Exploratory Measurement Invariance: A New Method Based on Item Response Theory

Adam W. Meade, J. Kemp Ellington, & S. Bartholomew Craig
North Carolina State University

Methods of assessing measurement invariance can be time consuming and laborious. In this study, we show that a large sample typical of organizational research can be divided into sub-samples across which measurement invariance is not present. However, attempts to use this sub-group membership to predict the results of measurement invariance analyses across demographic groups did not prove useful. Implications for the broader conceptualization of measurement invariance are discussed.

Test items are said to be free of differential functioning (i.e. are invariant) if they are related to the constructs they measure in the same way across multiple conditions or groups of people. A great deal of recent research has attempted to identify group boundaries across which instruments do not function equivalently (Vandenberg, 2002). However, progress in this area has been slow as comparisons of differential functioning are typically quite laborious, usually involving comparisons between only two groups or conditions at a time.

Because it is difficult or impossible to anticipate whether a specific group boundary will be a source of differential functioning, researchers have had little choice but to test for differential functioning across all group differences. Previous research has examined differences by sex (Powers & Jones, 1984), race (Liang, Lawrence, & Bollen, 1987; Powers & Jones, 1984), time (Vandenbega & Self, 1993), reporting relationship (Factea & Craig, 2001), culture (Ghorpade, Hattrup, & Lackritz, 1999; Ployhart, Wiechmann, Schmitt, Sacco, & Rogg, 2002; Riordan & Vandenberg, 1994), administration method (Meade, Lautenschlager, Michels, & William, 2003), and organizational level (Craig & Kaiser, 2003; Maurer, Raju, & Collins, 1998). In large datasets with multiple demographic variables, the number of possible pair-wise comparisons can make the assessment of differential functioning extremely unwieldy. To improve efficiency, we propose an exploratory method for identifying likely sources of differential functioning in large data sets.

**Appropriateness of Item Responses**

Our exploratory method uses the idea of “appropriateness” to identify likely sources of differential functioning in large datasets. Appropriateness refers to the consistency of individuals' patterns of responses across items (Drasgow, 1982; Drasgow, Levine, & McLaughlin, 1987; Meijer & Sijtsma, 1995, 2001). For example, if a test-taker incorrectly answers a simple arithmetic question but correctly answers a difficult calculus problem, his or her response pattern would be inconsistent. In the past, appropriateness indices have been used primarily to detect faking (Ferrando & Chico, 2001; Zickar & Drasgow, 1996; Zickar & Robie, 1999) and errors in a testing context (Drasgow, Levine, & McLaughlin, 1991; Parsons, 1983). However, appropriateness indices more generally indicate the extent to which the item response model fits the data for each person.

Several reasons why a person's responses may have poor fit to the model estimated for a general sample have been discussed in detail elsewhere (see Drasgow & Guertler, 1987; Meijer, 1998; Reise & Flannery, 1996). However one possible cause of poor person-model fit that has received little prior attention is the notion that some persons may have a different conceptualization of the construct measured by the items than the sample as a whole.
whole (Reise & Flannery, 1996; Tellegen, 1988). For such persons, the item parameters and possibly the item response model as a whole, are not the same as they are for others with a more main-stream conceptualization of the construct being measured. Tellegen (1998) has termed this concept as “traitedness” – the degree of relevance a particular trait construct has to the description of an individual. The notion of traitedness grew out of personality research where some theorists realized that a given personality trait may not be equally applicable for understanding behavior across individuals (Baumeister & Tice, 1988; Bem & Allen, 1974). For example, imagine a person responding to a multi-factor personality scale. Some respondents may not perceive items measuring one of the traits in the same way as other respondents. If this person has low traitedness for this trait, then he or she may not perceive these items to be related to one another. For such a person, the behaviors that are measured by the items may not correlate with one another. Thus, for such a respondent, the items measuring this trait would not relate to one another in a meaningful way and a summary scale score would be meaningless for this individual.

Using this conceptual framework, Reise and Waller (1993) successfully used the most common appropriateness index ($Z_{app}$; Drasgow et al., 1985) to determine a sub-sample for which a measurement model did not apply. After examining the usefulness of appropriateness indices (which they referred to as an index of “scalability”), they argued that, by definition, persons with low scalability could not be compared to other persons in the sample for the trait in question. Though they did not use terms familiar in the measurement invariance literature it is clear that in their study a comparison of the psychometric properties of the items measuring the trait for persons with low scalability and the rest of the population would suffer from a lack of measurement invariance.

Extending this research, we also see the utility of appropriateness indices for locating a subgroup of people who view the survey differently than most others in a sample. Specifically, we believe that persons with aberrant response patterns (i.e. low appropriateness index values) will have different conceptualizations of the construct and thus yield different item parameter estimates than persons with appropriate responses. Thus, we believe that a sample of respondents can be split into two subgroups: (1) persons with responses that fit the measurement model of the trait being measured, and (2) persons with responses that can be considered aberrant from the rest of the sample. In addition, we believe that traditional tests of measurement invariance would reveal that the measure in question would not behave equivalently for these two subgroups.

However, we believe that classifying persons as members of this subgroup of aberrant responders can provide further utility. Specifically, by examining the relations between respondents’ aberrant-response group membership and their membership in various demographic groups, we expect to be able to identify which group boundaries are likely to be sources of differential functioning for a measure. For example, if men were classified as aberrant responders at a much higher rate than women, it is possible that measurement invariance comparisons between men and women would reveal that a measure does not function equivalently for the two samples. Conversely, if groups composed of different races had similar numbers of persons whose responses were classified as aberrant, then it is likely that measurement invariance would hold across those racial groups. In essence, we propose that response classification (aberrant or normal) can serve as a more computationally simple way to test for a possible lack of measurement invariance across many demographic groups as compared to more traditional pair-wise tests of measurement invariance.

### Method

**Participants**

In order to illustrate the proposed methodology, we obtained a large sample of 5,753 respondents to Prospector®, a leadership performance assessment questionnaire (Center for Creative Leadership, 1994). Data were collected as part of a leadership development process. The leadership development program required participants to obtain 360° feedback from qualified raters. The sample used in this study included peers, supervisors, subordinates, and others involved in the assessment process. Respondents were classified into 3631 males, 2122 females, while 4533 respondents listed their race as White, 357 as African American, 448 as Asian, and 182 as Hispanic.

**Measures**

Prospector® contains 47 items designed to measure eleven facets of leadership behavior in two broad clusters: “engages in opportunities to learn” and “creates a context for learning”. Raters are asked to respond to items describing focal leaders’ behavior on a five point scale ranging from “strongly agree” to “strongly disagree.” Though Prospector is a well-validated instrument, the item response theory (IRT) methods used in this study assume strict

---

1 specific information about Prospector is available via the Internet at http://www.ccl.org
unidimensionality in order to obtain unbiased item and person parameters estimates. Therefore we conducted an exploratory factor analysis of the entire data sample in order to identify items that formed a single unidimensional factor. We retained items that loaded highly on the first factor (e.g. factor loadings > .45). Examining the content of these items indicated that they appeared to measure the ability to learn or grow from experience, though there was some slight variation in item content. With so many items, it is not uncommon to see many load onto the first factor, however we chose to retain only 21 items that conceptually fit with the content of the first factor.

*Appropriateness Index Computation*

We computed $Z_l$ (Drasgow et al., 1985) statistics using a SAS program developed specifically for this study. We used MULTILOG 7.03 (Thissen, 1991) to compute item and theta estimates needed to compute the $Z_l$ index. The $Z_l$ statistic is distributed as roughly a $Z$ distribution and, as per previous applications and recommendations (Drasgow & Levine, 1986), persons’ responses with $Z_l$ scores more than two standard deviations below the mean were classified as aberrant. Other response patterns were considered normal.

*Differential Functioning between Normal and Aberrant Subgroups*

Before this new methodology could be used to determine whether a measure would show invariance across demographic groups based on their aberrant/normal subgroup classification, it was first important to show that measures were not invariant for the aberrant and normal subgroups.

Items were tested for differential functioning (DIF) using the IRTLRDIF program (Thissen, 2001) comparing persons with normal response patterns ($Z_l$ scores $> -2.0$; Drasgow et al., 1985) and one with aberrant response patterns ($Z_l$ scores $<-2.0$). *Comparing Appropriateness Indices Across Groups*

Chi-square tests were conducted on demographic group membership by aberrant/normal subgroup membership for each race and the sex of the raters. Significant chi-square values indicate that different demographic groups (e.g. males/females) are significantly more likely to be classified as either normal or aberrant responders. However, in order to determine the efficacy of the proposed methods, it was important to show two things: (1) the measure in question shows DIF for demographic groups differentially associated with aberrant/normal subgroup membership; and (2) the measure in question does not show DIF for demographic groups not differentially associated with aberrant/normal subgroup membership. We again used the IRTLRDIF program to make these comparisons.

**Results & Discussion**

Results from the tests of the normal and aberrant respondents revealed that nearly every item exhibited DIF. Upon closer inspection, it appears that the $b$ parameters differed for all 21 items while the $a$ parameters differed for only two of the items.

Chi-square values associated with aberrant/normal group membership were significant for the race variable, but not the sex variable. Follow-up specific two-group comparisons by race resulted in a significant $\chi^2$ value for Asian-White and Hispanic-White comparisons, and were marginally significant for Black-Hispanic comparisons. Thus, the measure was expected to show DIF across these groups, but not across other comparisons (e.g. White-Black, etc.).

*DIF Analyses by Demographic Groups*

Contrary to our expectations, the number of items showing DIF in comparisons of demographic groups seemed to have little relation to the prevalence of aberrant/normal group membership. Across all comparisons, the number of DIF items ranged from 1 to 8. However, the number of DIF items found for Asian-White (5), Hispanic-White (3) and Black-Hispanic (1) comparisons were typical of those of other comparisons where DIF was not expected.

**Discussion**

In this study, we identified two important findings. First, for a sample not atypical of those commonly encountered in organizational research, a sub-sample was identified for which a measure shows very poor measurement invariance properties as compared to the rest of the sample. Second, membership in this subgroup does not seem particularly useful in predicting the fault-lines by which more traditional measurement invariance tests are likely to encounter DIF.

The most likely explanation for our finding that aberrant/normal subgroup membership gives little indication as to possible groups that might show a lack of measurement invariance lies with the properties of the appropriateness index used to classify respondents. Specifically, appropriateness measurement indices can flag a person’s responses as aberrant for a number of reasons. Some persons may perceive items as far more “difficult” or “easy” than others. For this data, responses would be classified as aberrant both if persons readily agreed to items that others tended to not agree with, or if persons disagreed with items with which others typically
agreed. Additionally, persons responses would likewise be flagged as aberrant if their responses tended to not to be associated with the latent trait at all. As a result, there can be multiple causes of poor person-model fit.

The implications of multiple causes of poor model fit are readily apparent. Specifically, responses might be deemed aberrant due to a host of different response tendencies. As a result, the sub-sample of aberrant responders could have high variability in response patterns. This variability reduces the efficacy of using aberrant/normal subgroup membership as an indicator of likely DIF between demographic groups. For example, suppose that males were more likely to show aberrant response patterns than were females. It is possible that those males with aberrant responses could be evenly split between persons tending to not agree with items (e.g. response 1) on which others chose middle response options (e.g. response 3) and those that tended to agree (e.g. response 5) with items with which others chose middle response options. When item parameters are then estimated for the male sample as a whole, the effects of these lenient and severe response tendencies would cancel out. Thus the male and female samples may show measurement invariance.

**Implications**

Tests of measurement invariance have become a bit of a hot topic in recent years. This is largely due to the severe implications of attempting to compare groups for which a measure is not invariant. However, we suggest that it is time to move past routine tests of measurement invariance. In this study we have shown that for a large sample, there exists a sub-sample for which measurement invariance does not hold (as compared to the remainder of the sample). More importantly, we believe that our sample was typical of most surveys used in organizational research. In other words, we believe that for almost any sample of respondents, there will exist some subset for which a measure does not function in the intended way. As a result, scores for these persons are not directly comparable to those of the remainder of the sample.

If any given sample can be divided into groups for which a measure performs as desired and a smaller subgroup for which a measure does not behave as intended, then the interpretation of the sample statistics as a whole are not reflective of the entire sample. In this case, comparisons of individual persons within the same sample may or may not be valid. Typical measurement invariance tests are conducted across groups (e.g. male vs. female or white vs. African-American) based on the psychometric properties of the samples as a whole. However the comparability of scores within a given group (e.g. males) would seem to be more important than comparability of scores between groups (e.g. males and females) as a whole.

Currently the emphasis in measurement invariance research is very much on comparisons across samples that may or may not be of particular interest (e.g., comparing genders, races, persons of different reporting relationships, etc.) to ensure that the psychometric properties of measure are equivalent across these groups. However, we believe that these comparisons are not questions of psychometric importance for most samples, but instead are sociopolitical questions. Questions of equal item functioning across genders or races may be of sociopolitical interest and importance in ensuring fairness in testing practices. However, a more fundamental psychometric question is whether the modeled relationship between item responses and the latent trait is tenable for any given individual.

As such, we advocate a paradigm shift in measurement invariance research. Instead of comparing the functioning of items across groups of persons based on some demographic grouping variable (e.g. races), measurement invariance tests should begin at the individual level. In other words, before the sample properties of a measure are compared across samples, it should first be shown to function well within a sample.

While the importance of isolating cases for which the measure does not function appropriately awaits further research, it is not difficult to speculate on some of the potential ramifications of such research. For instance, it is plausible that response pattern aberrancy could serve as a moderator of predictor-criterion relationships. If there are persons in the sample for whom a measure does not function properly, it is unlikely that the observed scores for these persons would be as predictive of a criterion (Reise & Waller, 1993). Moreover, if a measure is used in a selection testing context, it would be important to show that scores are directly comparable (i.e., not aberrant) for individuals directly affected by hiring decisions. Thus, if one employee is to be selected from a larger pool of applicants, test scores for those persons in direct competition should be shown to function equivalently in order to ensure testing fairness.

**References**


Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search
for cross-situations consistencies in behavior. Psychological Review, 81, 506-520.


Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.


Author contact info:

**Adam W. Meade**  
Assistant Professor  
Department of Psychology  
North Carolina State University  
Campus Box 7801  
Raleigh, NC 27695-7801  
Phone: 919-513-4857  
Fax: 919-515-1716  
E-mail: adam_meade@ncsu.edu  
[http://www4.ncsu.edu/~awmeade](http://www4.ncsu.edu/~awmeade)

**J. Kemp Ellington**  
Doctoral Student  
Department of Psychology  
North Carolina State University  
Campus Box 7801  
Raleigh, NC 27695-7801  
Phone: 919-515-2251  
Fax: 919-515-1716  
E-mail: kemp@nc.rr.com

**S. Bartholomew Craig**  
Assistant Professor  
Department of Psychology  
North Carolina State University  
Campus Box 7801  
Raleigh, NC 27695-7801  
Phone: 919-513-0518  
Fax: 919-515-1716  
E-mail: bart_craig@ncsu.edu