PSYCHOMETRIKA—VOL. 72, NO. 4, 461–473
DECEMBER 2007
DOI: 10.1007/S11336-007-9039-7

INVARIANCE IN MEASUREMENT AND PREDICTION REVISITED

ROGER E. MILLSAP

ARIZONA STATE UNIVERSITY

Borsboom (Psychometrika, 71:425–440, 2006) noted that recent work on measurement invariance (MI) and predictive invariance (PI) has had little impact on the practice of measurement in psychology. To understand this contention, the definitions of MI and PI are reviewed, followed by results on the consistency between the two forms of invariance in the general case. The special parametric cases of factor analysis (strict factorial invariance) and linear regression analyses (strong regression invariance) are then described, along with findings on the inconsistency between the two forms of invariance in this context. Two numerical examples of inconsistency are reviewed in detail. The impact of violations of MI on accuracy of selection is illustrated. Finally, reasons for the slow dissemination of work on invariance are discussed, and the prospects for altering this situation are weighed.

Key words: measurement invariance, predictive invariance, factorial invariance, test bias, selection accuracy.

Introduction

Borsboom (2006) argued that modern advances in psychometrics have failed to penetrate the actual practice of measurement among psychologists. He gave a number of examples of advances that have not had much impact on practice. One of these examples concerned work by myself and others on measurement invariance and its relation to invariance in prediction. Borsboom pointed out that this work has been generally ignored, particularly in the formulation of testing standards (AERA, APA & NCME, 1999; Society for Industrial/Organizational Psychology, 2003). Not long after the publication of Borsboom (2006) (and independently of that publication), I was asked by a colleague who is an industrial/organizational psychologist to explain why my work on invariance has had no visible impact on testing practice in that field. It is sobering to be asked such questions, and to ponder what answers may exist.

I must concede that both Borsboom (2006) and my colleague are correct: The body of work on invariance in measurement and prediction has yet to have much impact on measurement practice. For example, it is still true that many psychologist's views about bias in testing are based primarily on studies that compare test/criterion regressions across populations (Hunter & Schmidt, 2000; Neisser, Boodoo, Bourchard, Boykin, Brody, Ceci, Halpern, Loehlin, Perloff, Sternberg, & Urbina, 1996; Sackett, Schmitt, Ellington, & Kabin, 2001). This basis for judging bias is enshrined in current testing standards, as noted by Borsboom. Yet conclusions about test bias that rely primarily on invariance in test/criterion regressions or correlations are demonstrably flawed (Millsap, 1995). The many empirical primary and meta-analytic studies on invariance in regressions and correlations, while providing useful information, cannot fully support conclusions about bias or lack of bias in measurement. Furthermore, some relatively simple diagnostic procedures that could be used to more fully examine bias in prediction and measurement are not being used.

This paper is based on the Presidential Address given at the International Meeting of the Psychometric Society in Tokyo, Japan, on July 11, 2007. This research was supported by National Institute of Mental Health grants 1P30 MH 068685-01A1 and RO1 MH64707-01.

Requests for reprints should be sent to Roger E. Millsap, Department of Psychology, Box 871104, Arizona State University, Tempe, AZ 85287-1104, USA. E-mail: millsap@asu.edu

What is the explanation for this gap between psychometric theory and actual measurement practice? Borsboom (2006) provided some cogent explanations, as did some of the commentators on that paper (e.g., Clark, 2006). Here I will offer my perspective. In what follows, I first review the concepts of measurement invariance and predictive invariance. The relationship between the two forms of invariance are then reviewed, both for the general nonparametric case and for the specific examples of factorial invariance and regression invariance. Two specific cases in which these forms of invariance are inconsistent are then fully described. A numerical example is given to illustrate these cases, along with their implications for selection accuracy. The final section returns to the question of why these results are not more widely known, and what might be done to disseminate psychometric advances.

What Is Measurement Invariance?

At its root, the notion of measurement invariance (MI) is that some properties of a measure should be independent of the characteristics of the person being measured, apart from those characteristics that are the intended focus of the measure. This definition requires elaboration because it brings together some disparate concepts. First, what do we mean by a “measure”? MI is not tied to any specific type of test or item. It could apply to individual test items, blocks of test items or testlets, subtests, or whole tests. It could also be applied to ratings or judgments made by a set of raters in relation to a set of rates (this application is not considered further here). Finally, no particular scale properties for the measure are assumed; MI could apply to discrete nominal or ordinal scores, or to continuous interval scores.

A second consideration lies in the measurement “properties” that are expected to be independent of examinee characteristics. We don’t expect that all properties of a measure will be invariant. The average score on a measure will generally vary as a function of many examinee characteristics, for example. Similarly, the reliability of a measure is not viewed as an invariant property of the measure, given that variation in true scores may itself be different across groups of examinees. On the other hand, if a measure fits a common factor model, we expect that the unstandardized factor loading(s) for the measure will be invariant under fairly broad conditions (Bloxom, 1972; Meredith, 1964a, 1964b). As shown below, the key to distinguishing which properties of a measure should be invariant from those properties that are not is the definition of invariance in terms of conditional probabilities.

The definition of MI also requires one to distinguish between characteristics of the person that are the “focus” of the measure, and those characteristics that are irrelevant to this focus (Ackerman, 1992; Kok, 1988; Shealy & Stout, 1993; Stout, 1990). It should be obvious that the notion of measurement invariance is rendered vacuous if one takes the position that the “focus” of the measure is simply defined by the content alone (e.g., “intelligence is what intelligence tests measure”). On this view, there can be no question of invariance because the focus of the test is defined solely by its content. Score differences between individuals who take the same test might be attributable partly to measurement error, but there can be no coherent definition of “bias” in measurement because there is no clear definition of characteristics that are irrelevant to the focus of the test. MI requires some a priori definition of the intended focus of the measure: What is it that we are trying to measure? In psychological measurement, the characteristics that are the focus of a measure are usually formally defined as latent variables. The question then becomes: What are the intended latent variables to be targeted by the measure?

Nothing in the definition of MI requires the intended latent variables to be unitary, with only one intended latent dimension. Some confusion exists on this point (e.g., Hunter & Schmidt, 2000). It is true that some of the latent variable models used to investigate violations of MI routinely assume unidimensionality, examples being models based on unidimensional item response theory (IRT). In contrast, studies of factorial invariance have been conducted for almost 70 years (e.g., Thomson & Lederman, 1939) using multiple factor models. Hence there is no methodological requirement of unidimensionality in studies of MI. For studies in which a single latent variable is defined as the intended latent variable for the measure, the presence of additional latent variables underlying the measure may indeed trigger violations of MI. Such findings are probably appropriate if the additional latent variables cannot be explained or identified through post hoc analyses.

With these considerations in mind, the definition of measurement invariance is traditionally expressed using conditional probability. Let **X** be the $q \times 1$ vector of random variables representing scores on the observed measures under study. Let **W** be an $r \times 1$ vector of the intended latent variables for **X**. Let **V** be an $s \times 1$ vector of measured variables that define the person characteristics of interest that should be irrelevant to **X** once **W** is considered. In many cases, $s = 1$ and **V** is a scalar group identifier that defines demographic variables such as gender or ethnicity. Measurement invariance (MI) of **X** in relation to **W** and **V** is defined to hold if and only if

$$P(\mathbf{X}|\mathbf{W}, \mathbf{V}) = P(\mathbf{X}|\mathbf{W}) \tag{1}$$

for all **X**, **W**, **V**, where $P(A|B)$ is the conditional probability function for A given B (Lord, 1980; Mellenbergh, 1989; Meredith & Millsap, 1992). Here this probability can be expressed either as a discrete conditional probability for discrete **X**, or as a conditional probability density function for continuous **X**. The general notation in the above is intended to apply to either case, depending on the context.

Investigations of invariance in (1) arise in many contexts, varying with the type of measure **X** and the type of model that describes the relation of **X** and **W**. When **X** fits a common factor model with common factors **W**, MI in (1) implies factorial invariance. Factorial invariance has a long history (Ahmavaara, 1954; Thomson & Lederman, 1939; Thurstone, 1947). Factorial invariance itself is ordinarily weaker than invariance in (1) because only the first and second conditional moment structure is studied in factorial invariance investigations, while (1) requires invariance in conditional distributions. When **X** consists of item scores that fit one of the models in item response theory (IRT), investigations of invariance in (1) evaluate differential item functioning (DIF) (Thissen, Steinberg, & Wainer, 1988). Here **W** is a continuous latent variable, typically unidimensional. Alternatively, **W** might be defined as a latent class identifier, with **X** fitting a latent class model across populations defined by **V**. A traditional focus of invariance studies has been in cognitive and achievement tests in educational settings, but invariance studies in other research contexts are becoming more common. In translated measures, invariance investigations examine equivalence of measures across language groups (Drasgow & Probst, 2004; Hambleton, Merenda, & Spielberger, 2006). Measures of attitudes, personality attributes, and other noncognitive attributes are also studied for invariance (Byrne, 1994; Hofer, Horn, & Eber, 1997; Pentz & Chou, 1994). A further application lies in randomized or quasi-experimental studies in which the groups are defined experimentally by the treatment received, and the goal is to check whether the relation of **X** to **W** is altered by the treatment (Millsap & Hartog, 1988; Riordan, Richardson, Schaffer, & Vandenberg, 2001).

What Is Predictive Invariance?

Much of the literature on "test bias" in applied psychological research has focused on predictions based on test scores, rather than on measurement invariance in (1) (Cleary, 1968; Jensen, 1980). For example, suppose that we partition $\mathbf{X} = (Y, \mathbf{Z})$ with Y a scalar criterion measure of interest, and **Z** a $p \times 1$ vector of predictor variables with $p = q - 1$. To illustrate,

Y might be a measure of job performance and $\mathbf{Z}$ might be a set of selection measures used to select prospective employees. Or Y could be grade point average and $\mathbf{Z}$ could represent the SAT Verbal and Math scores. In this prediction context, a different notion of invariance arises based on the relationship of Y to $\mathbf{Z}$, and whether this relationship varies depending on other person characteristics. Defining $\mathbf{V}$ as before, we can define predictive invariance (PI) for $\mathbf{Z}$ in relation to Y and $\mathbf{V}$ as existing if and only if

$$P(Y|\mathbf{Z},\mathbf{V}) = P(Y|\mathbf{Z}) \tag{2}$$

for all $Y, \mathbf{Z}, \mathbf{V}$ (Meredith & Millsap, 1992; Millsap, 1995). For example, suppose that Y is a binary variable that indicates whether a person scores above a fixed threshold on an observed criterion measure (e.g., dollar sales), Z is a selection test score for salespersons, and V is a gender identifier. Under predictive invariance, the probability that anyone exceeds the threshold on dollar sales, given their selection test score, is the same regardless of gender. Conversely, if invariance in (2) fails to hold, it means that within groups defined by a common selection test score Z, there are gender difference in the probability of surpassing the dollar sales threshold. A similar description applies if Y is taken as actual dollar sales. In this case, a violation of PI occurs when there are gender differences in the distribution of dollar sales within groups defined by a common selection test score Z.

Several points should be noted about the definition of PI in (2). First, no latent variables appear in (2). In many applications, $\mathbf{Z}$ will ultimately be used as a basis for decisions involving selection or access to resources (e.g., clinical treatment). The decisions are made on the basis of $\mathbf{Z}$, rather than any latent variables that underlie $\mathbf{Z}$. For this reason, the relationship of Y to $\mathbf{Z}$ is of intrinsic interest. Second, PI in (2) is more stringent than is desired in many applications when Y is continuous. In this case, standard practice is to model the relationship of Y to $\mathbf{Z}$ using linear regression, followed by an investigation of whether the regression functions vary over $\mathbf{V}$. Studies of "differential prediction" are studies of this form of PI; this topic has a long history (Cleary, 1968; Humphreys, 1952; Potthoff, 1966).

Investigations of PI arise in many domains. As noted, traditional studies of differential prediction in educational measurement and industrial/organizational psychology using cognitive tests as predictors are familiar examples (e.g., Bridgeman & Lewis, 1996). Less familiar are studies of PI in clinical prediction in relation to groups defined by culture, language, gender, or age (e.g., Krakowski & Czobor, 2004). Here the criterion may be a binary diagnosis, for example. Also in a clinical context, investigations of how treatment effects vary depending on baseline status are really investigations of PI, where the outcome measure is Y, the treatment groups define $\mathbf{V}$, and $\mathbf{Z}$ is the baseline measure (Brown & Liao, 1999). Within an analysis of covariance in this setting, PI would imply no treatment effect and no treatment by baseline interaction.

How Are MI and PI Related?

While much has been written about each form of invariance or special cases of the two, very little has appeared on their relationship, particularly in the most general case. One special case that has received attention is when: (1) $\mathbf{X}$ fits a common factor model with $p = 1$; and (2) Y and $\mathbf{Z}$ are related via a linear regression (Birnbaum, 1979; Humphreys, 1986; Linn, 1984; Linn & Werts, 1971; Millsap, 1995, 1997, 1998). The general question of interest is whether the existence of one form of invariance implies that the other form of invariance must hold. Further, if there is no logical equivalence between the two forms of invariance, under what conditions are they consistent or inconsistent? One motivation for such questions is practical. PI is more easily investigated than MI because no latent variables are involved. If PI has implications for

MI, this fact greatly simplifies the investigation of MI. A second motivation lies in how we should interpret the many existing studies of PI in areas such as industrial/organizational psychology. In applied settings, one can ask whether the existence of MI should matter if PI is found to hold. To make this concrete, if the prediction of a job performance measure Y given a selection test score $\mathbf{Z}$ is the same regardless of $\mathbf{V}$, why should we care about MI? After all, no systematic under- or over-prediction of Y is happening in this case in relation to $\mathbf{Z}$ and $\mathbf{V}$. As will be illustrated below, in this purely practical context we should still care about MI because violations of MI can lead to systematic inaccuracy in selection, even though PI holds.

The General Case

Some results are available on the relationship between MI and PI for the general case in which no particular parametric model is assumed either for the relation of $\mathbf{X}$ to $\mathbf{W}$, or the relation of Y to $\mathbf{Z}$ (Meredith & Millsap, 1992; Millsap & Meredith, 1992). These results are useful in setting broad conditions for the consistency or inconsistency between the two forms of invariance.

The first result identifies one set of conditions under which MI and PI are consistent. Suppose that

$$P(Y|\mathbf{Z},\mathbf{W}) = P(Y|\mathbf{Z}) \tag{3}$$

for all $Y, \mathbf{Z}, \mathbf{W}$. Then if MI holds for $\mathbf{X}$ in relation to $\mathbf{W}$ and $\mathbf{V}$ (i.e., (1) holds), it must be true that (2) holds, or that PI holds for $\mathbf{Z}$ in relation to Y and $\mathbf{V}$. The condition in (3) was denoted “Bayes sufficiency” for $\mathbf{Z}$ in relation to Y and $\mathbf{W}$ in Meredith and Millsap (1992; see also Lehman, 1986). It is a sufficiency condition in the sense that (3) implies that all of the information in $\mathbf{W}$ that is relevant to Y is contained in $\mathbf{Z}$. Once we condition on $\mathbf{Z}$, Y and $\mathbf{W}$ are conditionally independent. The classic example of (3) exists when $\mathbf{Z}$ is the sum of a set of binary test items that fit a Rasch model, Y is a single item score variable that is part of the sum in $\mathbf{Z}$, and $\mathbf{W}$ is the latent variable underlying all test items. Then (3) is known to hold under the structure of the Rasch model. The sufficiency principle is relevant for the Mantel–Haenszel method of DIF detection, and explains why the item being studied for DIF must be included in the sum $\mathbf{Z}$ (Zwick, 1990).

Unfortunately, Bayes sufficiency is violated under some conditions that are often assumed to hold in predictive studies. For example, suppose that Y and $\mathbf{Z}$ are disjoint sets of variables (i.e., Y is not contained in $\mathbf{Z}$) and that

$$P(Y|\mathbf{Z},\mathbf{W},\mathbf{V}) = P(Y|\mathbf{W},\mathbf{V}) \tag{4}$$

for all $Y, \mathbf{Z}, \mathbf{W}, \mathbf{V}$. Then it can be shown that if MI in (1) holds, PI in (2) cannot hold generally. PI must be violated for some combination of Y, $\mathbf{Z}$, and $\mathbf{W}$. The condition in (4) is typically denoted as “local independence” of Y and $\mathbf{Z}$ given $\mathbf{W}$ and $\mathbf{V}$. Local independence is a condition that is invoked in almost all latent variable models for Y and $\mathbf{Z}$. For example, if $\mathbf{X}$ fits a common factor model with common factors $\mathbf{W}$ and with multivariate normality for ($\mathbf{X}$, $\mathbf{W}$) within populations defined by $\mathbf{V}$, then (4) holds. In prediction applications, the assumption that Y and $\mathbf{Z}$ fit a latent variable model under local independence is more natural than the Bayes sufficiency condition in (3). Local independence is generally inconsistent with Bayes sufficiency, and under local independence in (4), MI is inconsistent with PI.

The above two results involving Bayes sufficiency and local independence are not the only results that may be achieved for the general, nonparametric case. Under some conditions, local independence for $\mathbf{Z}$ and Y is compatible with PI, but these conditions imply that MI fails to hold for $\mathbf{Z}$, or that $\mathbf{Z}$ is biased as a measure of $\mathbf{W}$. These nonparametric results are useful by virtue of their generality, but applied researchers ordinarily work within specific parametric modeling traditions. We turn now to these special cases, and their implications for the relationships between the two forms of invariance.

The Linear Case: Factor Analysis and Regression

As noted earlier, one specific case in which the relation between PI and MI has received attention is the case of: (1) $\mathbf{X}$ fits a common factor model in relation to the single common factor $\mathbf{W}$; and (2) the regression of Y on $\mathbf{Z}$ is linear. For the factor model, let $\mathbf{X}_k$ be the $q \times 1$ vector of measured variables within the kth population as defined by $\mathbf{V}$, $k = 1, \dots, K$. We assume that

$$\mathbf{X}_k = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}_k W_k + \mathbf{u}_k, \tag{5}$$

where $\boldsymbol{\tau}_k$ is a $q \times 1$ vector of measurement intercept parameters, $\boldsymbol{\Lambda}_k$ is a $q \times 1$ vector of factor loading parameters, W_k is a scalar factor score random variable, and $\mathbf{u}_k$ is a $q \times 1$ vector of unique factor random variables. We assume that, for all k,

$$\mathrm{Cov}(W_k, \mathbf{u}_k) = \mathbf{0}, \qquad \mathrm{E}(W_k) = \kappa_k, \qquad \mathrm{Var}(W_k) = \phi_k, \qquad \mathrm{E}(\mathbf{u}_k) = \mathbf{0}, \qquad \mathrm{Cov}(\mathbf{u}_k) = \boldsymbol{\Theta}_k, \tag{6}$$

with $\boldsymbol{\Theta}_k$ being a $q \times q$ diagonal matrix. Given the partitioning $\mathbf{X}'_k = (Y_k, \mathbf{Z}_k)$, we can define an analogous partitioning for $\boldsymbol{\tau}_k$ and $\boldsymbol{\Lambda}_k$ as

$$\boldsymbol{\tau}_k = \begin{bmatrix} \tau_{yk} \\ \boldsymbol{\tau}_{zk} \end{bmatrix}, \qquad \boldsymbol{\Lambda}_k = \begin{bmatrix} \lambda_{yk} \\ \boldsymbol{\Lambda}_{zk} \end{bmatrix}. \tag{7}$$

Note that under this factor model

$$\mathrm{E}(\mathbf{X}_k | W_k) = \boldsymbol{\tau}_k + \boldsymbol{\Lambda}_k W_k, \qquad \mathrm{Cov}(\mathbf{X}_k | W_k) = \boldsymbol{\Theta}_k. \tag{8}$$

MI therefore implies that for $\boldsymbol{\Delta}_k = (\boldsymbol{\tau}_k, \boldsymbol{\Lambda}_k \boldsymbol{\Theta}_k)$ we must have $\boldsymbol{\Delta}_k = \boldsymbol{\Delta}$ for all k; there are no group differences in the parameter set $\boldsymbol{\Delta}$. Invariance in $\boldsymbol{\Delta}_k$ is denoted as *strict factorial invariance* in the literature (Meredith, 1993). Strict factorial invariance by itself need not imply MI unless certain distributional assumptions are met. Conditions of factorial invariance that are weaker than strict invariance are often of interest in applications. *Metric invariance* (Horn & McArdle, 1992; Thurstone, 1947) or *pattern invariance* (Millsap, 1995) is often put forth as a minimum requirement for invariance: $\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}$ for all k. Finally, note that invariance in the parameters (κ_k, ϕ_k) that determine the distribution of W_k is not required for factorial invariance.

Turning to the relation between Y and $\mathbf{Z}$, it is assumed that within the kth group defined by $\mathbf{V}$,

$$Y_k = \beta_{0k} + \boldsymbol{\beta}'_{1k} \mathbf{Z}_k + e_k, \tag{9}$$

with β_{0k} being the regression intercept, $\boldsymbol{\beta}_{1k}$ the $p \times 1$ vector of regression coefficients, and e_k being a residual random variable. It is assumed that

$$\mathrm{E}(Y_k | \mathbf{Z}_k) = \beta_{0k} + \boldsymbol{\beta}'_{1k} \mathbf{Z}_k, \qquad \mathrm{Var}(Y_k | \mathbf{Z}_k) = \sigma^2_{ek}. \tag{10}$$

PI implies that the parameters $\boldsymbol{\Omega}_k = (\beta_{0k}, \boldsymbol{\beta}_{1k}, \sigma_{ek})$ are invariant: $\boldsymbol{\Omega}_k = \boldsymbol{\Omega}$ for all k. In practice however, interest focuses chiefly on the intercept and regression coefficient parameters. For example, the condition $\boldsymbol{\beta}_{1k} = \boldsymbol{\beta}_1$ for all k is denoted *slope invariance* (Millsap, 1995). If slope invariance holds and the regression intercepts are also invariant ($\beta_{0k} = \beta_0$), this condition will be denoted *strong regression invariance*, by analogy with strong factorial invariance (Meredith, 1993).

Having defined the parametric models for both measurement and prediction involving $\mathbf{X}_k$, we can return to the issue of consistency between MI and PI. For example, given strict factorial invariance, must we also have strong regression invariance? Conversely, given strong regression invariance, must we also have strict factorial invariance? Fortunately, we have fairly complete answers to these questions. Here we will examine two situations in which the two forms of invariance are inconsistent.

Case One: Strict Factorial Invariance Without Strong Regression Invariance

Given the above model definitions, it has long been known that strict factorial invariance may hold and yet strong regression invariance may fail to hold (Birnbaum, 1979; Humphreys, 1986; Linn, 1984). In this situation, the slope parameters are invariant ($\boldsymbol{\beta}_{1k} = \boldsymbol{\beta}_1$ for all k) but the regression intercept parameters vary across groups. Millsap (1998) gave several theorems that applied to this case. It is assumed that not only does strict factorial invariance hold, but invariance in the common factor variance must hold as well. Under these assumptions, for $k = 1, 2$ for example, we have $\beta_{01} > \beta_{02}$ if and only if $\kappa_1 > \kappa_2$. The direction of the intercept difference is determined by the factor mean difference: the group with the larger factor mean will have the larger intercept. If a common regression line is imposed on the two groups, the group with the higher intercept will show systematic underprediction via the common line. In many applications, this group is the majority or reference group. This scenario matches empirical findings in studies of PI in which the only group difference in the regression lies in the intercepts, and that difference appears to favor the lower scoring group (i.e., the group with the lower factor mean) (Gottfredson, 1994; Hartigan & Wigdor, 1989; Jensen, 1980; Sackett & Wilk, 1994; Schmidt, Pearlman, & Hunter, 1980). These results, combined with the model that implies them, have been used to illustrate why group differences in regression intercepts need not indicate any problem in the predictor measure (Z), given that the results occur under strict factorial invariance (e.g., Jensen, 1980).

Several points should be noted with regard to this Case One scenario however. First, the model that is assumed to be responsible for the results is highly restrictive: both strict factorial invariance and invariance in the common factor variance are needed. The model has some strong implications. It implies that the covariance matrix for $\mathbf{X}_k$ is identical across groups, as is the correlation matrix. Furthermore, group differences in the observed means on $\mathbf{X}_k$ should all be in the same direction, apart from sampling error. In other words, across the vector of measured variables $\mathbf{X}_k$, one group should consistently have higher means. Second, the factor model that underlies the results is testable with ordinary structural equation modeling software, even when only a single predictor is used ($p = 1$). The factor model has $df = p(p+2)$, and so $p = 1$ implies $df = 3$. The test of fit is described and illustrated using real data in Millsap (1998). Rejection of the model would suggest that some other explanation for the group difference in regression intercepts must be sought. Although this test has been available for almost a decade, and many studies have examined strong regression invariance during this period, no further published examples of its use have appeared. In other words, while this restrictive model is often assumed to hold and to explain empirical findings, the model itself is never subjected to any empirical tests.

Case Two: Strong Regression Invariance Without Strict Factorial Invariance

In contrast to the previous example, it is possible to have strong regression invariance while also using a predictor that is systematically biased in the measurement sense. To illustrate this case, assume that $p = 1$ and that a single factor W underlies Y and Z. Within the factor model, we will assume that the factor loadings, unique factor variances, common factor variance, and criterion measurement intercept are all invariant:

$$\boldsymbol{\Lambda}_k = \boldsymbol{\Lambda}, \qquad \boldsymbol{\Theta}_k = \boldsymbol{\Theta}, \qquad \phi_k = \phi, \qquad \tau_{yk} = \tau_y, \tag{11}$$

for $k = 1, 2$. No restrictions are placed on the common factor means κ_k or on the predictor measurement intercept τ_{zk}. If the predictor intercept is invariant, strict factorial invariance would hold for both Y and Z. By permitting the predictor measurement intercept to vary, we permit the violation of MI for the predictor Z. It can be shown that these assumptions imply that the slopes for the regression of Y on Z are invariant, but the regression intercepts may differ (Millsap, 1998). The regression intercepts can be written

$$\beta_{0k} = [\tau_y - \beta_1 \tau_{zk}] + [\lambda_y - \beta_1 \lambda_z]\kappa_k \tag{12}$$

for $k = 1, 2$. From the above expression, it is clear that for any pair of factor means (κ_1, κ_2), it is always possible to find a pair of predictor measurement intercepts so that the regression intercepts are invariant. The group difference in the regression intercepts can be written as

$$\beta_{01} - \beta_{02} = [\lambda_y - \beta_1 \lambda_z](\kappa_1 - \kappa_2) - \beta_1[\tau_{z1} - \tau_{z2}]. \tag{13}$$

Millsap (1998) showed that the first term in brackets is positive as long as $\Theta_z > 0$ and $\lambda_y > 0$. These requirements are ordinarily met in applications. Suppose then for the sake of argument that $\kappa_1 > \kappa_2$. Then as this factor mean difference gets larger, the difference in regression intercepts also grows. Now consider the measurement intercepts, and suppose that $\tau_{z1} > \tau_{z2}$, or that the predictor measure is biased in favor of the higher scoring group. Given the usual case in which $\beta_1 > 0$, it is clear that the bias in the predictor operates to reduce the difference in the regression intercepts. In fact, the bias in the predictor can shrink the regression intercept difference to zero, resulting in strong regression invariance without MI. The apparent strong invariance in the regression will mask the measurement bias in the predictor measure.

Unlike the first case in which factorial invariance held but strong regression invariance was violated, this second case has received no attention in the literature. The factor model underlying this second case is not testable with $p = 1$, but is testable with $p > 1$. Millsap (1998) illustrated the test in real data for $p = 2$. In testing the fit of these models, a logical sequence would be to test for strict factorial invariance first. If that model is rejected, the next model is the one underlying the case just illustrated. No papers reporting the use of this test procedure have appeared in the literature since Millsap (1998), although the test is easily done with structural equation modeling software that performs multiple-group CFA.

Example. To illustrate the two cases just described, suppose that $p = 1$ and that we have the following invariant factor loadings, common factor variance, and unique factor variances:

$$\mathbf{\Lambda} = \begin{bmatrix} .4 \\ .6 \end{bmatrix}, \qquad \phi = 1, \qquad \mathbf{\Theta} = \begin{bmatrix} .24 & 0 \\ 0 & .22 \end{bmatrix}. \tag{14}$$

Suppose also that the invariant criterion measurement intercept is $\tau_y = .50$. Under this parametrization, the correlation between Z and Y is .50 in both groups, and the invariant regression slope is $\beta_1 = .41$. Without loss of generality, we can fix $\kappa_2 = 0$ and then manipulate the remaining factor model parameters $(\kappa_1, \tau_{z1}, \tau_{z2})$ to study their impact on the regression intercepts.

We begin with the first case of strict factorial invariance. We set $\kappa_1 = 1.0$ so that group one has the larger factor mean. We also set $\tau_{21} = \tau_{22} = .50$. Theorem 2 in Millsap (1998) establishes that group one will have the larger regression intercept, and this is in fact what is found:

$$\beta_{01} = .449, \qquad \beta_{02} = .295. \tag{15}$$

The parameter values for Y and Z in the two groups are

$$\boldsymbol{\mu}_1 = \begin{bmatrix} .9 \\ 1.1 \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix} .40 & .24 \\ .24 & .58 \end{bmatrix}. \tag{16}$$

In this case, the use of a common regression line to model both groups will lead to systematic underprediction for members of group one, and overprediction for members of group two. No violation of factorial invariance exists however. The prediction error pattern is produced by the unique factor variance in the predictor. If there is no unique variance in the predictor, the regression intercept difference vanishes. One way to reduce the unique factor variance is to improve the reliability of the predictor, but the unique variance includes more than measurement error.

Turning now to the second case of strong regression invariance, suppose that we keep the factor means as above, but set the predictor measurement intercepts to

$$\tau_{21} = .500, \qquad \tau_{22} = .124. \tag{17}$$

The predictor measurement intercept difference favors members of group one: at any given factor score, members of group one are expected to score higher on Z than members of group two. In other words, the predictor is biased in favor of group one. The regression lines will be identical however, with common intercepts and slopes:

$$\beta_{01} = \beta_{02} = .449, \qquad \beta_{11} = \beta_{12} = .410. \tag{18}$$

The parameter values for the Y and Z measures are

$$\boldsymbol{\mu}_1 = \begin{bmatrix} .9 \\ 1.1 \end{bmatrix}, \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} .5 \\ .124 \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} .40 & .24 \\ .24 & .58 \end{bmatrix}. \tag{19}$$

We can note here that the group difference in means on the predictor is $1.1 - .124 = .976$, and the difference in measurement intercepts is .376. Hence about 38% of the mean difference is accounted for by the measurement bias favoring group one. This fact is not apparent in the regression lines however.

Selection Accuracy Implications

In addition to these parametric results, we can investigate what would happen if Z was used to select individuals under each of the two cases. To answer this question, some distributional assumptions are needed. We will assume that $\mathbf{X} = (Y, Z)$ is bivariate normal within each of the two groups, with the relevant parameters being given in (16) for the first case, and (19) for the second case. We also assume that in the combined population that is a mixture of groups one and two, selection proceeds as a simple top-down selection based on Z. Note that the combined population is a mixture of two bivariate normal distributions, and does not itself have a bivariate normal distribution. Millsap and Kwok (2004) describe an algorithm for calculating cutpoints and quadrant probabilities in the mixture of two bivariate normal distributions under selection. This algorithm is used to obtain the results reported here.

We first consider how selection under different top-down selection percentages in the combined mixture affects the proportions selected from groups one and two. Table 1 provides this information. In this table, the first column lists the overall percent selected (selection ratio) in the combined population (e.g., 5% is top five percent). The remaining columns give the resulting proportions selected in the separate groups, along with the percentage of group two members among those selected, all done separately by case. The general trend is that the proportions selected under Case Two for the lower scoring group two are smaller in comparison to Case One. This trend holds regardless of the selection percentage in the combined population. This result is the direct effect of the smaller latent intercepts in group two in Case Two, and the resulting smaller mean on Z. The reduced mean on Z in group two reduces the relative proportion selected from that group in any top-down selection based on Z.

Next, we consider the accuracy of selection under the two cases. In addition to the parameters for the measured variables (Y, Z) in (16) and (19), we also know the parameters that determine the factor score distribution in each case, under an assumption of normality for the factor scores.

TABLE 1.
Proportions selected in each group under two cases.

	Case One			Case Two		
% Selected	G1	G2	%G2	G1	G2	%G2
5%	.0847	.0153	15.30	.0952	.0048	4.80
10%	.1620	.0380	19.00	.1853	.0147	7.35
15%	.3691	.1309	26.18	.4282	.0718	14.36

TABLE 2.
Accuracy measures under 25% selection ratio.

	Sensitivity		PPV	
	G1	G2	G1	G2
Case One	.7097	.6335	.7626	.5004
Case Two	.7785	.4381	.7209	.6309

If we assume bivariate normality for (W, Z) in each group, we can compare Case One and Case Two to evaluate the impact of the measurement intercept difference on the *sensitivity* and *positive predictive value* associated with selection based on Z. Millsap and Kwok (2004) described the logic underlying these calculations in this selection context. Consider the bivariate distribution of (W, Z) and the parameters governing this distribution under bivariate normality. We can regard W as the quantity that defines the examinee's actual standing on the latent variable measured by Z. We will select examinees using the fallible measure Z because W is unknown. If we select the top 25% of the examinees based on Z in the combined group one and group two, we thereby identify the subset of the combined population that is the "selected" subset. We can find the needed cutpoint on Z that would locate the top 25% in the combined population. Similarly, we can identify a cutpoint on W within the combined population that would mark off the top 25% in the distribution of W. The two cutpoints effectively divide the bivariate distribution of (W, Z) into four quadrants, which can be labeled as *true positives*, *false positives*, *true negatives*, and *false negatives*. This division into four quadrants can be done separately for groups one and two, using the same cutpoints in both groups. Once the relative proportions in the four quadrants are known, we can calculate sensitivity and positive predictive value (PPV) separately by group. Sensitivity in this context is defined as the conditional probability of selection given that the examinee is above the cutpoint on W (i.e., should be selected). PPV is the conditional probability that an examinee is above the cutpoint on W, given that they are selected. All of these calculations can be done once for the model under Case One and once for the model under Case Two.

We will examine results for the 25% selection ratio only. Table 2 gives the sensitivity and PPV results for Cases One and Two separately. For sensitivity, the sensitivity shows an increase from Case One to Case Two for group one, and a sharp decrease in sensitivity from Case One to Case Two in group two. It appears that the group difference in measurement intercepts in Case Two is reducing the sensitivity of the measure Z. In this sense, the violation of invariance in Case Two is reducing the accuracy of Z as a selection instrument. An opposite trend is found for PPV, although not as dramatic. The increase in PPV for group two in Case Two comes about due to the increased selectivity in that group, as shown in Table 1.

These selection results are clearly only one set of results among many scenarios that could be considered. Some of these scenarios would only reveal trivial inaccuracies while others would reveal more dramatic effects. The main point to be conveyed is that an exclusive focus on PI ignores the larger context that must be considered for a full evaluation of selection accuracy. The above analysis of selection accuracy could be undertaken for any empirical application without difficulty, using parameter values and distributional assumptions deemed appropriate for the data at hand. No such applications will be found in the current literature on selection in applied psychology however.

Discussion

Having described some of the main results on the relationship between MI and PI, as well as some numerical examples, we can return to the original question of why these results are not more widely known and used. In her comments on Borsboom (2006), Clark (2006) argued that the current literature on this topic is written in unnecessarily technical language, preventing most psychologists from understanding the material. Borsboom (2006) also commented on the mismatch between the often highly technical papers that report psychometric advances, and the generally weak mathematical training of most psychologists. It is certainly true that psychometricians (myself included) could do more to make their findings understood by psychologists, but I don't believe that this problem fully explains why the work on invariance is not more widely used. The level of mathematics required to understand the conclusions offered by this work is not high; one does not need to follow the details of the proofs to understand their conclusions. Furthermore, it is not necessary for all or even most psychologists to understand all of the mathematics before the conclusions begin to influence practice. The dissemination of a methodological advance is usually led by a relatively small group of applied researchers who do have the background needed to understand the methodology. There are a sufficient number of such researchers in fields such as industrial/organizational psychology. Structural equation modeling is a good example of a technical advance that, while initially available only in technical journals, was eventually disseminated widely and is now part of the standard curriculum for graduate students in psychology.

A larger barrier to the wider dissemination of recent work on invariance is the conviction that questions about bias in measurement have already been settled scientifically. This conviction seems to be widely held, at least among many influential psychologists who work in applied areas. Hunter and Schmidt (2000) conclude that "We trust the literature on test bias, and we know that the literature on item bias is unsound from a technical standpoint. Thus, on the basis of our review of this literature, it is our current working hypothesis that there is no item bias" (p. 157). The literature on "test bias" that is noted here is the literature on predictive invariance; the literature on "item bias" is the DIF literature. Sackett, Schmitt, Ellingston, and Kabin (2001) note the "extensive body of research" showing that "standardized tests do not underpredict the performance of minority group members" (p. 303). In a report by a Task Force created by the Board of Scientific Affairs of the American Psychological Association, Neisser, Boodoo, Bourchard, Boykin, Brody, Ceci, Halpern, Loehlin, Perloff, Sternberg, and Urbina (1996) state that because intelligence tests are used as "predictors," the relevant question is "whether the tests have a predictive bias against Blacks." They conclude that no such bias exists: "the actual regression lines for Blacks do not lie above those for Whites" (p. 93). It appears that a scientific consensus has already been forged on the issue of bias.

What is the foundation of this consensus? It is built upon the large number of predictive studies that show either invariance in regressions and correlations, or relatively small differences in regression lines and correlations. In the case of correlation differences, large meta-analytic studies in the employment sector have not revealed substantial differences in correlations by ethnicity (Schmidt & Hunter, 1998). Empirical studies of MI have contributed little to this consensus. In fact, studies of MI are viewed by some as unnecessary in light of the predictive work already completed (Hunter & Schmidt, 2000). Given the scientific consensus, it is clear why the work on invariance in measurement and prediction has made few inroads. This work undercuts the foundation for the consensus by showing that bias in measurement is entirely consistent with little or no predictive bias. A test that shows predictive invariance, yet is biased from the measurement standpoint, can produce systematic selection errors when used in selection, as illustrated earlier.

The future of work on invariance in measurement and prediction cannot be known, but it is probably best to take a long view. The current consensus on bias was formed based on evidence that spans decades. It may be decades before the work on invariance has comparable impact. The invariance example contains some lessons for psychometricians who are concerned about the dissemination of psychometric advances throughout psychology. First, psychometricians should encourage psychologists to think integratively about measurement. Prediction and measurement are separable theoretically, but when tests are used in selection, the exclusive focus on prediction tends to obscure any concerns about the tests as measurements. It is extremely rare to find an empirical PI study that also examines MI empirically, using the same data. No particular barrier exists to conducting such studies however. Second, psychometricians should encourage critical thinking about measurement among psychologists. As noted earlier, the models that underlie some of the inconsistency between MI and PI are testable in most cases, but these tests are underutilized. The selection accuracy analyses that were illustrated earlier could be implemented in any bivariate prediction study, and would reveal whether the impact of any measurement bias is trivial or substantial. Finally, psychometricians need to present their research in terms that can be appreciated by the larger psychological community. Journals such as *Psychometrika* have an important role to play in this effort. Important technical work will always be valued, but work that can ultimately influence practice must be translated into language that practitioners can appreciate.

References

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*, 67–91.

Ahmavaara, Y. (1954). The mathematical theory of factorial invariance under selection. *Psychometrika*, *19*, 27–38.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education Joint Committee on Standards for Educational and Psychological Testing (1999). *Standards for educational and psychological testing*. Washington: AERA.

Birnbaum, M.H. (1979). Procedures for the detection and correction of salary inequities. In T.R. Pezzullo & B.E. Brittingham (Eds.), *Salary equity* (pp. 121–144). Lexington: Lexington Books.

Bloxom, B. (1972). Alternative approaches to factorial invariance. *Psychometrika*, *37*, 425–440.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.

Bridgeman, M.H., & Lewis, C. (1996). Gender differences in college mathematics grades and SAT-M scores: A reanalysis of Wainer and Steinberg. *Journal of Educational Measurement*, *33*, 257–270.

Brown, C.H., & Liao, J. (1999). Principles for designing randomized preventive trials in mental health: An emerging developmental epidemiology paradigm. *American Journal of Community Psychology*, *27*, 673–710.

Byrne, B.M. (1994). Testing for factorial validity, replication, and invariance of a measuring instrument: A paradigmatic application based on the Maslach Burnout Inventory. *Multivariate Behavioral Research*, *29*, 289–311.

Clark, L.E. (2006). When a psychometric advance falls in the forest. *Psychometrika*, *71*, 447–450.

Cleary, T.A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Drasgow, F., & Probst, T.A. (2004). The psychometrics of adaptation: Evaluating measurement equivalence across languages and cultures. In R.K. Hambleton, P.F. Merenda, & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 265–296). Hillsdale: Lawrence Erlbaum.

Gottfredson, L.S. (1994). The science and politics of race-norming. *American Psychologist*, *49*, 955–963.

Hambleton, R.K., Merenda, P.F., & Spielberger, C.D. (2006). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale: Lawrence Erlbaum.

Hartigan, J.A., & Wigdor, A.K. (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington: National Academy Press.

Hofer, S.M., Horn, J.L., & Eber, H.W. (1997). A robust five-factor structure of the 16PF: Strong evidence from independent rotation and confirmatory factorial invariance procedures. *Personality and Individual Differences*, *23*, 247–269.

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

Humphreys, L.G. (1952). Individual differences. *Annual Review of Psychology*, *3*, 131–150.

Humphreys, L.G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Psychological Bulletin*, *71*, 327–333.

Hunter, J.E., & Schmidt, F.L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law*, *6*, 151–158.

Jensen, A.R. (1980). *Bias in mental testing*. New York: Free Press.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent models* (pp. 263–275). New York: Plenum.

Krakowski, M., & Czobor, P. (2004). Gender differences in violent behaviors: Relationship to clinical symptoms and psychosocial factors. *American Journal of Psychiatry*, *161*, 459–465.

Lehmann, E.L. (1986). *Testing statistical hypotheses*. New York: Wiley.

Linn, R.L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement*, *21*, 33–47.

Linn, R.L., & Werts, C.E. (1971). Considerations for studies of test bias. *Journal of Educational Measurement*, *8*, 1–4.

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

Mellenbergh, G.J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Meredith, W. (1964a). Notes on factorial invariance. *Psychometrika*, *29*, 177–185.

Meredith, W. (1964b). Rotation to achieve factorial invariance. *Psychometrika*, *29*, 187–206.

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

Meredith, W., & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289–311.

Millsap, R.E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*, 577–605.

Millsap, R.E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Millsap, R.E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research*, *33*, 403–424.

Millsap, R.E., & Hartog, S.B. (1988). Alpha, beta, and gamma change in evaluation research: A structural equation approach. *Journal of Applied Psychology*, *73*, 574–584.

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

Millsap, R.E., & Meredith, W. (1992). Inferential conditions in the statistical detection of measurement bias. *Applied Psychological Measurement*, *16*, 389–402.

Neisser, U., Boodoo, G., Bourchard, T.J., Boykin, A.W., Brody, N., Ceci, S.J., Halpern, D.F., Loehlin, J.C., Perloff, R., Sternberg, R.J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist*, *51*, 77–101.

Pentz, M.A., & Chou, C. (1994). Measurement invariance in longitudinal clinical research assuming change from development and intervention. *Journal of Consulting and Clinical Psychology*, *62*, 450–462.

Potthoff, R.F. (1966). *Statistical aspects of the problem of biases in psychological tests* (Institute of Statistics Mimeo Series No. 479). Chapel Hill, NC: Department of Statistics, University of North Carolina.

Riordan, C.R., Richardson, H.A., Schaffer, B.S., & Vandenberg, R.J. (2001). Alpha, beta, and gamma change: A review of past research with recommendations for new directions. In L.L. Neider & C. Schriesheim (Eds.), *Equivalence in measurement* (pp. 51–98). Greenwich: Information Age Publishing.

Sackett, P.R., & Wilk, S.L. (1994). Within-group norming and other forms of score adjustment in preemployment testing. *American Psychologist*, *49*, 929–954.

Sackett, P.R., Schmitt, N., Ellington, J.E., & Kabin, M.B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative action world. *American Psychologist*, *56*, 302–318.

Schmidt, F.L., & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of over 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.

Schmidt, F.L., Pearlman, K., & Hunter, J.E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology*, *33*, 705–724.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*, 159–194.

Society for Industrial/Organizational Psychology (2003). *Principles for the application and use of personnel selection procedures*. Bowling Green: Society for Industrial Organizational Psychology.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293–325.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale: Lawrence Erlbaum.

Thomson, G.H., & Lederman, W. (1939). The influence of multivariate selection on the factorial analysis of ability. *British Journal of Psychology*, *29*, 288–305.

Thurstone, L.L. (1947). *Multiple factor analysis*. Chicago: University of Chicago Press.

Zwick, R. (1990). When do item response function and Mantel–Haenszel definitions of differential item functioning coincide? *Journal of Educational Statistics*, *15*, 185–197.

Manuscript received 30 AUG 2007
Final version received 30 AUG 2007
Published Online Date: 16 NOV 2007