changes in the societies, leading to large differences in norms across individuals.

Country-class 5. Similar to country-class 4, the last country cluster did not appear to come from a particular geographical/cultural region. However, almost all individuals came from individual-class 4. Individual-class 4 consists of individuals that reported the lowest "respect" and "pride" relative to other individual-classes. It is the typical pattern of Asian/Collectivist groups with low pride, and low on sadness and worry.

Summary of Results

Structure of affect. When one examines the experience of individual emotions, the pleasant-unpleasant dimension was clearly revealed. This dimension was strongly and consistently related to the experience of physical pain and to a moderate degree love and life satisfaction. Although the emotion terms sampled did not vary highly on the activation dimension, the external vectors showed that emotional activation was related to the experience of love and higher life satisfaction. In contrast, age was related to increased emotion deactivation. Overall, this revealed that a two dimensional structure of affect was present across all the individual-classes, with positive and negative items consistently lying on the opposite ends of the valence continuum; whereas the activation component was congruent with external criteria like love, life satisfaction and age. Given the wide sampling available, these results confirm past findings that the affect structure across different cultures conforms to two primary dimensions – valence and activation (Russell, 1983, 1991; Russell et al., 1989). Nevertheless, our results also suggest that these affect structures are also distinct, such that the specific locations of the item loadings differs. Further, the underlying dimensions have differential strengths of relations to various external criteria.

Individual-classes. The individual-classes uncovered in the present study indicated that the organization of affect in individuals goes beyond whether they are generally happy or unhappy. What we found is that the four latent classes mixed emotions in a more intricate way than simply the amount of pleasant versus unpleasant emotions a person feels. The first latent class was generally composed of “unhappy individuals”, but there was a notable lack of stress in this group. The third individual-class came closest to what is typically thought of as “happiness,” as they also had the highest life satisfaction, except that this group experienced high levels of stress. The pattern in these two groups indicates that reported feelings of stress are caused by different factors than the other negative emotions. Perhaps a successful but hectic lifestyle in prosperous nations is most associated with reports of stress. The second individual-class might be characterized as emotional, with high levels of both positive and negative affect. Because positive affect predominates for most people, the group can be characterized as on balance as happy, but with a substantial amount of negative feelings. Here the discrepancy of the individual-classes from a simple happy-unhappy continuum becomes quite clear. The fourth individual-class was the most complex in that it experienced substantial enjoyment but low levels of pride and respect, and moderate levels of some types of negative affect but low levels of other types. This is a pattern that is often associated with Asian societies, which suppress individualistic emotions such as pride. Thus, it appears that the latent classes to some degree represent how “happy” or “unhappy” people are in general, which may reflect circumstances in their nations, but also reflects the norms for feelings in the respective societies. All emotions may be amplified in general, or specific emotions may be dampened.

Country-classes. We found that country-classes were generally interpretable and appeared to share certain commonalities (e.g., socio-economic status, historic-cultural roots, or geographic region). Further, country-classes did not consist only of homogenous groups of individuals. Rather, four individual-classes, in which the same measurement model applied, spanned these country classes. This implied that there were qualitative differences in how individuals viewed self-reported emotions. Nevertheless, some country-classes had a substantial proportion (e.g., 0.69, 0.76, 0.77, and 0.94) of individuals from specific individual-classes. While not strictly homogenous, such countries did consist of fairly homogenous individual-classes. The endorsement profiles associated with each individual-class showed the specific types of emotions likely to be experienced. A cursory examination of the dominant individual-class underlying the country-class suggests a logical correspondence. For example, country-classes with lower socio-economic status and more conflicts had relatively lower endorsements of positive emotions; on the other hand, relatively developed countries had more positive emotions, particularly enjoyment.

Discussion

This paper explicated the MMM-IRT model for identifying latent classes at both the individual- and hierarchical-level, stating the statistical and conceptual assumptions. For illustrative purposes, we applied this methodology to self-reported emotions of individuals from 116 countries. Results showed that these individual- and country-classes were interpretable and uniquely clarify how self-reported emotion is structured among groups of individuals and countries across the globe. In this section, we discuss the relevance of MMM-IRT to important organizational issues including cross-cultural, multilevel and measurement equivalence research. Because MMM-IRT also integrates both of person- and variable-centered approaches, there is additional utility beyond either, and we elaborate on these issues. Finally, we present new areas for research in the MMM-IRT methodology.

Cross-cultural research

With the rise of organizational internationalization, organizational scientists are emphasizing the need to examine cross-cultural issues (Adler, 1983; Gelfand, 2000; Gelfand, Raver, & Holcombe Ehrhart, 2002). A significant topic in cross-cultural research is the comparability of constructs (Riordan & Vandenberg, 1994) and the generalizability of theoretical models across cultures (e.g., Wasti, Bergman, Glomb, & Drasgow, 2000). Often, cross-cultural studies have relied on cross-national comparisons, implicitly assuming that country membership segregates cultural subpopulations. However, with globalization and cultural connectivity, it has been argued that assuming culture as “geographically localized” is less tenable (Hermans & Kempen, 1998). And yet, despite waning enthusiasm for the use of country as a proxy to culture (e.g., Matsumoto & Yoo, 2006), countries/regions are theoretically and historically important in shaping cultural identities (Chao & Moon, 2005), and the use of country groupings has some validity.

Another challenge in cross-cultural research is ascertaining the degree to which cultural dimensions are universal or relativistic (see Tay, Woo, Klafehn, & Chiu, in press). Universalistic approaches assume that psychological or cultural dimensions are invariant across countries and only *quantitative* differences occur (e.g., individualism-collectivism; Hofstede, 1984). In contrast, relativistic approaches view cultures as *qualitatively* distinct (e.g., unique personality dimension in Chinese subpopulation; F. M. Cheung et al., 2001). Within the field of organizational research, this tension is manifest in a diverging emphasis on the qualitative assessment of organizational culture (e.g., Schein, 1992) or the use of questionnaire measures for quantitative comparisons (e.g., Cooke & Szumal, 2000; O'Reilly, Chatman, & Caldwell, 1991). Can these two polarized views be resolved by deriving qualitatively distinct interpretations of constructs that are culturally universal (cf. Poortinga & Van Hemert, 2001)? Indeed, the trend toward growing cultural complexity and hybridization may lead to subgroups that share common meanings not necessarily bounded by national or organizational boundaries (Hermans & Kempen, 1998; Miller, 1997).

While not a panacea for all cross-cultural conundrums, MMM-IRT does go beyond traditional ME procedures in addressing these challenges. Foremost, due to its methodological assumptions, ME procedures utilize country groupings as a priori cultural subpopulations. In contrast, the use of MMM-IRT allows researchers to use less restrictive assumptions; all members in a country do not necessarily share the same measurement model. Further, as illustrated by our example, one can derive qualitatively distinct classes of individuals that span countries, revealing universalistic, yet possibly idiosyncratic, frames-of-references on a construct(s). For instance, Table 5 shows that country-class 1 is primarily defined by individual class 1 (76%), revealing an idiosyncratic frame-of-reference for countries in country-class 1, but not necessarily limited to only such countries. This reasoning stems from past research which used multigroup LC analysis[5] on norms for experiencing emotions (Eid & Diener, 2001), which ascertained relatively universal or idiosyncratic latent classes if there were roughly equal proportions or disproportionate numbers of country membership within a LC respectively.

The results from our illustration suggest other issues for consideration as well. First, given that most countries have large proportions of individuals that share the same measurement model, there is some correspondence with dimensional approaches that assume all individuals within the country share the same measurement model. However, MMM-IRT also shows that there are non-negligible proportions of individual-classes that span across countries and country-classes. If there are indeed individuals who share a similar frame-of-reference across countries, would delineating individuals by country membership, as with ME procedures, result in a compartmentalized view of how emotion structures differs? For instance, one may conclude that country A has a different measurement model than country B, but fail to determine if there are groups of individuals who do share the same measurement model across both countries. Also, with a larger number of countries, the probability of non-ME is more likely and ME procedures become less feasible. Yet, discerning common motifs among countries are important. Are there commonalities among countries? Can these countries be

[5] Multigroup LC models can be seen as a form of multilevel LC models. Multigroup LC models assume that hierarchical units (e.g., countries) are fixed effects, whereas multilevel LC models assume that hierarchical units are random effects (see Vermunt, 2003). We note that in instances when there are only a few hierarchical units, we can treat them as fixed effects in the MMM model. In this case, no latent classes for hierarchical units are assumed.

grouped together while taking into account measurement issues? We suggest that MMM-IRT can be used to explore these important theoretical questions.

Multilevel research

Multilevel theory has become a cornerstone for explicating many organizational issues (Klein et al., 1994; Klein & Kozlowski, 2000; Kozlowski & Klein, 2000), ranging from organizational climate (Glick, 1985) to leadership (Yammarino & Bass, 1991). Articulating the correct theoretical levels-of-analysis (Klein et al., 1994), the structure and function of higher-level constructs (Morgeson & Hofmann, 1999) and composition models (Chan, 1998) are fundamental for conceptual clarity and appropriate statistical analysis, all of which are inextricably intertwined and inevitably impinge on research findings and theoretical models. While levels issues have grown in clarity with respect to multilevel regression models, MMM-IRT is an alternative analytic technique that integrates measurement and latent class issues within a single model. As a result, it stands to generate several critical lines of inquiry in multilevel research.

Level-of-analysis. In the framework of multilevel regression models, individual scores are generally taken as direct reflections of one's standing on the individual-level construct; analogously, composite scores (e.g., mean scores within hierarchical units) show the quantitative ordering of hierarchical-units on the higher-level construct. MMM-IRT extends the levels-of-analysis conceptualization. In this procedure, latent constructs are directly estimated at both levels, but these latent constructs reflect qualitative similarity. At the individual-level, the construct reflects homogenous measurement classes. Thus, not all individuals share a similar construct referent by which to compare and manipulate observed scores. However, individuals within a common class can be quantitatively compared. At the higher-level, the construct demarcates the hierarchical-unit similarity.

Structure-and-function of higher-level constructs. While individual- and hierarchical-level constructs are both present in MMM-IRT, the structure and function of these constructs are substantially different. Recent theoretical developments have proposed that structure emerges from joint interactions within the collective (synonymous with the hierarchical unit) and eventually exerts an independent influence on lower-level units (Morgeson & Hofmann, 1999). Concomitantly, in cultural research, it has been suggested that observable externalizations of interactions (Hannerz, 1992) can act on work-related and psychological constructs within the collective (e.g., political and economic infrastructure for a country; or the codification of formal organizational behavior). In either case, it is assumed that higher-level units all reference the same higher-level construct.

Notably, in MMM-IRT, the hierarchical-construct bears some similarity in that it takes into account *nested dependencies* that may arise because of joint interactions or observable externalizations. The first major departure is that the MMM-IRT analysis is founded upon on item-level information, where qualitatively distinct individual-level measurement models are inferred. A second difference is that in MMM-IRT, the hierarchical-construct is posited to be the basis for how hierarchical units are similar or different on this assortment of individual-classes. In this regard, the hierarchical-construct does not represent a “hermeneutically” invariant construct (a hierarchical-construct that is interpreted in the same manner) on which hierarchical units only differ quantitatively. It is a latent classification of hierarchical units. Given the nature of the hierarchical-construct, a subsequent question would be: What processes underlie its emergence? We speculate that the structure may arise because of migratory trends, where information or individuals flow more freely within the latent classification of hierarchical units; such cross-unit interactions give rise to common construct perceptions. For example, our results show that Latin America, Portugal and Spain are similar; which may be a result of actual migration of individuals, leading to common practices and languages. As an analog to observable externalizations, hierarchical-units establish protocols, policies and infrastructure that exert a homogenizing influence. Or, it may be that these hierarchical units share common occurrences. Our illustration shows that country-class 1 consisted primarily of countries with lower wealth and ongoing internal conflicts; whereas country-class 3 consisted mainly of richer countries that have more stability.

The function, or “causal outputs” (Morgeson & Hofmann, 1999, p. 254), of MMM-IRT higher-level constructs, is then the homogenizing influence, as alluded to earlier. We note that this influence should be understood in terms of the measurement classes at hand. For instance, the homogenizing influence among teams may be a consequence of team types, which leads to sweeping similarities in shared perceptions of a team climate measure. Certainly, these theoretical endeavors into structure and function are preliminary and we urge researchers to develop fuller and richer explications of MMM-IRT theoretical models.

Composition models. Closely related to understanding the structure and function of collective constructs, composition models are critical for multilevel research because they explicate the functional relationships among constructs that reference the same content, but are constituted at different levels (e.g., individual, team, organization, country). Chan (1998) developed a typology for different composition models, clarifying the meanings of the higher level construct in each case. These models focus on the meaning of higher level constructs depending on the operations on observed scores (e.g., mean, variance) and type of scores (e.g., an individual’s self-endorsements versus an individual’s normative perceptions) in lower level units. As mentioned earlier, MMM-IRT goes beyond the use of observed scores to the analysis of item-level responses. Some methodological limitations to the use of observed scores have been noted by researchers. First, it confounds true scores and the measurement model (Drasgow, 1987). Second, comparing individuals with a different measurement model is “tantamount to comparing apples to spark plugs” (Vandenberg & Lance, 2000, p. 9) because the construct is qualitatively different from one class of individuals to the next. Instead, delineating individuals who share a common measurement model, and applying the relevant composition models among these individuals may be one way to overcome epistemic differences. Because the measurement model is equivalent in each individual-class, we can obtain the means (or variances) of the observed scores for each individual-class within each hierarchical-class. This procedure stemming from MMM-IRT results may allow for a more sensitive procedure ensuring that composite scores are equivalent.

Measurement Equivalence Research

Issues related to measurement equivalence have been mentioned in both cross-cultural and multilevel research. However, ME is fundamental for accurate substantive interpretations of the cross-group comparisons that most interest organizational scientists (Vandenberg & Lance, 2000). Because MMM-IRT analysis can generate new methodological queries in ME research itself, it warrants fuller development.

The rationale for conducting ME research is to ensure that qualitative differences on the construct(s) of concern do not exist between groups. Organizational and cross-cultural researchers ordinarily wish to establish ME prior to examining group differences. But what if ME is not achieved? Given the difficulties associated with establishing ME, recommendations include relaxing the assumption of full ME (i.e., using partial ME models), deleting non-equivalent items, or giving a post-hoc interpretation of non-equivalent items (G. W. Cheung & Rensvold, 1999; Vandenberg, 2002). Researchers have also used item parceling to meet the ME requirement (see Meade & Kroustalis, 2006). While procedures for dealing with non-ME are useful when attempting to answer specific questions regarding mean differences between manifest groups, we suggest that for other purposes, the MMM-IRT model is a less restrictive framework, and yet provides a more nuanced interpretation of how hierarchical units differ. Specifically, the MMM model assumes that qualitative differences can exist within a manifest group. Establishing that different hierarchical units have different proportions of individual-classes allows one to make comparisons regardless of whether differences are primarily quantitative or qualitative. Using our results in Table 5 as an example, (a) the differences between countries within country-class 5 is primarily quantitative because a majority of individuals have the same measurement model; (b) differences between countries across country-classes are primarily qualitative where there is little overlap in the individual-classes, as in country-class 1 and 5; (c) finally, one can examine quantitative differences for each individual-class across the countries even where countries have different proportions of individual-classes.

Lines of inquiry for ME research. We found that countries have mixtures of individual-classes, with each individual-class defined by a distinct measurement model. Thus, manifest groupings do not partition latent heterogeneity in an exact fashion. Instead, it is possible that such groupings share subpopulations with common measurement models. Although ME research assumes measurement homogeneity within manifest groups, it is not known how and when within-group unobserved heterogeneity can cause problems. Little research has explicitly examined what happens when a manifest group has two or more latent subpopulations with differing measurement models. Are there conditions in which a common measurement model can be applied? For example, the size of the dominant latent subpopulation may be an important factor. Further, to what extent does a shared latent subpopulation across manifest groups impact the detection of non-ME? Recent simulation research has shown that when latent groupings that share distinct measurement models diverge from observed groupings, standard DIF techniques have lower power to detect DIF associated with true differences due to the latent group measurement models (De Ayala, Kim, Stapleton, & Dayton, 2002). More research needs to examine the correspondences and divergences between MMM-IRT and traditional ME applications. For an alternative to the traditional ME procedures, refer to the use of MM-IRT with covariates by Tay, Newman & Vermunt (in press) in this special issue.

Integrating Person-centered and Variable-centered Approaches

Dimensional techniques like factor analysis are variable-centered approaches, where the interest is in accounting for the relationships among the observed variables. Accordingly, latent dimensions are posited to account for the covariation between variables. For example, the Five-Factor model of personality (McCrae & Costa, 1987) is fundamentally based on a variable-centered approach. On the other hand, classification techniques like latent class analysis have been considered as person-centered approaches, where the prime interest is in accounting for relationships among individuals (B. Muthén & L. Muthén, 2000). It is assumed that different types of individuals are responsible for the variability in responses. At this point, a clarification is necessary to avoid any confusion. In our view, there are two forms of person-centered approaches: (a) idiographic approaches (Allport, 1962) are concerned with how variables are organized within the individual and the focus is on intra-individual variability over time; and (b) taxometric approaches where the aim is to cluster individuals who have similar attributes is generally based on cross-sectional data (see Meehl, 1992). In our illustration and in many other applications, because the data are most likely cross-sectional, we view MMM-IRT primarily as a taxometric rather than an idiographic person-centered approach.

In mixed-measurement IRT modeling, which is the underlying basis for MMM-IRT, both taxometric person-centered and variable-centered approaches are integrated (e.g., Lubke & Muthén, 2005). Specifically, at the individual-level, latent groups of individuals are inferred, and within these groups, a latent dimension is posited to account for the relationship among observed variables. The practical utility is that individuals within each individual-class can be compared quantitatively; but qualitative distinctions can also be made about the types of responses made. Qualitative distinctions between individuals are emphasized in wide-ranging organizational issues, these include examining personality clusters within the context of Person-Environment (P-E) fit (De Fruyt, 2002), biographical data profiles to improve predictive validity (Schmitt et al., 2007), or profiles in coping with sexual harassment (Cortina & Wasti, 2005). However, it is not known the degree to which observed scores used to determine clusters and external validities are comparable from a measurement standpoint.

Through the use of mixed-measurement techniques, and MMM-IRT more generally, one can determine the adequacy of a single measurement model to the entire sample. For instance, it has been found that not all individuals perceive and respond to the '?' category on a personality scale in the same way (Hernandez et al., 2004). However, because score comparisons are valid within each responder class, the predictive validity of such classes can still be examined (Maij-de Meij et al., 2008).

MMM-IRT: Areas for Methodological Research

We suggest that there are several important directions for MMM-IRT research. These consist of two general areas: data requirements for analysis and model-data fit strategies. To use MMM-IRT in research, it is first necessary to determine the numbers of items and sample sizes required at both the lower- and higher-levels. For example, recent multilevel latent class simulations have centered on 5 to 30 individual-level units and 30 to 500 hierarchical-level units (Lukociene et al., under review; Lukociene & Vermunt, in press). Because the MMM-IRT model additionally comprises a measurement model for each class, more research should determine if sample sizes commonly encountered in organizational research can be used to fit the model.

Regarding model-data fit strategies, there are several procedures for ascertaining the correct number of individual- and hierarchical-level classes because classes at both levels are dependent. These strategies include (a) the two-step approach used in this paper which follows the multilevel latent class strategy where the best number of individual-classes is first obtained while setting hierarchical-classes to 1. At step 2, the procedure is repeated for hierarchical classes setting the individual-classes to the best number previously obtained (Vermunt, 2003); (b) a more recent three-step approach whereby after the second step, one re-estimates third step in which the number of individual-level classes is re-estimated fixing hierarchical-classes to the value found in step 2 (Lukociene et al., under review); and (c) an exhaustive approach where all combinations of individual-classes and hierarchical-classes are estimated and one selects the model with the lowest information criteria value (Bijmolt, Paas, & Vermunt, 2004). However, such an approach is only practicable when the estimation of each model is fairly quick, and becomes much harder with large sample sizes. Thus, more work needs to be done in this area.

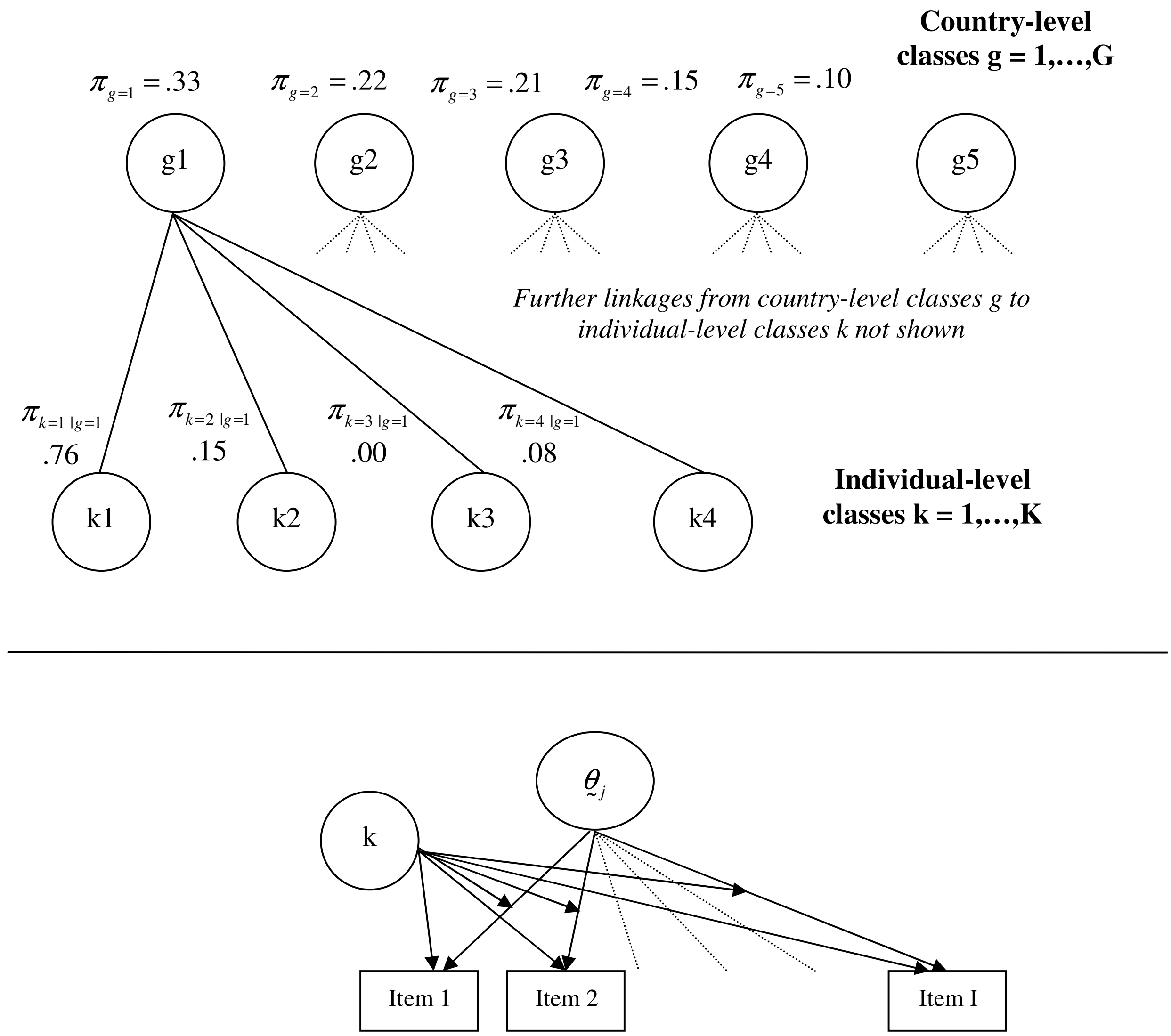
There have been a number of information criteria indices that have been proposed for use in latent class modeling, ranging from the Akaike information criteria (AIC) (Akaike, 1974), BIC, consistent AIC (CAIC) (Bozdogan, 1987), to AIC3 (Bozdogan, 1993). Although simulations have examined nested models of the MMM-IRT (e.g., multilevel latent class models), less is known about the effectiveness of information criteria for identifying the true model. Recent research has shown that for the BIC performs well for larger sample sizes, but the AIC performs better for smaller sample sizes in multilevel latent class models (Lukociene & Vermunt, in press). However, because BIC tends to underestimate the number of classes, and the AIC tends to overestimate it, the AIC3 has been recommended as a compromise. We suggest that more research should be undertaken in this direction for MMM-IRT models.

Summary and conclusion

Although organizational scientists have relied primarily on observed heterogeneity (e.g., membership in hierarchical units) to distinguish subpopulations for separate analyses, we propose that latent heterogeneity exists in our data and may not consistently correspond to these manifest hierarchies. Where the goal is to obtain relatively homogenous subpopulations that share the same measurement model while taking into account nested dependencies, we propose that the MMM-IRT model can be applied. This approach utilizes information from observed hierarchical groups and item responses to infer individual-level measurement classes and hierarchical-classes simultaneously. Such an approach has multiple theoretical and methodological advantages for cross-cultural comparisons, multilevel research, and the study of measurement invariance. In conclusion, we encourage organizational scientists to consider the use of this model for the study of a wide range of substantive and methodological issues.

Figure 1

Graphical presentation of the MMM model: Data structure and item response model



Note. Diagram above shows the model of latent heterogeneity (hierarchical-level and individual-level latent classes) within the data while the diagram below depicts the probability of item response with respect to individual-class and trait standings where the individual-class *k* not only moderates the relationship of the latent trait to the indicators but also affects the indicators directly. We note that the summation of the country-class or individual-class probabilities may not add to 1 due to rounding.

Figure 2

Structure of affect with dimensional representations

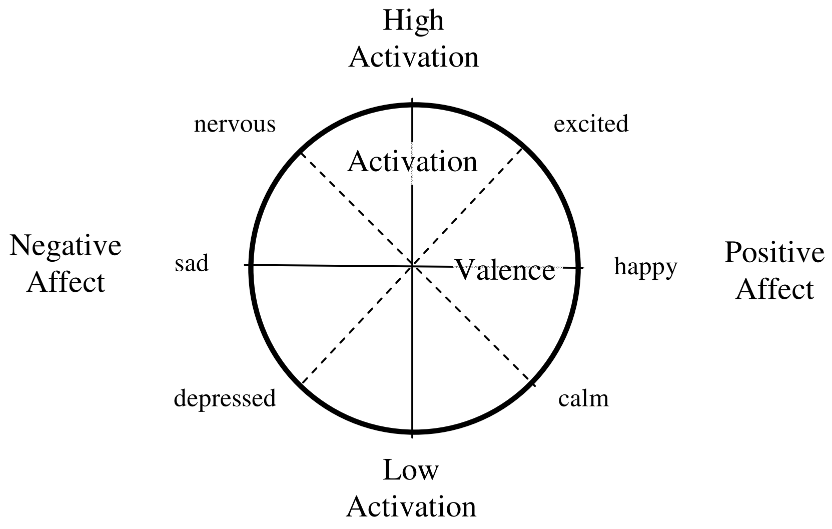


Figure 3

Endorsement profiles by individual-level latent class

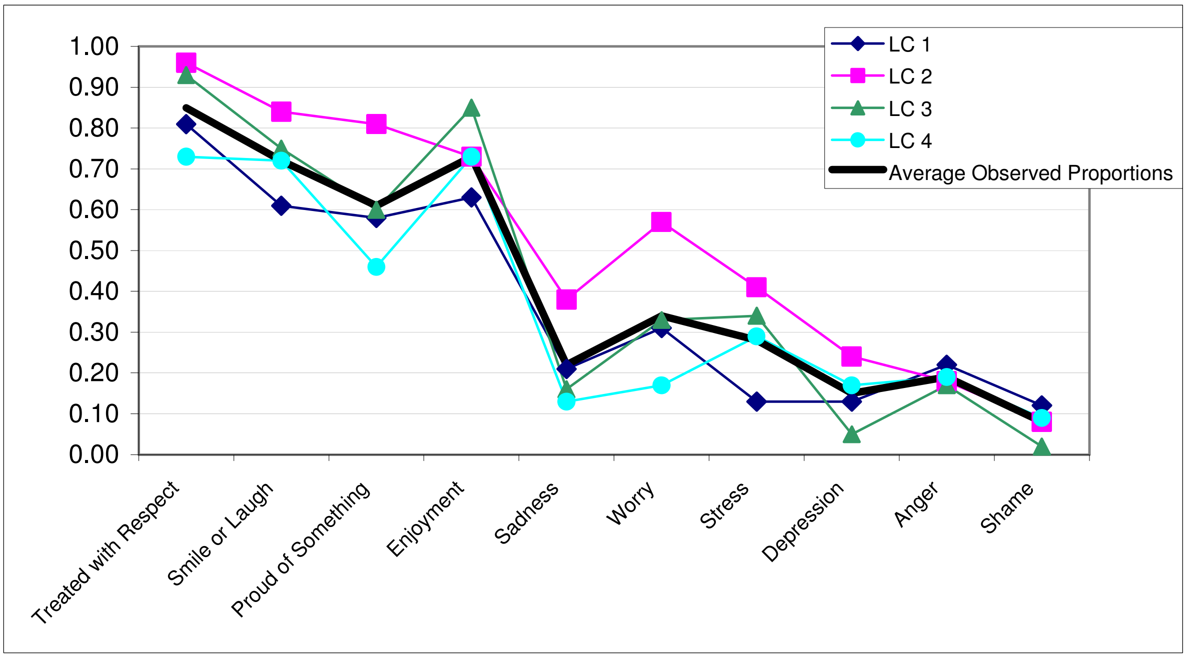


Figure 4

Dimensional representations of self-reported emotions for individual-classes

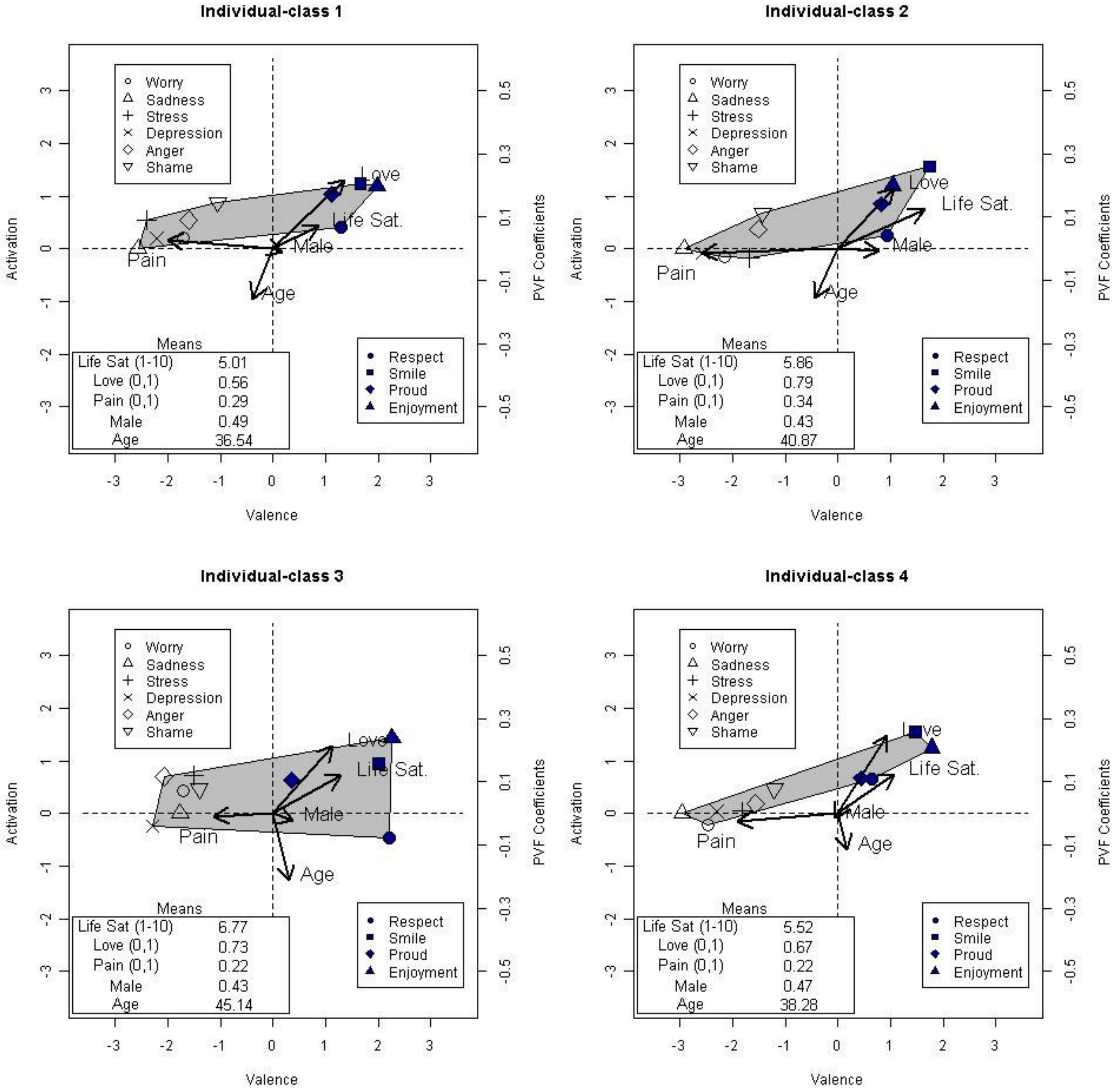
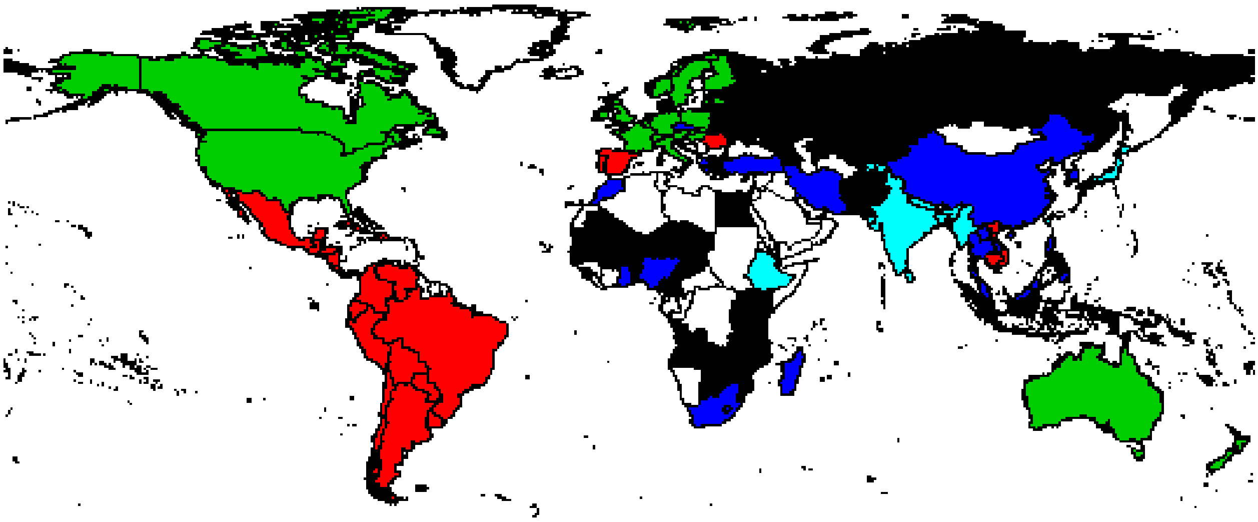


Figure 5

Geographical representation of 5 country-classes for 116 countries



Note. Black=Country-class 1; Red= Country-class 2; Green= Country-class 3; Blue = Country-class 4; Turquoise= Country-class 5

Table 1

MMM-IRT: Applications to, and generative questions for organizational topics

Organizational topic	Substantive Issues	Application of MMM-IRT	Generative Theoretical and Methodological Questions
Cross-cultural research	- With globalization, the country-as-culture paradigm for cross-cultural comparisons is increasingly being questioned. Can we instead obtain measurement classes that span countries? - Are cultural constructs universal or idiosyncratic? - Are there ways to ascertain commonalities among countries aside from quantitative comparisons?	- MMM-IRT takes into account nested dependencies and infers individual-level measurement classes that span across countries. - An examination of the proportions of measurement classes within each country can shed light on the degree of idiosyncrasy and universality for the construct of interest. - Countries are grouped to the extent they share similar measurement class proportions.	-To what degree do individuals from different societies (e.g., East versus West) share common frames-of-reference on cultural constructs (e.g., individualism-collectivism)?
Multilevel research	- *Level-of-analysis.* How can we directly estimate hierarchical constructs? - *Composition models.* Observed score aggregates are commonly used without testing for measurement invariance among hierarchical units because of small numbers of lower-level units.	- MMM-IRT allows a direct statistical inference of hierarchical-level latent classifications. - Given a reasonably sized total sample with sufficient lower-level units within hierarchical units, we can determine if multiple groups share a common measurement basis for comparisons.	- *Structure and function of collective constructs.* Can we develop a fuller explication of the nature of collective constructs produced by MMM-IRT? -*Composition models.* Can MMM-IRT provide a framework by which scores of individuals sharing the same measurement model are aggregated despite being in disparate hierarchical units? What would be the implications for composition models?
Measurement Equivalence (ME)	- There are difficulties in interpreting quantitative differences among groups when non-ME occurs. Can we approach ME by examining commonalities of individuals among hierarchical units? -With large numbers of hierarchical units, conducting multiple-groups measurement equivalence is difficult.	- MMM-IRT model is a less restrictive framework, in that not all individuals within the hierarchical unit necessarily share the same measurement model. - It can provide a more nuanced interpretation of how hierarchical units differ, by examining if there are common classes among hierarchical units. -Because MMM-IRT uses a multilevel framework, one can incorporate many hierarchical units in the examination of common measurement classes.	-To what degree do differences in individual-level class proportions affect measurement invariance? For example, how does the size of measurement class or magnitude of difference between measurement models affect results from ME procedures?

Table 2

Different structural models within the general latent variable framework

Levels-of-analysis	Data Type				
		Non-nested data			
Hierarchical		-			
Individual		No measurement model		Measurement Model	
		Single Class	Multiple Classes	Single Class	Multiple Classes
	Categorical	-	Latent Class Analysis (LCA)	Item Response Theory (IRT)	Mixed-measurement IRT
	Continuous		Latent Profile Analysis (LPA)	Factor Analysis (FA)	Mixed-measurement FA
		Nested data			
Hierarchical		Single Dimension			
Individual		No measurement model		Measurement Model	
		Single Class	Multiple Classes	Single Class	Multiple Classes
	Categorical	-	Multilevel LCA	Multilevel IRT	Multilevel mixed IRT
	Continuous		Multilevel LPA	Multilevel FA	Multilevel mixed FA
Hierarchical		Multiple Latent Classes			
Individual		No measurement model		Measurement Model	
		Single Class	Multiple Classes	Single Class	Multiple Classes
	Categorical	-	Multilevel mixed LCA	Multilevel mixed IRT	Multilevel mixed IRT
	Continuous		Multilevel mixed LPA	Multilevel mixed FA	Multilevel mixed FA

Table 3

Means, standard deviations and percentage missing of positive and negative emotion variables

	Mean	SD	% Missing
Positive Emotions			
Treated With Respect	0.85	0.35	3.73%
Smile or Laugh	0.72	0.45	2.75%
Proud of Something	0.61	0.49	4.13%
Enjoyment	0.73	0.44	1.49%
Negative Emotions			
Sadness	0.22	0.41	1.07%
Worry	0.34	0.47	0.93%
Stress	0.28	0.45	1.22%
Depression	0.15	0.35	1.39%
Anger	0.19	0.39	1.10%
Shame	0.08	0.27	1.29%

Table 4

Results of MMM analysis on self-reported emotions of 116 countries

	Log-Likelihood	BIC	No. of Parameters
1-IClass 1-CClass	-544874.58	1090088.74	29
2-IClass 1-CClass	-543540.66	1087772.19	59
3-IClass 1-CClass	-542813.82	1086669.79	89
4-IClass 1-CClass[a]	-542574.66	1086542.77	119
5-IClass 1-CClass	-542456.97	1086658.67	149
6-IClass 1-CClass	-542396.25	1086888.52	179
7-IClass 1-CClass	-542306.67	1087060.65	209
4-IClass 1-CClass	-542574.66	1086542.77	119
4-IClass 2-CClass	-537220.20	1075880.69	123
4-IClass 3-CClass	-534258.24	1070003.60	127
4-IClass 4-CClass	-533357.92	1068249.80	131
4-IClass 5-CClass[b]	-532274.76	1066130.32	135
4-IClass 6-CClass	-532478.93	1066585.49	139

Note. [a]Low BIC values showed that 4 individual-classes (IClass) parsimoniously partitions the latent heterogeneity at the individual-level well. [b]Low BIC values showed that 5 country-classes (CClass) parsimoniously partitions the latent heterogeneity at the country-level given 4 individual-classes. This was the final MMM model selected.

Table 5

Distributions of individual-classes across country-classes

		Country-class 1	Country-class 2	Country-class 3	Country-class 4	Country-class 5
	County-class Size	0.33	0.22	0.21	0.15	0.10
Individual Class 1	Description Generally low positive affect – this class of individuals report much less smiling/laughter and enjoyment. There is higher anger and shame, moderate worry, sadness and depression, but low stress.	**0.76**	0.11	0.02	0.00	0.01
Individual Class 2	Generally very high positive affect. However, this class of individuals also report very high negative affect as well, especially on worry, sadness, stress and depression.	0.15	**0.69**	0.16	0.24	0.05
Individual Class 3	Moderate levels of positive affect with highest experience of enjoyment. Moderately low negative affect, this class of individuals experiences a notable amount of stress despite having the lowest depression, anger, and shame.	0.00	0.16	**0.77**	0.25	0.00
Individual Class 4	Moderate experience of positive affect; but this class of individuals has the lowest experience of respect and pride. They have moderate negative affect, but report the lowest worry and sadness.	0.08	0.05	0.05	**0.51**	**0.94**
	Countries	Afghanistan Angola Armenia Azerbaijan Belarus Benin Botswana Burkina Faso Burundi Cameroon Chad Egypt Estonia Georgia Haiti Indonesia Kazakhstan Kenya Kyrgyzstan Lithuania Malawi Mali Mauritania Moldova Mozambique	Argentina Bolivia Brazil Cambodia Chile Colombia Costa Rica Cuba Dominican Republic Ecuador El Salvador Guatemala Honduras Mexico Nicaragua Panama Paraguay Peru Portugal Puerto Rico Romania Spain Trinidad & Tobago Uruguay Venezuela	Australia Austria Belgium Canada Cyprus Denmark Finland France Germany Hungary Ireland Israel Italy Latvia Netherlands New Zealand Norway Poland Slovakia Slovenia Sweden Switzerland United Kingdom United States	China (Beijing) Czech Republic Ghana Greece Hong Kong/Macau Iran Jamaica Korea, (South) Laos Madagascar Malaysia Morocco Nigeria Philippines South Africa Thailand Turkey	Bangladesh China (Taiwan) Ethiopia India Japan Myanmar (Burma) Nepal Rwanda Singapore Sri Lanka Tajikistan

Niger	Vietnam
Pakistan	
Russia	
Senegal	
Sierra Leone	
Tanzania	
Togo	
Uganda	
Ukraine	
Uzbekistan	
West Bank & Gaza (Palestine)	
Zambia	
Zimbabwe	

Note. Country names are sorted in alphabetical order. We note that the margins of cell probabilities may not add to 1 due to rounding.

References

Adler, N. J. (1983). Cross-cultural management research: The ostrich and the trend. *Academy of Management Review, 8*, 226-232.

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

Allenby, G. M., & Rossi, P. E. (1999). Marketing models of consumer heterogeneity. *Journal of Econometrics, 89*, 57-78.

Allport, G. W. (1962). The general and the unique in psychological science. *Journal of Personality, 30*, 405-422.

Bijmolt, T. H., Paas, L. J., & Vermunt, J. K. (2004). Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. *International Journal of Research in Marketing, 21*, 323-340.

Bozdogan, H. (1987). Model selection for Akaike's information criteria. *Psychometrika, 53*, 345-370.

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.

Cantril, H. (1965). *The pattern of human concerns*. New Brunswick, NJ: Rutgers University Press.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 1998*, 234-246.

Chao, G. T., & Moon, H. (2005). The cultural mosaic: A metatheory for understanding the complexity of culture. *Journal of Applied Psychology, 90*, 1128-1140.

Cheung, F. M., Leung, K., Zhang, J. X., Sun, H. F., Gan, Y. Q., Song, W. Z., et al. (2001). Indigenous Chinese personality constructs: Is the Five Factor Model complete? *Journal of Cross-Cultural Psychology, 32*, 407-433.

Cheung, G. W., & Rensvold, R. W. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management, 25*, 1-27.

Cho, S.-J., & Cohen, A. S. (in press). A multilevel mixture IRT model with applications to DIF. *Journal of Educational and Behavioral Statistics*.

Church, A. T., Katigbak, M. S., Reyes, J. A. S., & Jensen, S. M. (1999). The structure of affect in a non-western culture: Evidence for cross-cultural comparability. *Journal of Personality, 67*, 505-534.

Cooke, R. A., & Szumal, J. L. (2000). Using the Organizational Culture Inventory to understand the operating cultures of organizations. In N. M. Ashkanasy, C. P. M. Wilderom & M. F. Peterson (Eds.), *Handbook of organizational culture and climate* (pp. 147-162). Thousand Oaks, CA: Sage.

Cortina, L. M., & Wasti, S. A. (2005). Profiles in coping: Responses to sexual harassment across persons, organizations, and cultures. *Journal of Applied Psychology, 90*, 182-192.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

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.

De Fruyt, F. (2002). A person-centered approach to P-E fit questions using a multiple-trait model. *Journal of Vocational Behavior, 60*, 73-90.

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.

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.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269-286.

The Gallup Organization. (2009, August 28). *Gallup world poll: Methodology*. Retrieved August 28, 2009, from http://www.gallup.com/consulting/worldpoll/108079/Methodological-Design.aspx

Gelfand, M. J. (2000). Cross-cultural industrial and organisational psychology: Introduction to the special issue. *Applied Psychology: An international review, 49*, 29-31.

Gelfand, M. J., Raver, J. L., & Holcombe Ehrhart, K. (2002). Methodological issues in cross-cultural organizational research. In S. Rogelberg (Ed.), *Handbook of industrial and organizational psychology research methods* (pp. 216-246). New York: Blackwell.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10*, 601-616.

Green, D. P., Goldman, S. L., & Salovey, P. (1993). Measurement error masks bipolarity in affect ratings. *Journal of Personality and Social Psychology, 64*, 1029-1041.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis*. Cambridge: University Press.

Hannerz, U. (1992). *Cultural complexity: Studies in the social organization of meaning*. New York: Columbia University Press.

Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface- and deep-level diversity on work group cohesion. *Academy of Management Journal, 41*, 96-107.

Hermans, H. J. M., & Kempen, H. J. G. (1998). Moving cultures: The perilous problems of cultural dichotomies in a globalizing society. *American Psychologist, 1998*, 1111-1120.

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.

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

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Levels issues in theory development, data collection, and analysis. *Academy of Management Review, 19*, 195-229.

Klein, K. J., & Kozlowski, S. W. J. (2000). From micro to meso: Critical steps in conceptualizing and conducting multilevel research. *Organizational Research Methods, 3*, 211-236.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco, CA: Jossey-Bass.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Larsen, R. J., & Diener, E. (1985). A multitrait-multimethod examination of affect structure: Hedonic level and emotional intensity. *Personality and Individual Differences, 6*, 631-636.

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

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

Lukociene, O., Varriale, R., & Vermunt, J. K. (under review). The simultaneous decision about the number of lower- and higher-level classes in multilevel latent class analysis.

Lukociene, O., & Vermunt, J. K. (in press). Determining the number of components in mixture models for hierarchical data. In A. Fink, L. Berthold, W. Seidel & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence*. Springer: Berlin-Heidelberg.

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.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*, 224-253.

Matsumoto, D., & Yoo, S. H. (2006). Toward a new generation of cross-cultural research. *Perspectives on psychological science, 1*(3), 234-250.

McCrae, R. R., & Costa, J. P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*, 369-403.

Meehl, P. E. (1992). Facctors and taxa, traits and types, differences of degree and differences in kind. *Journal of Personality, 60*, 117-174.

Mesquita, B., & Walker, R. (2003). Cultural differences in emotions: A context for interpreting emotional experiences. *Behavioral Research and Therapy, 41*, 777-793.

Miller, J. G. (1997). Theoretical issues in cultural psychology. In J. W. Berry, Y. H. Poortinga & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, pp. 85-128). Boston: Allyn & Bacon.

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Journal, 24*, 249-265.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analysis: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental research, 2000*, 882-891.

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

Muthén, L. K., & Muthén, B. (1998-2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Muthén & Muthén.

Newman, D. A., Hanges, P. J., Duan, L., & Ramesh, A. (2008). A network model of organizational climate: Friendship clusters, subgroup agreement, and climate schemas.

In D. B. Smith (Ed.), *The people make the place: A festschrift for Benjamin Schneider* (pp. 101-126). New York: Erlbaum.

O'Reilly, C. A., III, Chatman, J., & Caldwell, D. (1991). People and organizational culture: A profile comparison approach to assessing person-organization fit. *Academy of Management Journal, 34*, 487-516.

Poortinga, Y. H., & Van Hemert, D. A. (2001). Personality and culture: Demarcating between the common and the unique. *Journal of Personality, 69*(6), 1033-1060.

Rabe-Hesketh, S., Pickles, A., & Skrondal, A. (2001). GLLAMM: A general class of multilevel models and a Stata program. *Multilevel modelling newsletter, 13*, 17-23.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

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.

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. San Francisco: Jossey-Bass.

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.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161-1178.

Russell, J. A. (1983). Pancultural aspects of the human conceptual organization of emotions. *Journal of Personality and Social Psychology, 45*, 1281-1288.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological Bulletin, 110*, 426-450.

Russell, J. A., & Feldman Barrett, L. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology, 76*, 805-819.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology, 1989*, 848-856.

Schein, E. H. (1992). *Organizational culture and leadership: A dynamic view*. San Francisco: Jossey-Bass.

Schmitt, N., Oswald, F. L., Kim, B. H., Imus, A., Merritt, S., Friede, A., et al. (2007). The use of background and ability profiles to predict college student outcomes. *Journal of Applied Psychology, 92*, 165-179.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. London: Chapman & Hall.

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.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Steenkamp, J.-B. E. M., & Ter Hofstede, F. (2002). International market segmentation: Issues and perspectives. *International Journal of Research in Marketing, 19*, 185-213.

Tay, L., Newman, D. A., & Vermunt, J. K. (in press). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods.*

Tay, L., Woo, S. E., Klafehn, J., & Chiu, C.-Y. (in press). Conceptualizing and measuring culture: Problems and solutions. In E. Tucker, M. Viswanathan & G. Walford (Eds.), *The handbook of measurement: How social scientists generate, modify, and validate indicators and scales*: Sage.

Thayer, R. E. (1978). Toward a psychological theory of multidimensional activation (arousal). *Motivation and Emotion, 2*, 1-34.

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. (2003). Multilevel latent class models. *Sociological methodology, 33*, 213-239.

Vermunt, J. K. (2008a). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research, 17*, 33-51.

Vermunt, J. K. (2008b). Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics, 37*, 285-299.

Vermunt, J. K., & Magidson, J. (2000a). *Latent GOLD 4.0* [computer program]. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2000b). *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). *LG-Syntax user's guide: Manual for Latent GOLD 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.

Wasti, S. A., Bergman, M. E., Glomb, T. M., & Drasgow, F. (2000). Test of the cross-cultural generalizability of a model of sexual harassment. *Journal of Applied Psychology, 85*, 766-778.

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219-235.

Wedel, M. (2002). Introduction to the special issue on market segmentation. *International Journal of Research in Marketing, 19*, 181-183.

Wedel, M., & Kamakura, W. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Dordrecht: Kluwer.

Wedel, M., Kamakura, W., Arora, N., Bemmaor, A., Chiang, J., Elrod, T., et al. (1999). Discrete and continuous representations of unobserved heterogeneity in choice modeling. *Marketing Letters, 10*, 219-232.

Weiss, H. M., & Cropanzano, R. (1996). Affective events theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. *Research in Organizational Behavior, 19*, 1-74.

Yammarino, F. J., & Bass, B. M. (1991). Person and situation views of leadership: A multiple levels of analysis approach. *Leadership Quarterly, 2*, 121-139.

Yik, M. S. M., & Russell, J. A. (2003). Chinese affect circumplex: I. Structure of recalled momentary affect. *Asian Journal of Social Psychology, 2003*, 185-200.

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

The MMM-IRT is parameterized as follows. The probability $P(\underset{\sim}{y}_c)$, of observing a set of responses in hierarchical unit *c*, in this case countries *c,* is shown by

$$P(\underset{\sim}{y}_c) = \sum_{g=1}^{G} \pi_g P(\underset{\sim}{y}_c \mid g),$$

where *g*, *g*=1,…,*G*, is the group-level (i.e., country) latent class. Similar to the standard LC model, the probability of a hierarchical unit (i.e., country) belonging to latent class *g* is denoted as π_g. The conditional probability $P(\underset{\sim}{y}_c \mid g)$ of observing a set of responses n_c within each country can be written as,

$$P(\underset{\sim}{y}_c \mid g) = \prod_{j=1}^{n_c} P(y_{cj} \mid g),$$

where the probability of observing an individual's response vector in country *c* is statistically independent of other individuals' responses given country-class *g*. The conditional probability can be expanded, and written as

$$P(y_{cj} \mid g) = \sum_{k=1}^{K} \pi(k \mid g) \int P(y_{cj} \mid \underset{\sim}{\theta}_j, k) f(\underset{\sim}{\theta}_j) d\underset{\sim}{\theta}_j .$$

We see here that the individual-class probability $\pi(k \mid g)$ is contingent on country-class *g*. It is important to note, however, that the item response likelihood

$$P(y_{cj} \mid \underset{\sim}{\theta}_j, k) = \prod_{i=1}^{I} P(y_{cji} \mid \underset{\sim}{\theta}_j, k),$$

of an individual *j*, *j*=1,… n_c in country c to a set of items *I* is dependent only on individual-class k and individual standing on the traits $\underset{\sim}{\theta}_j$; it is not a function of country-class g. This parameterization implies that each individual-class *k* is uniquely defined by its own measurement model $P(y_{cj} \mid \underset{\sim}{\theta}_j, k)$, and is invariant across countries and country-classes.

Further, local independence is assumed here: within each individual-class and given $\underset{\sim}{\theta}_j$,

responses are statistically independent.