Contents lists available at ScienceDirect

Human Resource Management Review

journal homepage: www.elsevier.com/locate/humres

Measurement invariance: Review of practice and implications ☆

Neal Schmitt *, Goran Kuljanin

Department of Psychology, Michigan State University, E. Lansing, MI 48824-1116, United States

ARTICLE INFO

ABSTRACT

A review of efforts to assess the invariance of measurement instruments across different respondent groups using confirmatory factor analysis (CFA) is provided for the years since the Vandenberg and Lance [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, 4–69.] review. Investigators are more frequently reporting tests of scalar invariance and tests for differences in latent factor means and partial invariance. Efforts have been made to assess, the impact of the choice of a referent indicator in multi-group studies, the appropriateness of forming partials as indicators of a latent construct, the degree of convergence of item response theory and CFA analyses of measurement differences across groups, and the implications of findings of invariance. In this context, a demonstration of the impact of partial invariance on estimated group differences in reliability and means is provided and discussed.

Increased globalization of many business enterprises as well as immigration has generated the need to use research instruments with individuals in different cultures. Occasionally this also necessitates translation of the instrument to a different language. What one hopes in these instances is that individuals who have the same observed score on these instruments have the same standing on the construct underlying the measurement device. Cultural and language differences, as well as other differences, in the populations being measured necessitate an examination of the degree to which the instrument measures the same construct across these cultural and language groups. Unless measurement invariance is established, conducting cross-group comparisons of mean differences or other structural parameters is meaningless. The degree to which instruments are invariant across use in different situations and with different groups of people has been greatly facilitated by the development of several analytic techniques including item response theory and confirmatory factor analysis (CFA) (Stark, Chernyshenko, & Drasgow, 2006). In this paper, our focus is the use of CFA models in the study of measurement invariance. Researchers have used this strategy to analyze models that include only consideration of covariances, but the strategy can be, and is, extended to include covariance and mean structures in many of the papers in our review.

Vandenberg and Lance (2000) provided an important review of research that addressed measurement invariance. In their review, they pointed to a number of inconsistencies in the manner in which invariance had been assessed. Thus, the purpose of this review is to examine the use of CFA in assessing measurement invariance since the Vandenberg and Lance (2000) review to determine if researchers' practices have changed. We also review the literature that addresses problems/questions with the method itself. Finally, we present the example of an analysis that examines the impact of lack of invariance in one part of the model

☆ The authors would like to acknowledge the helpful comments of Robert Vandenberg and Eugene Stone-Romero on an earlier version of this paper.

* Corresponding author. Tel.: +1 517 355 9563.
E-mail address: schmitt@msu.edu (N. Schmitt).

1053-4822/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.hrmr.2008.03.003

on estimates of other model parameters. The CFA model of two latent factors that are measured with three indicators each is represented in Fig. 1. In the model depicted in this figure, each indicator is represented by the following regression equation:

$$I = c + fF + u \tag{1}$$

where I represents an observed measure or indicator, c is a constant or the intercept in the equation, F is a latent or unobserved variable, f is a regression coefficient or factor loading in CFA terms, and u is the residual or uniqueness associated with the regression.

1. Invariance defined

A measure is invariant when members of different populations who have the same standing on the construct being measured receive the same observed score on the test. A test violates invariance when two individuals from different populations who are identical on the construct score differently on it. In the CFA model, a series of tests are used to establish that there is invariance across populations. The sequence is outlined in Vandenberg and Lance (2000) and a variety of other sources. Byrne and Stewart (2006) provide a didactic treatment of tests for invariance that includes consideration of a hierarchical factor model. The usual steps in these analyses are as follows though not all need be included depending on the purpose of the research:

1. The first step is a test of the difference of the variance–covariance matrices relating items in the measure across the groups being compared. If this test indicates a lack of difference, the usual conclusion is that measures are invariant and no further tests are employed.
2. The second step, called a test for configural invariance, requires a demonstration that the same factors and pattern of factor loadings explains the variance–covariance matrices associated with the groups' responses. This means that the factor structure implied in Fig. 1 is the same for two or more groups of respondents. The values of the parameters in this model may vary, however. Separate analyses of subgroup variance–covariance matrices are conducted to determine if a common model can reasonably represent both matrices or a factor pattern based on previous research is used as a baseline model. This model and the parameters estimated in the model are used as the baseline against which other more restrictive models of the data are compared.
3. Called a test of metric equivalence, the next step involves a test that the values of the factor loadings of each variable on each factor are the same across groups. In terms of Fig. 1, values of the factor loadings (f_1 to f_6) are constrained to equality for each of the groups being compared. Demonstration of the equivalence of factor loadings was labeled “strong invariance” whereas configural invariance is sometimes called weak invariance (Horn & McArdle, 1992) though the terms “strong” and “weak” have often been used to refer to other forms of invariance as well (e.g., Meredith, 1993). With the establishment of metric invariance, researchers interested only in construct validity issues or the interrelationships between latent factors may proceed to test the invariance of factor variances and covariances.

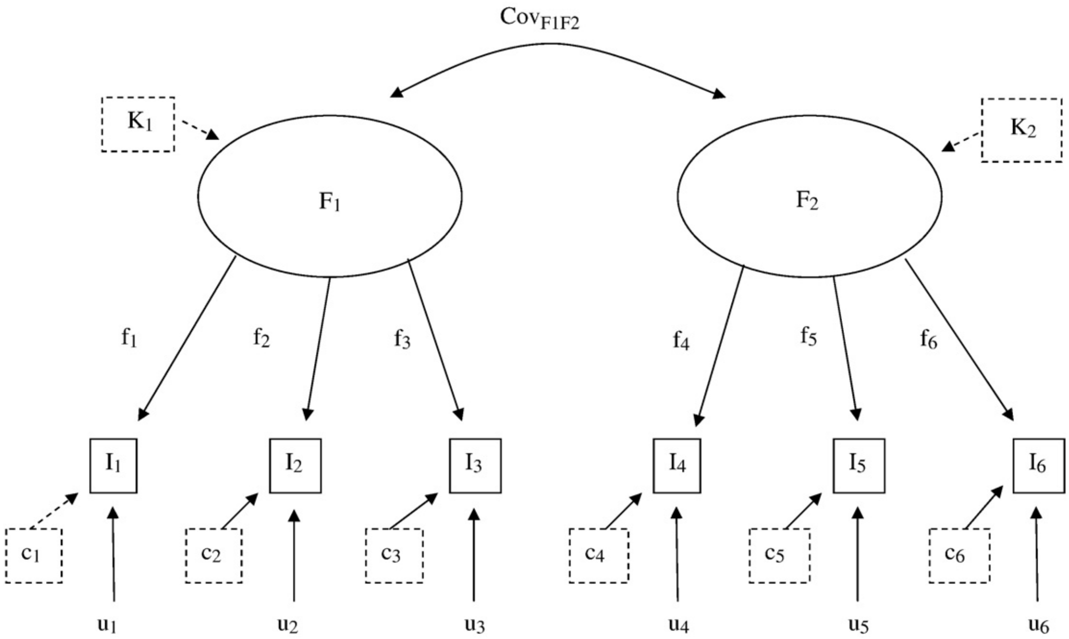

Fig. 1. Confirmatory factor analytic model with two latent factors (F_1 and F_2) each of which is represented by three measures or indicators (I_1 to I_3 and I_4 to I_6). The regression of the indicators on the factors also includes an intercept (c_1 to c_6) and a residual or uniqueness (u_1 to u_6): $I=c+fF+u$. The factors also covary (cov$_{F1F2}$). Each latent factor has a mean (K) and variance.

4. Those researchers interested in testing subgroup latent factor mean differences should proceed to test for scalar invariance. Scalar invariance requires that the intercepts of the regression equations of the observed variables on the latent factors are equivalent across groups. In Fig. 1, this would mean that the constants (c_1 to c_6) are equivalent across groups.
5. The fifth step involves a test of the equality of the uniquenesses (u_1 to u_6) associated with each observed variable. This would mean that the residuals of the regression equations for each indicator (I) are equivalent across groups. This demonstration is sometimes considered a test of the equality of the reliability associated with the measurement of the observed variables, but Vandenberg and Lance (2000) point out that this is legitimate only if the latent factor variances are equal (see also Rock, Werts, & Flaugher, 1978). The first five steps in this evaluation process address issues of measurement invariance whereas the subsequent steps described below involve relationships between the latent factors themselves and are often referred to as issues related to structural invariance. Invariance of uniquenesses is referred to by Meredith (1993) as "strict invariance" and is recognized by most researchers as difficult to achieve and not really necessary to test differences in factor structure or latent means, the substantive questions most often of interest to researchers (Widaman & Reise, 1997; Byrne & Stewart, 2006).
6. The first issue usually addressed in the structural part of this model is whether factor variances are equal across groups. The variances of F_1 and F_2, the latent factors, would be constrained to equality across groups.
7. This is followed by an evaluation of a further constraint that factor covariances (cov_{F1F2} in Fig. 1) are equal.
8. Finally, a test of the equivalence of factor means may represent the most important of the substantive research questions addressed by most researchers. In this case, one mean (k_1 or k_2) is set to 0 and the significance of the other k parameter represents a test of the difference of latent means.

When invariance cannot be established at one of these steps, researchers may also proceed with the test of a model that includes separate estimates of a subset of the subgroup parameters (e.g., some factor loadings, some item intercepts). Such models and the test of the fit of these models are referred to as tests of partial invariance.

Not all of these tests are performed by all researchers and they are certainly not all relevant in any given context. It is also the case that some researchers seem to feel obligated to perform tests that do not seem relevant to their research objectives (see the discussion of tests of equality of uniquenesses below). In the next section of this paper, we provide a review of the papers that have employed CFA analyses to evaluate measurement invariance since the Vandenberg and Lance (2000) review.

2. Assessments of factor invariance

To identify articles that used CFA in assessing factor invariance, we did a search of papers published since 2000 that used the term "measurement invariance." This produced approximately 88 papers. Of these, 75 conducted empirical analyses of measurement invariance using CFA methods. The remainder was discussions or critiques of the CFA method (e.g., Borsboom, 2006). Each of the empirical articles were read to determine the types of invariance considered, the content area addressed by the measures involved, the groups compared, and whether or not the measure had been translated into a language other than the original in which it was developed. The results of the coding of these papers are presented in Table 1.

The substantive issues addressed in these studies were quite diverse. It is clear that measurement invariance is being addressed in a wide variety of research areas in the social and behavioral sciences as well as areas that address human resource issues. The large portion of these papers though did involve topics that human resource and organizational behavior researchers would judge to be in their domain of interest. We believe that the diversity of papers is encouraging as it is clear that researchers are now recognizing the importance of establishing measurement invariance prior to testing or exploring substantively interesting hypotheses about group differences. It is also the case that there are many good papers that are exploring the sources of measurement invariance itself in substantively interesting ways (e.g., Robert, Lee, & Chan, 2006; Anderson, Lievens, van Dam, & Born, 2006; Wicherts et al., 2004). In these papers, the authors test theoretical hypotheses that would predict findings of measurement variance across groups. The bulk of the papers, however, are simply focused on demonstrating measurement invariance to support the use of the instrument across gender, racial, cultural, linguistic, or other demographically diverse subgroups (e.g., Gregorich, 2006; Soh & Leong, 2001; Wicherts, Dolan, & Hessen, 2005).

Not included in the table is a tabulation of the studies that included an examination of the differences in the variance–covariance matrices across groups. This test was reported in 26% of the studies reviewed in Vandenberg and Lance (2000); it was virtually never reported in the studies we reviewed. It is possible that this test almost always yields statistically significant results (except in cases in which there are a very small number of cases) and researchers are no longer bothering to conduct it. Alternatively, they do not perform the test because they plan to pursue more detailed analyses of invariance regardless of its outcome.

Every study we reviewed described tests for configural and metric invariance. Vandenberg and Lance (2000) indicated that 88% reported tests of configural invariance and 99% provided the results of tests of the equality of factor loadings, or metric invariance. The test of configural invariance considered uniformly as the baseline model against which other more constrained models were tested. What was accepted as adequate evidence of configural invariance varied considerably across studies. Most often the researchers started with a hypothesized a priori factor model and simply reported that it provided adequate fit to the data in both samples. What constituted adequate fit was invariably subjective though most authors cited some previous research regarding adequate fit as the basis for accepting or positing a factor pattern matrix. However, other factor models may have been as good or better representations of the data; very little examination of these alternative possibilities is reported in most studies.

Certainly the largest change across the two reviews is the proportion of researchers that examined scalar invariance. In our review, 54% examined the degree to which item intercepts varied across groups although Vandenberg and Lance (2000) reported that only 12% did so. This must certainly be considered an important positive development in the use of CFA models in the

Table 1
Research on measurement invariance

Authors	Area	Groups/composed	Trans.?	Scalar	Uniq.	Factor var.	Factor cov.	Factor means	Part. inv.
Yoo (2002)	Consumer ethnocentrism	Gender/age	No	X	X	X	X	X	
Yoo and Donthu (2001)	Brand equity	Korean–American	Yes						X
Watkins and Canivez (2001)	WISC-III	Time of testing	No		X	X	X		
Munet-Vilaro, Folkman and Gregorich (2002)	Ways of coping	3 Latino Groups	Yes	X	X	X	X	X	
Toporek and Pope-Davis (2001)	Vocational identity	Race (W vs. B) and gender	No						
Thomas, Turkheimer and Oltmanns (2000)	Maudsley Obsessional Compulsive Inventory	Race (W vs. B)	No	X					
Soh and Leong (2002)	Individualism–collectivism	American–Chinese	No	X		X		X	
Wicherts et al. (2004)	Intelligence–Flynn effect	Dutch children across time	No	X				X	
Welkenhuysen-Gybels, Billiet, and Cambre (2003)	Ethnocentricism	Nine West European countries	?						X
Uelschy, Laroche, Tamilia and Yannopoulos (2004)	Dental satisfaction and service quality	US, English and French Canadians	Yes	X				X	
Scholderer, Brunso, Bredahl and Grunert (2004)	Food-related life styles	Eight West European countries	?	X	X	X	X		
Lubke, Dolan, Kelderman and Mellenbergh (2003)	Intelligence	Afr. American vs. Caucasian	No	X	X	X		X	
Levine et al. (2003)	Insomnia	Age groups	No	X	X	X		X	X
Guppy et al. (2004)	Cybernetic coping-stress	Students, police officers, and Social service workersv.	No						X
Eid, Langeheine and Diener (2003)	Satisfaction with life	Chinese vs. Americans	Yes						X
Byrne and Watkins (2003)	Self-concept	Australian vs. Nigerian adolescents	No				X		X
Bowden et al. (2004)	Intelligence	Normative vs. neurologically impaired	No	X	X	X	X	X	X
Atienza et al. (2003)	Satisfaction with life	Males vs. females	No		X	X			X
Antonakis et al. (2003)	Leadership	Males vs. females	No	X	X		X	X	X
Soh and Leong (2001)	Interests	Chinese vs. Americans	No						
Pousette and Hanse (2002)	Job characteristics, health and absenteeism	4 Swedish occupational groups	No	X	X	X		X	
Maydeu-Olivares, Rodriguez-Fornells, Gomez-Benito and D'Zurilla (2000)	Social problem solving	Spanish vs. English students	Yes						X
Motl et al. (2000)	Physical activity	Time and cohort differences	No		X	X			
McArdle, Johnson, Hishinuma, Miyamoto and Anrade (2001)	Depression	Hawaiian and non-Hawaiian males and females	No	X	X	X	X	X	
Land and Long (2000)	Coping	Caregiver groups	No				X		
Durvasula et al. (2001)	Vanity	Young adults: US, China, India, NZ	Yes	X	X	X	X	X	X
Cheung and Rensvold (2000)	Work orientation	US vs. Italians	?	X		X			
Cheng and Watkins (2000)	Self-esteem	Age and gender	No		X		X		X
Cervellon and Dube (2002)	Food preference	Fr. and Eng. Canadians, Chinese	Yes	X	X		X	X	X
Woehr, Arciniega, and Lim (2007)	Work ethic	Korean, Mexican, and US	Yes		X		X		X
Wang and Russell (2005)	Job satisfaction	Chinese vs. Americans	Yes		X		X		X
Tucker, Ozer, Lyubomirsky and Boehm (2006)	Life satisfaction	US, Russian student and community members	Yes	X	X				
Stark et al. (2006)	Simulated data	NA	NA	X					X
Sin et al. (2005)	Marketing orientation	Mainland Chinese vs. Hong Kong Chinese	Yes		X				
Salzberger and Sinkovics (2006)	Technophobia	Britain, Austria and Mexico workers	Yes						X
Roesch and Vaughn (2006)	Dispositional hope	Gender and ethnic status	No	X	X	X	X	X	X
Robert et al. (2006)	Individualism–collectivism	US, Singapore, and Koreans	Yes	X	X	X	X	X	X
Rijkeboer and van den Bergh (2006)	Schemas	Clinical vs. nonclinical	No		X		X		
Reeve and Lam (2005)	Cognitive ability	Retest groups	No	X	X	X	X	X	
Pellegrini and Scandura	Mentoring functioning	Satisfied vs. nonsatisfied groups	No						X
Anderson, Lievens, van Dam and Born (2006)	Assessment center	Males vs. females	No	X	X	X	X	X	
Motl, Dishman, Birnbaum and Lytle (2005)	Depression	Gender and time	No		X	X	X		
Meade, Michels and Lautenschlager (2007)	Personality	Internet vs. paper, choice vs. no choice	No	X	X	X	X	X	
Meade and Lautenschlager (2004a,b)	Simulation	Not applicable	NA	X		X			X
Marshall (2004)	Post-traumatic stress disorder	Spanish vs. English victims	Yes	X					

(continued on next page)

Table 1 (*continued*)

Authors	Area	Groups/composed	Trans.?	Scalar	Uniq.	Factor var.	Factor cov.	Factor means	Part. inv.
Mark and Wan (2005)	Patient satisfaction	Gender and ethnic status	No	X	X	X			
Makikangas et al. (2006)	Health	Time of testing and sample	No	X				X	X
Liu, Borg, and Spector (2004)	Job satisfaction	Language and culture	Yes		X	X	X		
Leone, Van der Zee, van Oudenhoven, Perugini and Ercolani (2005)	Personality	Italian and Dutch	Yes						X
Le, Casillas, Robbins and Langley (2005)	Student readiness	College vs. HS, race, gender	No				X		
Kim, Cramond, and Bandalos (2006)	Creativity	Gender and grade level	No		X	X	X		
Gomez and Fisher (2005)	Spiritual well-being	Gender	No			X	X	X	X
Gaudreau, Sanchez and Blondin (2006)	Affect	2 randomly selected samples	Yes		X	X	X		X
Frenzel, Thrash, Pekrun and Goetz (2007)	Academic emotions	German vs. Chinese	Yes	X				X	X
Du and Tang (2005)	Love of money	Gender and college major	Yes						
Doll, Deng, Raghunathan, Torkzadeh and Xia (2004)	Satisfaction with information systems	Employee and customer groups	No				X		X
Dolan et al. (2006)	Intelligence	Gender	No	X	X			X	X
Del Barrio, Carrasco and Holgado (2006)	Big Five Personality	Age and gender groups in Spain	No		X	X		X	
De Frias and Dixon (2005)	Memory	Age, gender and time	No						
Cole et al. (2006)	Leadership	Web-based vs. paper	Yes	X	X	X	X	X	
Bowden et al. (2006)	Intelligence	Age	No	X	X				X
Bonaccio and Reeve (2006)	Cognitive ability	High and low neuroticism	No	X	X	X	X	X	
Anderson et al. (2005)	Child feeding practices	Blacks and Hispanics	No		X	X	X		
Gregorich (2006)	Depression	Black and White men	No	X	X	X		X	X
Burns, Walsh, Gomez, and Hafetz (2006)	ADHD symptoms	US and Malaysian boys and girls	Yes	X	X	X	X	X	X
Weekley, Ployhart, and Harold (2004)	Personality and situational judgment	Applicant vs. incumbent	No						X
Bryne and Stewart (2006)	Depression	Hong Kong vs. American	?	X					X
Chen, Sousa, and West (2005)	AIDS health outcomes	Working vs. nonworking groups	No	X	X			X	
Grouzet, Otis, Pelletier (2006)	Academic motivation	Gender and time	No	X	X	X	X	X	
Chen and Tang (2006)	Unethical behavior	Psychology and business students	No						
Wicherts et al. (2005)	Intelligence	Minority vs. majority in Holland	No	X	X	X		X	X
Lievens and Anseel (2004)	Organizational citizenship	Peer vs. supervisor ratings	No		X	X	X		X
Feldt, Leskinen and Kinnunen (2005)	Sense of coherence	Unemployed vs. employed groups	No	X				X	
Dierdorff, Surface, Meade, Thompson and Martin (2006)	Organizational climate	Gender, military vs. civilian	No		X	X	X		X
Wang and Waller (2006)	Achievement and appearance	Chinese and American males and females	?	X			X	X	
Crockett, Randall, Shen, Russell and Driscoll (2005)	Depression	Latino and Anglo adolescents	No	X					X

Note. Trans = translation of instrument required for research, Scalar = scalar invariance, Uniq = uniqueness, Var = Variance, Cov. = covariance, Part. inv. = Partial invariance. NA = Not applicable; W = White, B = Black, Afr. = African, Fr = French, Eng. = English, and HS = high school. Because all studies reported tests of configural and metric invariance, columns for these tests were not included in the table.

examination of measurement equivalence. Establishing scalar invariance is necessary in those cases in which researchers also seek to compare the mean responses of the groups on the factors underlying the items. It may either be the case that researchers have become aware of the possibility of testing means and covariances in this model (Chan, 1998) or that the earlier review and some good examples of the use of the technique provided the impetus for change (Little, 1997; Widaman & Reise, 1997; Vandenberg & Self, 1993). Ployhart and Oswald (2004) also provided an excellent exposition of the steps required to test mean hypotheses about differences as well as the role mean difference tests play in establishing invariance. Their paper is also valuable didactically in that they draw parallels between these tests of mean and covariance structure and traditional regression analyses.

The last component of the measurement model is the unique variance (u_1 to u_6) associated with each measured variable. Tests of the invariance of the uniquenesses were completed and reported in 49% of the papers reviewed by Vandenberg and Lance (2000) and 55% of those reported in our Table 1. There still seems to be ambiguity about the need for tests of the invariance of uniquenesses in spite of several clear statements of their importance given various research objectives (Chan, 1998; Vandenberg & Lance, 2000). Some investigators perform and report these tests though they do not seem to be necessary nor are reasons for conducting them articulated (e.g., Cole, Bedeian & Field, 2006; Durvasula, Lysonski & Watson, 2001; Roesch & Vaughn, 2006). These tests are irrelevant if one's purpose is to compare latent means. However, if one's interest is in the comparison of observed mean differences between groups, then demonstrations of metric and scalar invariance are critical and sufficient. If a researcher has specific hypotheses about item uniquenesses or reliability (in the presence of latent factor invariance), this test would be appropriate. Otherwise, it seems superfluous and is not required in most or all cases.

Of the studies described in Table 1, slightly less than half reported aspects of the structural model (i.e., invariance of factor variances and covariances and factor means). Forty-seven percent reported tests of the invariance of factor covariances and variances and 41% reported tests of the differences of factor means. Similar figures for the studies reviewed by Vandenberg and Lance (2000) were 33%, 58%, and 21% for factor variances, covariances, and means, respectively. The increase in tests of factor means would seem to be a positive development because this comparison, as opposed to an evaluation of differences in observed factor means, provides a test of mean group differences that are not some function of measurement error. As in the case of the increase in tests of scalar invariance, researchers seem to have recognized and used this aspect of the CFA model more frequently in the years since the Vandenberg and Lance review. It should also be noted that establishing the invariance of the factor covariances and variances is not a precondition for examination of latent factor means. This is clearly stated in previous treatments of these tests (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), but researchers seem to proceed as though establishment of the invariance of factor covariances and variances is a precondition for a test of the difference of latent factor means.

The frequency with which tests of partial invariance were conducted was not noted in the Vandenberg and Lance (2000) review. In the set of studies summarized in Table 1, 50% reported some test of partial invariance. In most of these cases, the test of partial invariance was entirely empirical; that is, when researchers found evidence for a lack of invariance, they examined modification indices or the relative size of parameters across the groups being compared and allowed parameters to be freed across groups until they were satisfied that the remainder of the parameters were invariant across groups. Regrettably, there was no mention of the use of theory to guide such tests though some researchers did provide post hoc interpretations of the nature of parameters that showed subgroup differences (see Robert et al., 2006) for an especially thorough discussion of these differences). There are a number of issues related to tests of partial invariance that the literature has begun to address (i.e., the impact of allowing parameters to vary on subsequent tests of invariance, the choice of the referent indicator) that are discussed in the next section. Although we cannot make comparisons that would allow us to state that tests of partial invariance are more or less frequent, it is likely that they are becoming more frequent because most early tests and descriptions of the CFA model did not mention partial invariance.

In our examination of studies, we also noted whether or not the instrument involved in the study was translated from one language to another (see Column 4 of Table 1). Twenty of the 75 studies reviewed tested the equivalence of translated instruments, not surprising because it has proven difficult, if not impossible to translate some ideas from one language to another. Our hypothesis was that in those cases in which researchers were translating an instrument, they would be most interested in measurement equivalence rather than issues related to structural equivalence. We did find that researchers were more likely to investigate issues of partial invariance when the instrument was translated than otherwise (57% versus 42%) likely due to the fact that translated measures were more frequently found to lack equivalence across languages (see the papers in Table 1 in which translation of the instrument occurred and was being evaluated). They were less likely to assess factor invariance when the measure was translated than when it was not (33% versus 55%). Translation appeared to play a minimal role in the likelihood that other tests for invariance were conducted including those associated with structural invariance.

In summary, there does seems to be some change in the actual practices of researchers who address issues of measurement invariance since the Vandenberg and Lance (2000) review. The biggest change is that researchers are now much more likely to test for scalar invariance and to test for differences in factor means. It is also the case that in half of the studies reviewed investigators assessed partial invariance. Given what may be an increase in the use of models that include a provision for partially invariant parameters, the next section which includes an examination of new developments in CFA treatments of invariance is particularly important.

3. Continued development of methods for the assessment of measurement invariance

Because it is relatively rare that a researcher finds a measure that is invariant across all sets of parameters, it is not surprising that many more are presenting models that include partial invariance across participant groups in one or more sets of parameters. Researchers have begun to evaluate this practice and the extent to which allowing some parameters to vary across groups affects subsequent tests of parameters. Millsap and Kwok (2004) examined the impact of partial invariance on the selection of members of differing groups when a single factor model represents item responses. One complicating factor when examining the effect of partial invariance is the fact that variability in subgroup parameters across items may be contradictory (i.e., higher in one comparison for one group and lower in comparison for that group for another indicator). Millsap and Kwok's conditions all involved subgroup parameters that differed in the same direction across groups. Millsap and Kwok's results in their Tables 3–7 show that there were not huge changes in selection errors (most often less than 5%) even when only half the items were invariant and usually less than 10% when no items were invariant. Even with this reassuring evidence, some would argue lack of invariance indicates that different underlying factors explain responses in the groups compared and that selection using composites in these instances is inappropriate. The literature on factorial invariance does not provide any guidance with respect to what constitutes levels of invariance that would make a common interpretation of test scores across groups acceptable. This leaves it to the researcher to decide, whereupon they could do an analysis of the type Millsap and Kwok (2004) describe to determine the impact on selection of members of the groups compared. Much more research regarding the importance of partial invariance should be conducted so as to provide context within which to make decisions about the appropriate use of instruments across groups. This is certainly a complex issue if one remembers that one can have partial invariance involving factor loadings, intercepts, and uniquenesses (at the level of the measurement model) and that tests of different types of invariance are done sequentially with possible partial invariance at each level.

In the analysis of metric invariance, one of the indicators must be fixed to 1.00 as a referent indicator. Because this indicator is fixed to the same value across groups, it should involve an indicator that is invariant. Rensvold and Cheung (2001) have provided several ways to identify an appropriate reference indicator under these circumstances and Vandenberg (2002) has suggested the use of exploratory factor analysis (EFA) in each of the groups to identify a reference indicator that has equivalent loadings across respondent groups. Yoon and Millsap (2007) used modification indices to find items that were invariant across groups with a single factor model. They were successful in identifying an invariant indicator when most of the items were, in fact, invariant and the differences between the loadings of nonequivalent items were large and as sample size increased. Use of modification indices in this manner was not as effective under other conditions (i.e., small samples, small differences between factor loadings, and a small proportion of invariant items). As was true for the impact of partial invariance, it is not clear what impact choosing a reference indicator that varies in its relationship to a construct across groups would have on subsequent analyses. This issue should receive continued early attention; in the empirical papers we reviewed authors do not mention that any attention is given to this issue at all.

In many CFAs, researchers have used parcels comprised of several items each as indicators of a construct instead of single item indicators. This practice is preferred for a number of reasons including the fact that parcels have higher reliability and communality than single items and that their distributions more closely approximate normality than do single item indicators. In addition, models that use parcels typically fit better than models that use items as indicators of latent constructs. There are also a small number of studies that evaluate various methods of identifying parcels (e.g., Rogers & Schmitt, 2004) and Bandalos and Finley (2001) recommend parceling only when the parceled items are strictly unidimensional. Recently, Meade and Kroustalis (2006) questioned this practice in tests of measurement invariance. They reported finding 13 studies in which researchers used CFA to test measurement invariance when the indicators were comprised of parcels (see Anderson, Hughes, Fisher & Nicklas, 2005, for an example). The conclusion of a series of simulations which included known lack of invariance at the item level was that the use of parcels obscured findings that identified problematic items in terms of factor loadings or metric invariance. Unexpectedly, they found that items' intercepts were identified as lacking invariance when in fact intercepts were invariant. If establishing measurement invariance is the primary objective (as opposed to testing a structural model when one is reasonably confident in the quality of the measurements being used), then item indicators should be used. Further evaluation of the power of tests of item intercepts, however, should be undertaken.

In most tests of measurement invariance as in other model comparisons using structural equation modeling, there has been a heavy reliance on chi-square tests of the significance of different models. This means comparing models is inconsistent with researchers' use of other model fit indices when evaluating the appropriateness of a single model. Cheung and Rensvold (2002) have challenged the sole use of the chi-square difference test in evaluating differences in model fit. Starting with known differences in models across groups, Cheung and Rensvold compared changes in fit for 20 different indices. The conclusion generally drawn from this study is that the comparative fit index (Bentler, 1990) was the best index of change in fit and that fit changes of .01 or more were practically important (see Table 5, pp. 246–247 for critical values). A tangential, but important finding in light of the Meade and Kroustalis (2006) recommendation that item indicators be used is that fit indices with the exception of RMSEA are affected by the complexity of models so that with highly complex models fit indices will be lower than with low complexity models.

Researchers have recognized the relationship between CFA-based tests of measurement invariance and differential item functioning defined by item response theory (Reise, Widaman, & Pugh, 1993). Recently, several studies have compared the results of these two analytic techniques (e.g., Meade & Lautenschlager, 2004a; Raju, Laffitte, & Byrne, 2002; Stark et al., 2006). Although CFA treats the relationship between an underlying construct and item responses as a linear function and item response theory (IRT) analyses assume a logistic function, item loadings and intercepts in CFA provide similar information as do the item discrimination and item location parameters in IRT. There is no parameter in IRT analogous to the uniquenesses provided in CFA, but the standard error functions associated with trait level provide similar information. Currently, IRT software handles only unidimensional models, though significant advances have been made in developing multidimensional models and analytic solutions (e.g., Reckase, 1997; Yao & Boughton, 2007). Another difference between IRT and CFA approaches is the fact that in IRT, discrimination and location parameters are estimated and tested for group differences simultaneously, although CFA approaches usually involve the test of metric and scalar invariance in separate analytic steps. Stark et al. (2006) found that CFA and IRT were very similar in their

Table 2
Items in example analysis

Agreeableness
1. Make people feel at ease
2. Insult people
3. Feel little concern for others
4. Have a soft heart
5. Take time out for others
Conscientiousness
6. Get chores done right away
7. Leave my belongings around
8. Pay attention to details
9. Am always prepared
10. Am exacting in my work
Emotional Stability
11. Change my mood a lot
12. Get stressed out easily
13. Get upset easily
14. Have frequent mood swings
15. Am easily disturbed

Note. Items 2, 3, 7, and 11–15 were reverse scored.

Table 3
Standardized parameter estimates of the configural model of African-American and Caucasian responses to Agreeableness, Conscientiousness, and Emotional Stability items

	Factor loadings		Uniquenesses	
Variable	African-Am.	Caucasian	African-Am.	Caucasian
Agree 1	.41	.46	.84	.79
Agree 2	.33	.41	.89	.84
Agree 3	.34	.47	.89	.78
Agree 4	.57	.63	.67	.61
Agree 5	.53	.62	.72	.62
Consc 1	.49	.56	.76	.69
Consc 2	.36	.41	.87	.83
Consc 3	.43	.52	.82	.73
Consc 4	.55	.61	.69	.63
Consc 5	.45	.51	.80	.74
ES 1	.70	.78	.51	.40
ES 2	.62	.65	.62	.58
ES 3	.61	.69	.63	.53
ES 4	.78	.84	.40	.29
ES 5	.48	.54	.77	.71

Factor correlations

	Agreeableness	Conscientiousness	Emotional Stability
Agreeableness	1.00	.54	.26
Conscientiousness	.41	1.00	.23
Emotional Stability	.13	.12	1.00

Note. Agree, Consc, and ES refer to the Big Five dimensions of Agreeableness, Conscientiousness and Emotional Stability respectively. African-Am. refers to African-American.
Correlations for the African-American group are above the diagonal; those for the Caucasian group are below the diagonal.

ability to detect items that performed differently across groups. Not surprisingly CFA approaches did better with polytomous data and are the analytic method of choice when there are multiple dimensions. Stark et al. proposed a free-baseline model with Bonferroni corrected *p*-values and simultaneous estimation of loadings and intercepts that performed well in identifying problematic items without increasing Type I error rates. More research on the relative advantages and complementary nature of the information provided by IRT and CFA approaches, as well as the hybrid approach suggested by Stark et al. will certainly help users of these methods.

An issue which seems to be ignored by most of the literature on measurement equivalence is the degree to which lack of measurement equivalence across groups translates into diminished reliability and validity in groups. Millsap's work on this issue, specifically the impact of partial invariance, described above is an exception (Millsap & Everson, 1993; Millsap & Kwok, 2004). He has focused on the implications of variable subgroup parameters on decision accuracy. Meade and Lautenschlager (2004b) examined the identification of item factor loading differences for a limited number of conditions when those differences were known. Traditional CFA approaches were effective when there were large sample sizes, when differences in factor loadings were mixed as opposed to uniformly higher or lower in one group, and when there were a larger number of indicators (12 versus 6). More work on the practical and theoretical importance of nonequivalence of item parameters would clearly facilitate the interpretation of parameter differences and the implications of such differences for theory and for decision making based on the use of instruments.

Some researchers have also presented traditional statistics (i.e., observed means and coefficient alphas) for scales by subgroups in instances in which there were varying degrees of invariance (e.g., Gregorich, 2006; Robert et al., 2006) which should help in ascertaining the practical importance of nonequivalence of item parameters, but few have made comparisons between finding that parameters varied across groups and reliability or validity within these subgroups or the impact on estimated mean differences in

Table 4
Tests of scalar invariance and partial invariance

	χ^2 (*df*)	CFI	RMSEA
Metric invariance	1197.33(186)	.90	.074
Scalar invariance[a]	1348.51(201)	.88	.074
Consc. 2 free	1280.26(200)	.88	.073
Agree 3 free	1255.01(199)	.89	.073
Consc. 5 free	1243.01(198)	.89	.073
Agree 1 free	1232.57(197)	.89	.073
ES 5 free	1219.98(196)	.89	.072

[a] The chi-square difference between the metric invariance model and the scalar invariance model was statistically significant (χ^2_{diff}=151.18, *df*=15, $p<.01$, CFI_{diff}=.02).
Agree, Consc, and ES refer to the Big Five dimensions of Agreeableness, Conscientiousness and Emotional Stability respectively.

Table 5
Estimates of standardized mean differences of latent factors

	Fully invariant model	Partial invariant model	Fully free model[a]
Agreeableness	.09	.08	.06
Conscientiousness	−.07	−.12	−.02
Emotional Stability	.00	.02	.05

[a]To identify the factor mean parameter in this model, it is necessary to constrain the intercept of one indicator per latent variable. Accordingly, we constrained the intercepts of Items 2, 6, and 11 to equality across subgroups.
Negative signs indicate higher scores for Caucasians; positive scores indicate higher scores for African-Americans.

latent variables. As one example of the comparisons we think would be valuable, we provide the example analysis in the next section of the paper.

4. Lack of measurement equivalence and corresponding differences in reliability and mean differences

The data set for this illustration is based on the responses of 680 African-American and 1522 Caucasian college students to fifteen items from the short form of the IPIP (Goldberg, 1999). The items, contained in Table 2 are from measures of Conscientiousness, Agreeableness and Emotional Stability. Only five items were taken from each scale to keep our illustration simple. Variance–covariance matrices for the two groups were used as input to LISREL 8.72 and are the subject of the analyses reported below.

The initial analysis was a test of the difference in the variance–covariance matrix representing the responses of the African-American and Caucasian participants. This test revealed a statistically significant difference (χ^2=441.11, *df*=120, *p*<.01; RMSEA=.051, NNFI=.94, CFI=.97). Fit indices indicated that differences between these matrices are not large. A configural model that represented each item response as a function of a single latent construct (i.e., Agreeableness, Conscientiousness, and Emotional Stability) and a uniqueness component fit the data reasonably well (χ^2=1183.86, *df*=174, *p*<.01; RMSEA=.077, CFI=.90). This configural model was accepted because considerable prior research confirmed the discriminant and convergent validity of these items as measures of these three Big Five traits (Goldberg, 1999). As stated above, the configural model is considered the baseline model against which other hypotheses about subgroup differences are tested. Standardized parameter estimates for this model are provided in Table 3.

The first of the more constrained models usually tested is a test of equivalent factor loadings or metric invariance. This test yielded a nonsignificant change in χ^2 (χ^2_{diff}=13.47, *df*=12, *p*>.05). Given the nonsignificant difference in the chi-squares associated with these two nested models (configural and metric invariance models), we proceeded to test models which specified that the intercepts of the indicators across the two groups were equivalent (i.e., scalar invariance). The test for scalar invariance provided a worse fitting model than did the metric invariant model (χ^2_{diff}=151.18, *df*=15, *p*<.01, CFI_{diff}=.02). This is the point at which accepted practice is to explore partial invariance models. Accordingly, we evaluated a model in which intercepts for the second Conscientiousness item were estimated for both respondent groups as the subgroup intercepts for this item were most different. We continued freeing an additional intercept parameter until the chi-square and CFI for the resultant partially invariant model indicated approximately the same level of fit as the model in which all intercepts were free. Fit indices for these models are presented in Table 4. As can be seen, we freed five intercept parameters in this case each yielding a better fitting model using the chi-square difference test. CFI indices were the same out to two digits.

Developing partially invariant models in this manner raises at least two questions. One is the theoretical meaningfulness of differences in the intercepts for the items involved; that is, are the differences interpretable in light of what is known about the two groups. Similar differences within the IRT model have proven theoretically elusive (Holland & Wainer, 1993; Schmitt, Holland, & Dorans, 1993). On two of these items, the intercepts indicated more positive responses from African-Americans than Caucasians and the reverse was true for the other three items. If one reads the items involved in Table 2 (Items 1, 3, 7, 10, and 15), there is no readily interpretable reason for subgroup differences.

Table 6
Tests of the invariance of uniquenesses

	χ^2 (*df*)	CFI	RMSEA
Scalar Invariance (5 partial ests.)	1219.98(196)	.89	.072
Uniquenesses Invariant.[a]	1526.49(211)	.86	.079
Agree 3 free	1443.34(210)	.87	.076
ES 3 free	1398.81(209)	.88	.075
ES 1 free	1367.23(208)	.88	.074
ES 4 free	1333.42(207)	.88	.073
Consc 3 free	1309.59(206)	.89	.072

Agree, Consc, and ES refer to the Big Five dimensions of Agreeableness, Conscientiousness and Emotional Stability respectively.
[a] The chi-square difference between the scalar partial invariance model and the invariant uniquenesses model was statistically significant (χ^2_{diff}=306.51, *df*=15, *p*<.01, CFI_{diff}=.03).

Table 7
Differences in factor reliability based on different models

	African-American				Caucasian			
	Alpha	u^2 inv.	Part inv.	All free	Alpha	u^2 inv.	Part inv.	All free
Agree	.49	.61	.60	.58	.63	.61	.63	.64
Consc	.56	.61	.61	.60	.64	.61	.63	.64
ES	.77	.81	.79	.78	.83	.81	.84	.83

Note. u^2 inv. refers to a model in which all uniquenesses were constrained equal across groups. Part. inv. refers to a model in which the 5 uniquenesses described in Table 6 were freely estimated across groups and All free is a model in which all uniquenesses were estimated for both groups. Agree, Consc, and ES refer to the Big Five dimensions of Agreeableness, Conscientiousness and Emotional Stability respectively. Inv refers to invariant.

A second question relates to the effect that allowing partial invariance at one step in the process of testing for measurement invariance has on subsequent tests for differences in the structural or measurement components of the model. Intercept differences should have the greatest impact on tests for differences in latent factor mean differences. Hence we examined these differences under models in which factor means were estimated when (1) all intercept parameters with the exception of one for each construct were freely estimated; (2) all intercept parameters were constrained to equality; and (3) when the five intercepts designated as nonequivalent in Table 4 were freely estimated across groups. Estimated standardized factor means under these different models are provided in Table 5. As can be seen, there are minimal differences in estimated factor means. These differences, of course, might be larger or smaller given all intercept differences were in the same direction and/or the magnitude of such differences were larger. It should be noted that in the fully free model, one intercept parameter per factor was constrained to equality across groups to allow identification of the factor mean.

The model in which intercepts were treated as partially invariant (the last model in Table 4) was then compared with a model in which uniquenesses were constrained to equality across groups. The comparison of these two models yielded a statistically significant difference (χ^2_{diff}=306.51, df=15, p<.01, CFI$_{\text{diff}}$=.03). As was true for scalar differences, we sequentially estimated models in which an additional uniqueness was estimated in both groups so as to develop a model that was close in fit to the model in which all uniqueness parameters were freely estimated in both groups. In this case, uniquenesses were uniformly smaller in the Caucasian group than in the African-American group. The fit of the first five of these models is provided in Table 6.

Because uniquenesses are supposed to represent lack of reliability when factor variances are fixed, we then proceeded to fix factor variances and covariances (i.e., Steps 6 and 7 in the original model) and found nonsignificant and relatively small differences as a function of constraining these six parameters to equality across groups. We then estimated reliability of the latent factor scores (the squared sum of the factor loadings divided by this value plus the sum of the uniquenesses). The resultant differences in reliability across three models (all uniquenesses invariant, all uniquenesses freely estimated in both groups and a partially invariant model with five uniquenesses freely estimated in both groups) are provided for both groups in Table 7. For comparison purposes, we also provide coefficient alpha values for the scales in both groups. As can be seen, reliabilities for the Caucasian group are all very similar across models including values of coefficient alpha. However, for the African-American group, there are perhaps nontrivial differences in reliability estimates (up to .03) across the three SEM models and the value of alpha is quite a lot lower than the SEM estimates, particularly for the Agreeableness and Conscientiousness factors. As would be expected these reliability differences did impact the intercorrelations of the latent factors—mostly so for the relationship between Agreeableness and Conscientiousness for the African-American group. In this instance, the estimate of the correlation between these two latent factors was .54 when all uniquenesses were freely estimated and .49 when all were constrained to equality across groups. In all other cases for both groups of respondents, differences in estimated factor correlations were less than .02. Individual investigators will have to make their own judgments about the practical significance of the differences associated with the fully constrained, partially constrained, or freely estimated models of the uniquenesses, but they impress us as relatively inconsequential in this instance.

5. Conclusions

Our review of studies conducted since 2000 that have assessed measurement invariance suggests that examinations of scalar invariance and factor mean differences are much more frequent than they were in the literature reviewed by Vandenberg and Lance (2000). All investigators estimate configural and metric invariance, though assessments of configural invariance often seem relatively cursory. Few investigators test the significance of the difference of the variance–covariance matrices. It is also the case that there have been no studies of which we are aware in which someone has tested the sensitivity of this overall test of the difference of the variance–covariance matrices in the presence of invariance in other parts of the model or whether the overall test is too powerful and will virtually always lead to a conclusion that there are group differences even when group sizes are quite small. French and Finch (2006) have conducted simulations that focused on the accuracy of the common CFA steps; more similar work on the overall test of invariance should be performed. Finally, although no assessment of the frequency of tests of partial invariance was conducted by Vandenberg and Lance, slightly over half of the studies in our review did so.

There are several important assessments or developments in the use of confirmatory factor analyses to test for measurement invariance. These include guides for the use of the CFI index in comparing models (Cheung & Rensvold, 2002), the choice of referent indicators in multi-group analyses (Rensvold & Cheung, 2001), the use of parcels in measurement invariance studies (Meade & Kroustalis, 2006), and the impact of partial invariance on selection of members of different subgroups using different selection ratios

(Millsap & Kwok, 2004). In addition, there have been several comparisons of measurement invariance using SEM and IRT-based methods, and Stark et al. (2006) have proposed a model that incorporates aspects of both approaches.

Finally, because partial invariance tests appear to be more commonly applied than was the case in the literature reviewed by Vandenberg and Lance (2000), we provided some examples of the evaluation of the impact of partial invariance on the estimate of factor means, factor reliabilities and factor intercorrelations. Using this example, we concluded that partial invariance made little difference in the estimates of structural model parameters. This example is not provided as the final word on this issue, but merely as representative of studies that should be conducted in a much more systematic manner varying different aspects of the data and models. Simulations, like that of Millsap and Kwok (2004) are likely to provide the most definitive answers to the impact of evaluating and using partial invariance in our models. It should also be pointed out that there may be substantive reasons to test models of partial invariance and in these cases the concern about estimates of other aspects of the model may be less relevant.

References

Anderson, C. B., Hughes, S. O., Fisher, J. O., & Nicklas, T. A. (2005). Cross-cultural equivalence of feeding beliefs and practices: The psychometric properties of the child feeding questionnaire among Blacks and Hispanics. *Preventive Medicine*, *41*, 521−531.

Anderson, N., Lievens, F., van Dam, K., & Born, M. (2006). A construct-driven investigation of gender differences in a leadership-role assessment center. *Journal of Applied Psychology*, *91*, 555−566.

Antonakis, J., Avolio, B. J., & Sivasubramaniam, N. (2003). Context and leadership: An examination of the nine-factor full-range leadership theory using the Multifactor Leadership Questionnaire. *The Leadership Quarterly*, *14*, 261−295.

Atienza, F. L., Balaguer, I., & Garcia-Merita, M. L. (2003). Satisfaction with life scale: Analysis of factorial invariance across sexes. *Personality and Individual Differences*, *35*, 1255−1260.

Bandalos, D. J., & Finley, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides (Ed.), *New developments and techniques in structural equation modeling* (pp. 269−296). Mahwah, NJ: Lawrence Erlbaum.

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

Bonaccio, S., & Reeve, C. L. (2006). Differentiation of cognitive abilities as a function of neuroticism level: A measurement equivalence/invariance analysis. *Intelligence*, *34*, 403−417.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care*, *44*, 176−181.

Bowden, S. C., Cook, M. J., Bardenhagen, F. J., Shores, E. A., & Carstairs, J. R. (2004). Measurement invariance of core cognitive abilities in heterogeneous neurological and community samples. *Intelligence*, *32*, 363−389.

Bowden, S. C., Weiss, L. G., Holdnack, J. A., & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale-III. *Psychological Assessment*, *18*, 334−339.

Burns, G. L., Walsh, J. A., Gomez, R., & Hafetz, N. (2006). Measurement and structural invariance of parent ratings of ADHD and ODD symptoms across gender for American and Malaysian children. *Psychological Assessment*, *18*, 452−457.

Byrne, B. M., & Stewart, S. M. (2006). The MACS approach to testing for multigroup invariance of a second-order factor structure: A walk through the process. *Structural Equation Modeling*, *13*, 287−321.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology*, *34*, 155−175.

Cervellon, M., & Dube, L. (2002). Assessing the cross-cultural applicability of affective and cognitive components of attitude. *Journal of Cross-Cultural Psychology*, *33*, 346−357.

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*, *83*, 234−246.

Cheng, C. H. K., & Watkins, D. (2000). Age and gender invariance of self-concept factor structure: An investigation of a newly developed Chinese self-concept instrument. *International Journal of Psychology*, *35*, 186−193.

Chen, F. F., Sousa, K. H., & West, S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling*, *12*, 471−492.

Chen, Y., & Tang, T. L. (2006). Attitude toward and propensity to engage in unethical behavior: Measurement invariance across major among university students. *Journal of Business Ethics*, *69*, 77−93.

Cheung, G. W., & Rensvold, R. B. (2000). Assessing extreme and acquiesence response sets in cross-cultural research using SEM. *Journal of Cross-Cultural Psychology*, *31*, 187−212.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233−255.

Cole, M. S., Bedeian, A. G., & Field, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods*, *9*, 339−368.

Crockett, L. J., Randall, B. A., Shen, Y., Russell, S. T., & Driscoll, A. K. (2005). Measurement invariance of the center for epidemiological studies depression scale for Latino and Anglo adolescents: A national study. *Journal of Consulting and Clinical Psychology*, *73*, 47−58.

De Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment*, *17*, 168−178.

Del Barrio, V., Carrasco, M. A., & Holgado, F. P. (2006). Factor structure invariance in the children's Big Five questionnaire. *European Journal of Psychological Assessment*, *22*, 158−167.

Dierdorff, E. C., Surface, E. A., Meade, A., Thompson, L. F., & Martin, D. L. (2006). Group differences and measurement equivalence: Implications for command climate survey research and practice. *Military Psychology*, *18*, 19−37.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between the WAIS-III common factors and sex and educational attainment in Spain. *Intelligence*, *34*, 193−210.

Doll, W. J., Deng, X., Raghunathan, T. S., Torkzadeh, G., & Xia, W. (2004). The meaning and measurement of user satisfaction: A multigroup invariance analysis of the end-user computing satisfaction instrument. *Journal of Management Information Systems*, *21*, 227−262.

Du, L., & Tang, T. L. (2005). Measurement invariance across gender and major: The love of money among university students in People's Republic of China. *Journal of Business Ethics*, *59*, 281−293.

Durvasula, S., Lysonski, S., & Watson, J. (2001). Does vanity describe other cultures? A cross-cultural examination of the vanity scale. *Journal of Consumer Affairs*, *35*, 180−199.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. *Journal of Cross-Cultural Psychology*, *34*, 195−210.

Feldt, T., Leskinen, E., & Kinnunen, U. (2005). Structural invariance and stability of sense of coherence: A longitudinal analysis of two groups with different employment experiences. *Work and Stress*, *19*, 68−93.

French, B. F., & Finch, W. H. (2006). Confirmatory factor-analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*, 378−402.

Frenzel, A. C., Thrash, T. M., Pekrun, R., & Goetz, T. (2007). Achievement emotions in Germany and China: A cross-cultural validation of the Academic Emotions Questionnaire-Mathematics. *Journal of Cross-Cultural Psychology*, *38*, 302−309.

Gaudreau, P., Sanchez, X., & Blondin, J. (2006). Positive and negative affective states in a performance-related setting. *European Journal of Psychological Assessment*, *22*, 240−249.

Goldberg, L. R. (1999). A broad-bandwidth public-domain personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe*, *Vol. 7.* (pp. 7−28)Tilburg, The Netherlands: Tilburg University Press.

Gomez, R., & Fisher, J. W. (2005). The spiritual well-being questionnaire: Testing for model applicability, measurement and structural equivalencies, and latent mean differences across gender. *Personality and Individual Differences*, *39*, 1383−1393.

Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? Testing measurement invariance using the confirmatory analysis framework. *Medical Care*, *44*, S78−S94.

Grouzet, F. M. E., Otis, N., & Pelletier, L. G. (2006). Longitudinal cross-gender factorial invariance of the academic motivation scale. *Structural Equation Modeling*, *13*, 73−98.

Guppy, A., Edwards, J. A., Brough, P., Peters-Bean, K. M., Sale, C., & Short, E. (2004). The psychometric properties of the short version of the Cybernetic Coping Scale: A multigroup CFA across four samples. *Journal of Occupational and Organizational Psychology*, *77*, 39−62.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Erlbaum.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide in measurement invariance in aging research. *Experimental Aging Research*, *18*, 117−144.

Kim, K. H., Cramond, B., & Bandalos, D. L. (2006). The latent structure and measurement invariance of scores on the Torrance tests of creative thinking-figural. *Educational and Psychological Measurement*, *66*, 459−477.

Land, H., & Long, J. D. (2000). The structure of coping in AIDS caregivers: A factor analytically derived measure. *Journal of Applied Social Psychology*, *30*, 463−483.

Le, H., Casillas, A., Robbins, S. B., & Langley, B. (2005). Motivational and skills, social and self-management predictors of college outcomes: Constructing the student readiness inventory. *Educational and Psychological Measurement*, *65*, 482−508.

Leone, L., Van der Zee, K. I., van Oudenhoven, J. P., Perugini, M., & Ercolani, A. P. (2005). The cross-cultural generalizability and validity of the multicultural personality questionnaire. *Personality and Individual Differences*, *38*, 1449−1462.

Levine, D. W., Kaplan, R. M., Kripke, D. F., Bowen, D. J., Naughton, M. J., & Shumaker, S. A. (2003). Factor structure and measurement invariance of the Women's Health Initiative Insomnia Rating Scale. *Psychological Assessment*, *15*, 123−136.

Lievens, F., & Anseel, F. (2004). Confirmatory analysis and invariance of an organizational citizenship behaviour measure across samples in a Dutch-speaking context. *Journal of Occupational and Organizational Psychology*, *77*, 299−306.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53−76.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German job satisfaction survey used in a multinational organization: Implications of Schwarz's culture model. *Journal of Applied Psychology*, *89*, 1070−1082.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543−566.

Makikangas, A., Feldt, T., Kinnunen, U., Tolvanen, A., Kinnunen, M., & Pulkkinen, L. (2006). The factor structure and factorial invariance of the 12-item general health questionnaire (GHQ-12) across time: Evidence from two community-based samples. *Psychological Assessment*, *18*, 444−451.

Mark, B. A., & Wan, T. T. H. (2005). Testing measurement equivalence in a patient satisfaction instrument. *Western Journal of Nursing Research*, *27*, 772−787.

Marshall, G. N. (2004). Posttraumatic stress disorder symptom checklist: Factor structure and English–Spanish measurement invariance. *Journal of Traumatic Stress*, *17*, 223−230.

Maydeu-Olivares, A., Rodriguez-Fornells, A., Gomez-Benito, J., & D'Zurilla, T. J. (2000). Psychometric properties of the Spanish adaptation of the Social Problem-solving Inventory—Revised (SPSI-R). *Personality and Individual Differences*, *29*, 699−708.

McArdle, J. J., Johnson, R. C., Hishinuma, E. S., Miyamoto, R. H., & Anrade, N. N. (2001). Structural equation modeling of group differences in CES-D ratings of native Hawaiian and non-Hawaiian high school students. *Journal of Adolescent Research*, *16*, 108−149.

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

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies in establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361−388.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*, 60−72.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are internet and paper-and-pencil personality tests truly comparable? *Organizational Research Methods*, *10*, 322−345.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525−543.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied psychological measurement*, *17*, 297−334.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93−115.

Motl, R. W., Dishman, R. K., Birnbaum, A. S., & Lytle, L. A. (2005). Longitudinal invariance of the center for epidemiologic studies—Depression scale among girls and boys in middle school. *Educational and Psychological Measurement*, *65*, 90−108.

Motl, R. W., Dishman, R. K., Trost, S. G., Saunders, R. P., Dowda, M., Felton, G., et al. (2000). Factorial validity and invariance of questionnaires measuring social–cognitive determinants of physical activity among adolescent girls. *Preventive Medicine*, *31*, 584−594.

Munet-Vilaro, F., Gregorich, S. E., & Folkman, S. (2002). Factor structure of the Spanish version of the ways of coping questionnaire. *Journal of Applied Social Psychology*, *32*, 1938−1954.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods*, *7*, 27−65.

Pousette, A., & Hanse, J. J. (2002). Job characteristics as predictors of ill-health and sickness absenteeism in different occupational types—A multigroup structural equation modeling approach. *Work and Stress*, *16*, 229−250.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*, 25−36.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552−566.

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*, 535−549.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim, & L. L. Neider (Eds.), *Equivalence in measurement Research in management*, *Vol. I.*. (pp. 21−50) Greenwich, CT: Information Age.

Rijkeboer, M. M., & van den Bergh, H. (2006). Multiple group confirmatory factor analysis of the Young Schema-Questionnaire in a Dutch clinical versus non-clinical population. *Cognitive Therapy Research*, *30*, 263−278.

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

Rock, D. A., Werts, C. E., & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing psychometric properties of multiple variables across populations. *Multivariate Behavioral Research*, *13*, 403−418.

Roesch, S. C., & Vaughn, A. A. (2006). Evidence for the factorial validity of the dispositional hope scale. *European Journal of Psychological Assessment*, *22*, 78−84.

Rogers, W. M., & Schmitt, N. (2004). Parameter recovery and model fit using multidimensional composites: A comparison of four empirical parceling algorithms. *Multivariate Behavioral Research*, *39*, 379−412.

Salzberger, T., & Sinkovics, R. R. (2006). Reconsidering the problem of data equivalence in international marketing research: Contrasting approaches based on CFA and the Rasch model for measurement. *International Marketing Review*, *23*, 390−417.

Schmitt, A. P., Holland, P. W., & Dorans, N. J. (1993). In P. W. Holland, & H. Wainer (Eds.), *Evaluating hypotheses about differential item functioning* (pp. 281–316). Differential item functioning. Hillsdale, NJ: Erlbaum.

Scholderer, J., Brunso, K., Bredahl, L., & Grunert, K. G. (2004). Cross-cultural validity of the food-related lifestyles instrument (FRL) within Western Europe. *Appetite*, *4*, 197–211.

Sin, L. Y. M., Tse, A. C. B., Yau, O. H. M., Chow, R. P. M., Lee, J. S. Y., & Lau, L. B. Y. (2005). Relationship marketing orientation: Scale development and cross-cultural validation. *Journal of Business Research*, *58*, 185–194.

Soh, S., & Leong, F. T. L. (2001). Cross-cultural validation of Holland's theory in Singapore: Beyond structural validity of RIASEC. *Journal of Career Assessment*, *9*, 115–133.

Soh, S., & Leong, F. T. L. (2002). Validity of vertical and horizontal individualism and collectivism in Singapore: Relationships with values and interests. *Journal of Cross-cultural Psychology*, *33*, 3–15.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting DIF with CFA and IRT: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Steenkamp, J. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Thomas, J., Turkheimer, E., & Oltmanns, T. F. (2000). Psychometric analysis of racial differences on the Maudsley Obsessional Compulsive Inventory. *Assessment*, *7*, 247–258.

Toporek, R. L., & Pope-Davis, D. B. (2001). Comparison of vocational identity factor structures among African-American and White American college students. *Journal of Career Assessment*, *9*, 135–151.

Tucker, K. L., Ozer, D. J., Lyubomirsky, S., & Boehm, J. K. (2006). Testing for measurement invariance in the satisfaction with life scale: A comparison of Russians and North Americans. *Social Indicators Research*, *78*, 341–360.

Ueltschy, L. C., Laroche, M., Tamilia, R. D., & Yannopoulos, P. (2004). Cross-cultural invariance of measures of satisfaction and service quality. *Journal of Business Research*, *57*, 901–912.

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*, 4–69.

Vandenberg, R. J., & Self, R. M. (1993). Assessing newcomers' changing commitment to the organization during the first 6 months of work. *Journal of Applied Psychology*, *78*, 557–568.

Wang, M., & Russell, S. S. (2005). Measurement equivalence of the Job Descriptive Index across Chinese and American workers: Results from confirmatory factor analysis and item response theory. *Educational and Psychological Measurement*, *65*, 709–732.

Wang, P. Z., & Waller, D. S. (2006). Measuring consumer validity: A cross-cultural validation. *Psychology and Marketing*, *23*, 665–687.

Watkins, M. W., & Canivez, G. L. (2001). Longitudinal factor structure of the WISC-III among students with disabilities. *Psychology in the Schools*, *38*, 291–298.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement and subgroup differences. *Human Performance*, *17*, 435–461.

Welkenhuysen-Gybels, J., Billiet, J., & Cambre, B. (2003). Adjustment for acquiescence in the assessment of the construct equivalence of Likert-type score items. *Journal of Cross-Cultural Psychology*, *34*, 702–722.

Wicherts, J. M., Conor, V. D., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology*, *89*, 696–716.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention* (pp. 281–324). Washington, DC: American Psychological Association.

Woehr, D. J., Arciniega, L. M., & Lim, D. H. (2007). Examining work ethic across populations: A comparison of the Multidimensional Work Ethic Profile across three diverse cultures. *Educational and Psychological Measurement*, *67*, 154–168.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*, *31*, 83–105.

Yoo, B. (2002). Cross-group comparisons: A cautionary note. *Psychology and Marketing*, *19*, 357–368.

Yoo, B., & Donthu, N. (2001). Developing and validating a multidimensional consumer-based brand equity scale. *Journal of Business Research*, *52*, 1–14.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, *14*, 435–463.