A Study of Statistical Power and Type I Errors in Testing a Factor Analytic Model for Group Differences in Regression Intercepts

by

Margarita Olivera Aguilar

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved October 2010 by the
Graduate Supervisory Committee:

Roger E. Millsap, Chair
Leona S. Aiken
Craig Enders

ARIZONA STATE UNIVERSITY

December 2010

ABSTRACT

In the past, it has been assumed that measurement and predictive invariance are consistent so that if one form of invariance holds the other form should also hold. However, some studies have proven that both forms of invariance only hold under certain conditions such as factorial invariance and invariance in the common factor variances. The present research examined Type I errors and the statistical power of a method that detects violations to the factorial invariant model in the presence of group differences in regression intercepts, under different sample sizes and different number of predictors (one or two). Data were simulated under two models: in model A only differences in the factor means were allowed, while model B violated invariance. A factorial invariant model was fitted to the data. Type I errors were defined as the proportion of samples in which the hypothesis of invariance was incorrectly rejected, and statistical power was defined as the proportion of samples in which the hypothesis of factorial invariance was correctly rejected. In the case of one predictor, the results show that the chi-square statistic has low power to detect violations to the model. Unexpected and systematic results were obtained regarding the negative unique variance in the predictor. It is proposed that negative unique variance in the predictor can be used as indication of measurement bias instead of the chi-square fit statistic with sample sizes of 500 or more. The results of the two predictor case show larger power. In both cases Type I errors were as expected. The implications of the results and some suggestions for increasing the power of the method are provided.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Chapter 1

Introduction

Important decisions are made from the results of psychological tests, as in selecting people for a job, a graduate program, or a scholarship (Muchinsky, 1993; Sacket, Schmitt, Elligson, & Kabin, 2001). Because of the impact of these decisions, it is fundamental that the tests used show no bias against the different groups examined. In other words, the tests must be invariant in their psychometric functioning across the groups tested. The groups are usually defined in terms of demographic variables like gender and ethnic background. In psychological testing, two types of invariance have been studied: measurement invariance and predictive invariance.

Millsap (1998) proposed a confirmatory factor analytic model to test measurement and predictive invariance when group differences in regression intercepts exist. The model assumes full factorial invariance, and invariant common factor variances. If the model holds, group differences in regression intercepts can be explained in terms of differences in the common factor means. This model was tested in cases with one and two predictors with large sample sizes using real data.

The purpose of the present research is to study Type I errors and the statistical power for tests of the confirmatory factor analytic model proposed by Millsap (1998). Type I error rates and power are examined in simulated data with different sample sizes and with one and two predictors.

Chapter 2

Background Literature

Measurement invariance

Measuring individuals is a fundamental process when it is of interest to know their levels in a variable of interest, such as academic performance, personality, or attitudes. The variables of interest are usually latent, unobservable constructs that are assumed to cause the observed measures.

In order to make conclusions about the individuals, the test used must function equivalently across the groups studied; otherwise, the results of the tests have ambiguous interpretations (Borsboom, 2004). It is said that a test is measurement invariant if persons from different populations with identical values on the latent variables **W** of interest, have the same probability of obtaining a particular raw score at the item level or at the test level (Drasgow & Kanfer, 1985; Mellenbergh, 1989; Millsap, 2007).

Suppose that $\mathbf{X} = (Y, \mathbf{Z})$ is a $(p + 1) \times 1$ vector of observable random variables, where Y is a single criterion variable and **Z** is a set of p predictor variables. Further suppose that r latent variables underlie **X**, such that **W** is a $r \times 1$ vector of latent scores with $r < p + 1$. A classic and well known example can be found in the prediction of job performance. In a common factor model where a battery of cognitive tests is used as predictors and a measure of job performance is used as a criterion, it is often hypothesized that an underlying common factor,

Spearman's "g", would predominate (Gottfredson, 1988; Hunter, 1986; Ree, Earles & Teachout, 1994).

We will assume that two populations are being measured on **X** and that *V* is a variable that defines group membership. Usually the populations are defined in terms of demographic variables such as sex and ethnicity. In the employment example described before, differences between Whites and African-Americans in the prediction of job performance have been examined by comparing regression lines across groups. Measurement invariance means that there are no group differences in the relationship of a set of observed variables **X** to their underlying latent variables **W**. In the research addressing differences between White and African-Americans in the prediction of job performance some investigators have argued that any group differences observed are due to group differences in "g" (Jensen, 1992).

More formally, the definition of measurement invariance states that the relationship between **X** and **W** is independent of group membership (Mellenbergh, 1989; Millsap 1995, 2007), such that:

$$\Pr(\mathbf{X}|\mathbf{W} = w, V = v) = \Pr(\mathbf{X}|\mathbf{W} = w) \quad (1)$$

The above equation means that two persons with the same value in the latent variable **W** will have the same probability of achieving a particular score on **X** regardless of their group membership. If equation (1) does not hold then

measurement bias is said to exist. Under measurement bias two individuals with the same value in **W** will have probabilities of achieving scores on **X** that depend on group membership.

Different latent variable models have been proposed that describe the relationship between **W** and **X**, and hence, different approaches to testing measurement invariance exist. One of the models most widely used to describe the relationship between **W** and **X** is the common factor model. In this model **X** fits a common factor model with **W** being the common factors; factorial invariance is a form of measurement invariance in this model (Millsap, 1998).

Item response theory (IRT) provides another way to describe the relationship between **W** and **X**. IRT consist of a set of models that relates the probability of an item response to an examinee value on a latent variable, through a nonlinear monotonic function. Violations of measurement invariance in IRT are termed differential functioning in general; differential item functioning (DIF) refers to the study of invariance at the item level, and differential test functioning (DTF) is the study of measurement invariance at the test level (Stark, Chernyshenko, & Drasgow, 2006). In some cases, DIF is used to describe group differences in item properties that are evaluated without reference to a specific latent variable model (Holland & Wainer, 1993).

The focus of the present research was in measurement invariance under the common factor model.

The Common Factor Model

The most widely used model in studies of measurement invariance is the common factor model. In the case in which there is only one factor the model can be expressed as:

$$\mathbf{X}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i W_i + \mathbf{u}_i \tag{2}$$

where $\boldsymbol{\Lambda}_i$ is a (p + 1) x 1 factor pattern matrix for group i, W_i is a scalar common factor score, $\boldsymbol{\tau}_i$ is a (p + 1) x 1 latent intercept vector, $\mathbf{u}_i$ is the (p + 1) x 1 vector of unique factor scores.

If $\mathbf{X} = (Y, \mathbf{Z})$, the factor pattern and the latent intercepts can be partitioned as:

$$\boldsymbol{\Lambda}_i = \begin{Vmatrix} \lambda_{yi} \\ \boldsymbol{\Lambda}_{zi} \end{Vmatrix} \qquad \tau_i = \begin{Vmatrix} \tau_{yi} \\ \tau_{zi} \end{Vmatrix} \tag{3}$$

In this partitioning $\boldsymbol{\Lambda}_{zi}$ is a p x1 vector of factor loadings of the predictors on the common factor, λ_{yi} is a scalar containing the loading of the criterion on the common factor, τ_{zi} is a p x 1 latent intercepts of the predictors, and τ_{yi} is a scalar latent intercept of the criterion.

Under the factor model, the expected value of $\mathbf{X}_i$ given W_i and the conditional covariance of $\mathbf{X}_i$, can be expressed as:

$$E(\mathbf{X}|W = w) = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i w \qquad Cov(\mathbf{X}|W = w) = \boldsymbol{\Theta}_i \tag{4}$$

here $\boldsymbol{\Theta}_i$ is a diagonal matrix of unique variances for group *i*. For measurement invariance to exist, no differences between the groups should be found in $\boldsymbol{\tau}_i$, $\boldsymbol{\Lambda}_i$, and $\boldsymbol{\Theta}_i$. Invariance in the three parameters is known as complete factorial invariance or strict factorial invariance. Weaker forms of invariance are also possible; when invariance only holds for $\boldsymbol{\Lambda}$, this condition is called factor pattern invariance, metric invariance or weak factorial invariance; scalar or strong factorial invariance refers to invariance in $\boldsymbol{\Lambda}$ and $\boldsymbol{\tau}_i$ (Millsap, 2007; Widaman & Reise, 1997).

The parameters of the distribution of W_i are:

$$E(W_i) = \kappa_i \qquad \text{Var}(W_i) = \varphi_i \tag{5}$$

For measurement invariance to hold, it is not necessary to specify invariance in the parameters κ_i and φ_i. Under factorial invariance the populations of interest can differ in the distributions of W, and these differences are going to be reflected in the unconditional structure of $\mathbf{X}_i$:

$$E(\mathbf{X}) = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_i\kappa_i \tag{6}$$

$$\text{Cov}(\mathbf{X}) = \boldsymbol{\Lambda}_i\varphi_i\boldsymbol{\Lambda}_i` + \boldsymbol{\Theta}_i \tag{7}$$

If measurement invariance holds, the differences found in the observed variables **X**, are due to differences in the common factor *W* and not due to measurement bias.

Predictive invariance

In contrast to measurement invariance, predictive invariance is concerned with group differences in the relationship that holds only among observed measures. Cleary (1968) defined predictive bias as systematic errors in the prediction of a criterion for one of the groups studied. If there is predictive bias, the use of a single regression equation for describing two groups would lead to under-prediction for one group and over-prediction in the other.

Predictive invariance is more easily studied than measurement invariance because no latent variable model is needed. One of the domains in which this form of invariance has been of interest is in educational measurement. For example, Bridgeman and Lewis (1996) studied gender differences in the prediction of grades in college mathematics courses from SAT-M and from high school grade point average. In another study, Cleary (1968) studied differences in the prediction of college grades between black and white students.

In the present study focus is given to predictive invariance when the relationship among the measured variables is linear. A general definition for predictive invariance can be stated as:

$$\Pr(Y|\mathbf{Z}= z, V=v) = \Pr(Y|\mathbf{Z}= z) \quad (9)$$

Where Y is the criterion and $\mathbf{Z}$ is a set of predictor variables. If the relationship between Y and $\mathbf{Z}$ is linear then:

$$E(Y|\mathbf{Z} = z) = B_{0i} + \mathbf{B}_{1i}z \quad (10)$$

$$Var(Y|\mathbf{Z} = z) = \sigma^2_i \quad (11)$$

where σ^2_i is the residual variance in the ith group, B_{0i} is the regression intercept in the ith group, and $\mathbf{B}_{1i}$ is a p x 1 vector of regression slopes for the ith group. The definition of predictive invariance implies that the distribution of Y given $\mathbf{Z}$ is independent of group membership. Under predictive invariance, two individuals from different groups that have the same score in $\mathbf{Z}$ will have the same predicted score in Y.

Predictive invariance implies that the parameters B_0, $\mathbf{B}_1$, and σ^2 are invariant in the linear case. Slope invariance denotes invariance in $\mathbf{B}_1$ and regression intercept invariance refers to invariance in B_0 among the groups.

Relationship between measurement invariance and predictive invariance

Ideally, tests used for prediction will not only show predictive invariance but also measurement invariance. The question about the relationship between

both types of invariance emerges. If a test shows one type of invariance would the other type of invariance also hold?

As stated earlier, predictive invariance is more easily studied than measurement invariance because the former is only concerned with the relationships among observed variables and not with latent variables. So, if both types of invariance are related could one only study predictive invariance and still make conclusions about measurement invariance? Researchers usually assume that if predictive invariance holds the test is free of any bias (Sackett & Wilk, 1994).

There have been few attempts to study the conditions under which measurement and predictive invariance are consistent. In these studies the relationship between measurement invariance and predictive invariance has been examined mostly for situations in which **X** fits a common factor model with one latent variable $r = 1$, and for linear relationships among Y and **Z**.

For example, Millsap (1995, 1997) studied the relationship between factorial and slope invariance, and showed that the only scenario in which strict factorial invariance and slope invariance will both hold is when the groups have identical common factor variances. Also, the author examined the relationship between pattern and slope invariance and, as in the case of factorial and slope invariance, the conditions required for both forms of invariance to hold are stringent and often violated.

The relationship among regression intercept invariance and measurement invariance has also been studied (Millsap, 1998). Studies have shown that even when measurement invariance holds for a given data set, it is still possible that groups differ in their regression intercepts as a consequence of having fallible measures (Birnbaum, 1979; Linn, 1984; Millsap, 1998). One interpretation of this result is that the regression intercept differences are not representing measurement bias, but that they represent actual differences in the trait measured among the groups studied (Humphreys, 1986).

Millsap (1998) gave two theorems under which the regression intercept differences are not due to measurement bias. The conditions for these theorems are that slope invariance and factorial invariance must hold. Under these conditions the only parameter that can differ among the groups is κ_i, the factor mean. The theorems state that $\kappa_i > \kappa_2$ if and only if $\mathbf{B}_{01} > \mathbf{B}_{02}$. These theorems imply that under factorial invariance and slope invariance, the direction of the regression intercept differences must be the same as the direction in the difference in the common factor means. In other words, if measurement invariance holds, the group that has the larger factor means in $\boldsymbol{Z}$ and Y must also have the larger intercept in the regression for predicting Y.

One case in which the invariant common factor model will not hold is when one group has a larger mean in $\boldsymbol{Z}$ but a different group has the larger mean in Y, suggesting that the differences on the groups cannot be explained in terms of differences in common factor means. Another case in which the invariant

common factor model will not hold is when one group has larger common factor means in both ***Z*** and *Y*, and the regression for predicting *Y* shows that the larger intercept corresponds to the group with the larger mean. However, when conducting a reverse regression, that is, changing the roles of the predictor and the criterion so that now ***Z*** is being predicted from *Y*, the group with the larger common factor mean gets the smaller regression intercept (Birnbaum, 1979). If the invariant common factor model holds, the group with the larger means would have the larger intercepts in both forward and reverse regressions.

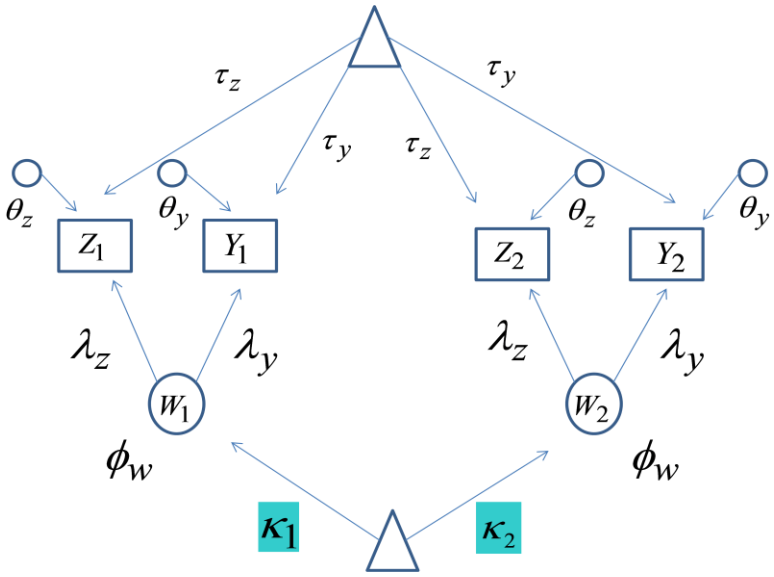
Millsap (1998) proposed a method for testing the conditions of the theorem using confirmatory factor analysis. The model has the requirements:

1) Factorial invariance, that is, invariance in the latent intercepts, the factor loadings, and the unique variances.
2) Invariance in the common factor variances φ must also hold under slope invariance, as established by the duality theorems (Millsap, 1995; 1997).

As a consequence of these requirements, slope invariance will hold. Slope invariance must hold since the study of intercept invariance is only meaningful under slope invariance.

The graphical representations of the models for the one and two predictor cases are given in Figure 1. The factor means are shadowed to indicate that those are the only parameters that are allowed to differ across the groups in the model.

One predictor case



Two predictor case

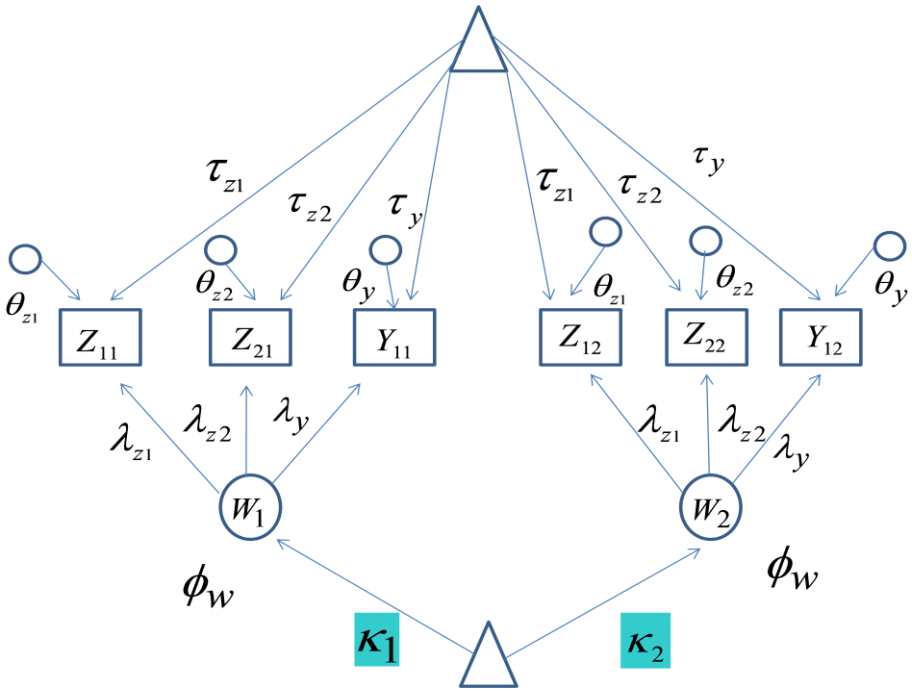


Figure 1. Measurement invariant models with group differences in the regression intercepts for one and two predictors. In each of the groups the latent variable *W* underlies the predictors ***Z*** and the criterion *Y*; all the parameters are restricted to be the same between groups except for the factor means **κ**.

The model also needs identification constraints such as fixing a predictor loading to one and fixing the factor mean to zero. The degrees of freedom for this model are p (p + 2).

Some real examples for testing the model using large sample sizes in the case of one predictor were provided in the study of Millsap (1998). In the first example, data from 12,424 examinees were analyzed. In this example it was shown that when measurement invariance holds, the group with the larger means in $\boldsymbol{Z}$ and Y also had the larger regression intercept in the forward and reverse regression.

In the second example data from 9,748 examinees were used to illustrate a case in which the invariant factor model did not fit the data. An inconsistent pattern with the invariant common factor model was found in the factor means; the group with the larger mean in $\boldsymbol{Z}$ had the smaller mean in Y. As a consequence, the group with the larger regression intercept in the forward regression was different from the group with the larger regression intercept in the reverse regression.

The third example with data from 68,766 examinees showed a case in which the invariant common factor model failed to fit the data even when the observed means were consistent with the model. The group with the larger mean in $\boldsymbol{Z}$ also had the larger mean in Y, however, the group with the larger intercept in the forward regression had the smaller intercept in the reverse regression.

It is important to note that in the examples just described only one predictor was studied. When the model did not fit the data there were no alternative models that could be tested. If the model failed to fit it was not possible to investigate the source of lack of fit because relaxing the invariance constraints would lead to identification problems. However, in the case of $p = 2$, a weaker model that relaxes constraints in the latent intercepts could be tested as shown in the next example.

The final example reported by Millsap (1998) was for the case of two predictors. The invariant common factor model failed to fit the data from 38,315 examinees. Having more than one predictor allows the possibility to relax some constraints to further examine the source of the violation to invariance. However, it should be noted that if the constraint in the invariance of the loadings is relaxed, it may lead to violations of slope invariance which complicates the interpretation of the intercepts. In this case, the only constraint that can be relaxed is invariance in the latent intercepts because it does not affect slope invariance. A model that relaxed the constraints of invariant latent intercepts was tested and showed adequate fit to the data.

In all of these examples, large sample sizes were used. The question about the power of the model to detect violations of measurement invariance remains in cases with small sample sizes. In the case with one predictor, it is expected that the low degrees of freedom and small sample sizes will produce loss of power (MacCallum, Browne & Sugawara, 1996).

Chapter 3

Research goals

A model to test factorial invariance in the presence of group differences in the regression intercepts was proposed and it was shown to be effective with low degrees of freedom and large sample sizes (Millsap, 1998). However, in many practical settings only small samples sizes (n=200) are available, and usually there is only one predictor of interest or there is only one predictor available.

The goal of the present research was to examine Type 1 errors and the statistical power of tests of fit for the model in cases with only one or two predictors, and different sample sizes. It was expected that the low degrees of freedom in the case of one predictor ($p = 1$) and the small sample sizes produce low power to detect violations of the invariant factor model. Also, it was expected that as the sample size or the number of predictors increase, the statistical power improves (MacCallum, Browne, & Sugawara, 1996).

Chapter 4

Method

Case 1. One predictor, p = 1

Data generation

The data were generated in Mplus version 5.1 using Monte Carlo simulations with 10,000 replications per condition. Data for two independent groups, each with one predictor and one criterion, were generated following a multivariate normal distribution.

Two different models were used to simulate the data. Model A consisted of an invariant factor model where the only parameter that differed between groups was the latent mean κ. Because measurement invariance holds in these data sets they were used as the comparison conditions. Under model B, not only the factor mean κ was different between groups but also the latent intercept for predictor *Z*, generating violations to measurement invariance. Both models, A and B, led to group differences in the regression intercepts.

Table 1 shows the values of the parameters shared by models A and B. it can be observed that the only values different across groups are the factor means. These parameter values were selected to reflect reliability values usually found in real data. In the common factor model the reliability of a variable is defined as the sum of the communality of the variable and the systematic variance specific to the variable. The communality for the criterion *Y* is .5, and for the predictor *Z* is .7.

So, the reliability of the variables would be expected to be higher than the communality values.

Table 1

Parameter values for the one predictor case

Parameter	Group 1	Group 2
κ_j	0	.3
Φ	.5	.5
τ_y	.3	.3
λ_z	1	1
λ_y	.6	.6
θ_z	.21	.21
θ_y	.18	.18

The variables manipulated in the simulations were the values of τ_z in model A, group difference in τ_z in model B, and the sample size:

a) Values of τ_z in model A. The data generated under model A had no group differences in τ_z. Three τ_z values were manipulated: .5, 1, and 1.5. This manipulation was important to examine if the Z latent intercepts were accurately estimated in the samples regardless of the population values.

b) Values of τ_z in model B. Under model B the data were generated assuming group differences in τ_z. The value of τ_z in group 1 was .5 in all conditions, while the values in group 2 were .7, 1 or 1.5, creating group differences in τ_z of 2, .5 and 1. Since the standard deviation of *Z* is .84 the difference in the latent intercepts represent a small (.24), medium (.6) and large (1.19) effect sizes following Cohen's d criterion.

   The different values of τ_z created group differences in the mean of *Z*, as calculated from equation 6. In group 1 the values of τ_z were.5 in all conditions under model B, so the mean of *Z* was always .5. On the other hand, the mean of *Z* in group 2 changed depending of the value of τ_z; the means of *Z* were 1, 1.3, and 1.8 for τ_z values of .7, 1, and 1.5 respectively. Thus, the *Z* mean differences between groups were .5, .8, and 1.3.

   The ratios of the group differences in τ_z to the group differences in the *Z* means are .4, .625, and .77. Any values of the latent intercepts that maintained these ratios could have been selected.

c) Sample size. The sample sizes used were 50, 100, 200, 500, 1000, 5000, and 20,000.

Analysis

An invariant factor model was fitted to the data sets generated under model A and under model B using a confirmatory factor analysis. As indicated before, under the invariant model the latent intercepts, the factor loadings, and the unique variances were constrained to be invariant in the two groups. Additionally, the factor variances were constrained to be the same in both groups. Thus, the only parameter that was allowed to differ between the groups was the factor mean.

For identification purposes, under group 1 the factor loading of *Z* was fixed to 1 and the factor mean was fixed to 0.

Fitting a factorial invariant model to the data generated under model A allowed examining Type 1 errors. Type 1 errors were determined by the percentage of samples in which the chi-square statistic incorrectly rejected the hypothesis that the invariant model was the true model. On the other hand, fitting a factorial invariant model to the data generated under model B allowed studying statistical power. Power was determined as the percentage of samples in which chi-square correctly rejected the hypothesis of invariance when in the data the *Z* latent intercepts were not invariant.

Case 2: p = 2

Data generation

Two independent groups, each with two predictors and one criterion, were simulated for case 2 following a multivariate normal distribution. Monte Carlo simulations with 10,000 replications were generated in Mplus version 5.1.

The same models described in the case of p=1 were used. Under model A the latent intercepts for Z_1 were modeled to be invariant using the values of Z in the case of p=1. Under model B the Z_1 latent intercepts were manipulated as in Z in the case of p=1. The latent intercepts of the second predictor were generated to be invariant under models A and B.

The sample sizes were, as in the case of p=1, 50, 100, 200, 500, 1000, 5000, and 20000. Thus, a total of 42 conditions were studied in the case of p=2.

Table 2 shows the values of the parameters in case 2; all the values are shared across groups except for the factor means. As in case 1 the values were selected to reflect values usually found in practice. The communalities for Z_1 and Y are the same as in the case of p=1, and the communality for Z_2 is .52.

Table 2

Parameter values for the two predictors case

Parameter	Group 1	Group 2
κ_j	0	.3
Φ	.5	.5
τ_{z2}	.6	.6
τ_y	.3	.3
λ_{z1}	1	1
λ_{z2}	.8	.8
λ_y	.6	.6
θ_{z1}	.21	.21
θ_{z2}	.30	.30
θ_y	.18	.18

Analysis

An invariant factor model was fitted to the data sets generated under model A and under model B using a confirmatory factor analysis. The latent intercepts, the factor loadings, the unique variances, and the factor variances were invariant across groups. The only parameter that was allowed to differ between the groups was the factor mean.

For identification purposes under group 1 the factor loading of Z_1 was fixed to 1 and the factor mean was fixed to 0.

Statistical power was studied when fitting the invariant factor model to the data generated under model B, and Type I errors were studied when fitting the invariant factor model to the data generated under model A.

Chapter 4

Results

Case 1, p=1

Type I errors

Type 1 errors were studied when fitting the invariant factor model to the data generated under model A as the percentage of samples in which the chi-square statistic incorrectly rejected the null hypothesis. The degrees of freedom for the model are 3 as calculated by p (p + 2), so the critical value for chi-square is 7.81. Type I errors are shown in Table 3. The results indicate that Type I errors are approximately what would be desirable at an alpha level of .05.

Table 3

Percentage of samples with $p < .05$ when p=1 under Model A (Type I errors)

N	$\tau_z = .5$	$\tau_z = 1$	$\tau_z = 1.5$
50	5.9	5.9	5.8
100	5.6	5.5	5.8
200	5	5.2	5.3
500	5.2	5.2	5.1
1000	5.3	4.9	5.3
5000	5	5.1	4.9
20000	5.4	5.1	5

Statistical Power

Statistical power was examined when fitting an invariant factor model to data generated under model B as the percentage of samples in which the chi square correctly rejected the null hypothesis. The degrees of freedom for the model are 3, and the corresponding critical value of the chi square is 7.81.

Table 4 shows the statistical power when p=1. The results indicate that the model has low power to reject the null hypothesis at any sample size. In approximately 95% of the samples the chi-square statistic indicated that the invariant factor model fitted the data even though this was not true.

Table 4

Percentage of samples with $p < .05$ when p=1 under Model B (Statistical power)

N	τ_z .5 vs .7	τ_z .5 vs 1	τ_z .5 vs 1.5
50	5.9	5.7	6.3
100	5.4	5.5	5.3
200	5	5	5.1
500	4.9	5.1	5.2
1000	5	5.3	5
5000	4.8	4.9	5.3
20000	4.8	5.3	5.2

Unique variance estimates for Z

Unexpected results were obtained regarding the values of the *Z* unique variances. As can be observed in Table 5, a substantial percentage of negative *Z* unique variances were obtained under models A and B. However, this proportion was larger when the invariant factor model was fitted to the data generated under model B than when it was fitted to the data generated under model A.

Not only the percentage of samples with negative unique variances was different between models A and B but also the impact of sample size in each model was different. While in model A the percentage of negative *Z* unique variances decreased as sample size increased, in model B it increased.

It is also interesting to note that in model A the percentage of negative *Z* unique variances did not change as a function of the values in τ_z. It should be recalled that under model A three different values of τ_z were manipulated: .5, 1 and 1.5. The percentage of negative unique variances for a specific sample size was the same in the three values of τ_z. For example, when the sample size was of 100 the percentage of samples with negative *Z* unique variances was 21% in the three values of τ_z.

In model B the percentage of negative *Z* unique variances increased as the difference in the latent intercepts increased; when the difference between the values of τ_z was .2, more than 60% of the samples had negative *Z* unique variances; the percentage of negative Z unique variances increased to more than

90% when the difference in the *Z* latent intercepts increased to .5; by the time the difference in the latent intercepts was 1 all the *Z* unique variances were negative.

Table 5

Percentage of samples with *Z* negative unique variances when p=1

N	Data generated for model A			Data generated for model B		
	$\tau_{z.}$ = .5	$\tau_{z.}$ = 1	$\tau_{z.}$ = 1.5	$\tau_{z.}$ .5 vs .7	$\tau_{z.}$ .5 vs 1	$\tau_{z.}$ .5 vs 1.5
50	24.7	25.2	25.1	59.4	91.2	99.9
100	20.8	21.1	20.8	66.2	97.8	100
200	13.7	14.1	13.5	73.3	99.8	100
500	4.50	4.40	4.40	83.1	100	100
1000	0.7	0.7	0.8	91.9	100	100
5000	0	0	0	99.9	100	100
20000	0	0	0	100	100	100

Origin of the negative Z unique variance: sample estimates

In order to explain the large percentage of samples with negative *Z* unique variances under model B, the distribution of the sample estimates was examined. It was found that the population differences in the *Z* latent intercepts caused a series of distortions in the sample estimates that led to the negative unique variances.

From equation 6 it can be seen that the *Z* latent intercept has a direct impact in the expected value of *Z*, so the distortions began with the parameter estimates that affect the expected values: the latent intercepts, the loadings, and the factor mean. The first parameter affected was the *Z* latent intercept in group 2, with sample estimates smaller than the population values. The underestimation of the *Z* latent intercept in group 2 caused the inflation of the factor mean in group 2, which in turn caused an underestimation in the *Y* loading.

The covariance structure was also affected due to violations of measurement invariance. In equation 7 it can be seen that the *Y* loading affects the variance of *Y* and the covariance of *Y* and *Z*. As a consequence of the underestimation of the *Y* loading, the sample estimates of the factor variance and the *Y* unique variance were inflated. Finally, the inflated values of the factor variance produced an underestimation of the *Z* unique variances.

Figure 2 shows the chain reaction that started with the population differences in the *Z* latent intercepts and ended up with the *Z* negative unique variances. In order to explain in detail the mechanism that led to the Z unique variances, the distribution of the sample estimates affected are presented next.

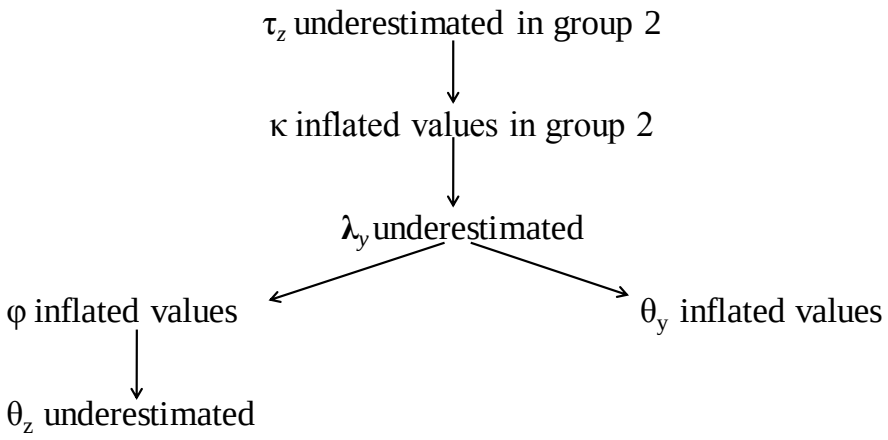


Figure 2. Chain reaction of the alterations in the sample estimates began by the differences in the *Z* latent intercepts when p=1.

Z latent intercepts τ_z. The three population values of the *Z* latent intercepts used to generate the data under model A were: .5, 1 and 1.5 In Table 7 it is shown that, when fitting the invariant factor model to model A, the sample estimates of the *Z* latent intercepts were close to the parameter values. It is important to note that the estimates of the Z latent intercepts for small sample sizes showed large variability as indicated by the large standard errors.

In contrast, the data under model B were generated to have population differences in the *Z* latent intercepts; in group 2 three different values of the Z latent intercepts were studied: .7, 1 and 1.5; in group 1 the value of the *Z* latent intercept was .5 in all conditions. As a consequence of fitting an invariant factor model to data simulated under model B, the sample values of the *Z* latent intercepts were forced to be the same. Table 6 shows that the sample estimates of

the *Z* latent intercept in all conditions under model B are .5, which corresponds to the true population value of τ_z in group 1.

Table 6

Average estimates of τ_z when p=1

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.491 (.116)	.497 (.083)	.5 (.06)	.5 (.038)	.5 (.027)	.5 (.012)	.5 (.006)
	1	.992 (.115)	.997 (.084)	1.01 (.06)	.999 (.038)	.999 (.027)	1 (.012)	1 (.006)
	1.5	1.49 (.117)	1.5 (.084)	1.5 (.06)	1.5 (.038)	1.5 (.027)	1.5 (.012)	1.5 (.006)
Model B	.5 vs .7	.493 (.117)	.497 (.083)	.501 (.061)	.501 (.037)	.5 (.027)	.5 (.012)	.5 (.006)
	.5 vs 1	.493 (.116)	.497 (.083)	.5 (.06)	.501 (.038)	.5 (.027)	.5 (.012)	.5 (.006)
	.5 vs 1.5	.492 (.116)	.5 (.083)	.499 (.06)	.5 (.038)	.5 (.027)	.5 (.012)	.5 (.006)

Note. Standard errors are in parenthesis.

In order to understand these results it should be considered that the *Z* latent intercept along with the *Z* loading and the factor mean directly affect the group means of *Z*, as can be seen from equation 6. For identification purposes the factor mean was fixed to zero and the *Z* loading was fixed to one in the first group. As a consequence, the only parameter left to estimate that affects the mean of Z in group 1 was the *Z* latent intercept. Since the factor mean was fixed to the

population value of zero, the sample value of the *Z* latent intercept in group 1 was the true population value of .5.

Under factorial invariance the *Z* latent intercept must be the same in both groups, so the second group got the value of .5, which underestimates the true population value.

Factor mean κ_2. As stated before, to identify the model in the CFA the factor mean was fixed to zero in the first group and it was freely estimated on the second group. The sample estimates of the factor mean are shown in Table 7.

Under model A, the values of κ_2 correspond to .3, which is the true population value of the factor mean in group 2. It is important to note the high variability of estimated values in small sample sizes; as the sample size increased the estimated values were more accurately measured as indicated by the decrease in the standard errors. Under model B, the estimated values of the factor mean increased as the population difference in the *Z* latent intercepts increased. For example, it can be observed in Table 7 that the sample estimate of the factor mean was 5 times larger than the value of the true population mean when the difference in the *Z* latent intercepts was 1.

Table 7

Average estimates of κ_2 when p=1. Population value κ_2=.3

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.318 (.158)	.305 (.116)	.301 (.085)	.3 (.053)	.3 (.037)	.3 (.017)	.3 (.008)
	1	.317 (.157)	.305 (.116)	.301 (.084)	.301 (.053)	.301 (.038)	.3 (.017)	.3 (.008)
	1.5	.319 (.16)	.305 (.117)	.3 (.084)	.3 (.054)	.3 (.038)	.3 (.017)	.3 (.008)
Model B	.5 vs .7	.513 (.161)	.505 (.118)	.499 (.085)	.499 (.054)	.5 (.037)	.5 (.017)	.5 (.008)
	.5 vs 1	.815 (.162)	.805 (.117)	.799 (.084)	.799 (.053)	.8 (.038)	.8 (.017)	.8 (.008)
	.5 vs 1.5	1.314 (.162)	1.301 (.118)	1.301 (.084)	1.3 (.053)	1.3 (.037)	1.3 (.017)	1.3 (.008)

Note. Standard errors are in parenthesis.

The increase in the estimates of the factor mean under model B can be explained as a consequence of the underestimation of the *Z* latent intercept in group 2. It should be recalled that under model B the mean of *Z* was bigger in group 2 than in group 1 in all conditions due to population differences in the *Z* latent intercepts and the factor mean. However, because of invariance constraints the *Z* latent intercept was fixed to be equal in both groups, underestimating the true population value in group 2. The *Z* loading was also fixed to be the same in both groups. As a consequence, the only parameter that could reflect the

population differences in means was the factor mean. To compensate for the underestimation of the sample estimates of the *Z* latent intercept in group 2 the sample values of the factor mean were larger than the true population values.

Y loading λ_y. The sample estimates of the *Y* loading are shown in Table 8. It can be observed that as the sample size increased the values of the *Y* loading got closer to the population value of .6 and the standard errors became smaller when the data was generated under model A.

Table 8

Average estimates of λ_y when p=1. Population value λ_y=.6

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.765 (.726)	.667 (.446)	.622 (.223)	.607 (.114)	.604 (.078)	.601 (.034)	.6 (.017)
	1	.768 (.738)	.666 (.458)	.621 (.21)	.605 (.114)	.603 (.076)	.601 (.034)	.6 (.017)
	1.5	.763 (.717)	.668 (.431)	.623 (.234)	.606 (.114)	.603 (.078)	.601 (.034)	.6 (.017)
Model B	.5 vs .7	.394 (.207)	.364 (.142)	.358 (.101)	.359 (.063)	.36 (.044)	.36 (.02)	.36 (.01)
	.5 vs 1	.238 (.113)	.228 (.087)	.222 (.065)	.224 (.041)	.224 (.029)	.225 (.013)	.225 (.006)
	.5 vs 1.5	.147 (.071)	.139 (.056)	.138 (.041)	.138 (.026)	.138 (.019)	.138 (.008)	.139 (.004)

Note. Standard errors are in parenthesis.

In contrast, the sample estimates of the *Y* loading under model B were different from the population values, and this difference became larger as the population difference in the *Z* latent intercepts increased.

To explain the small values of the *Y* loading it should be noticed that the expected values of *Y* are determined by the *Y* loading, the *Y* latent intercept, and the factor mean as shown in equation 6. Since the *Y* loading and the *Y* latent intercept are the same in the two populations, the population differences in the expected values of *Y* are due only to the factor means. However, as explained in the previous section, the sample estimate of the factor mean in group 2 was inflated because of the population differences in the *Z* latent intercepts. To compensate for the large estimated values of the factor mean, the sample estimate of the *Y* loading decreased as the population difference in the *Z* latent intercepts increased.

Factor variance Φ. The factor variance is a parameter that directly affects the covariance structure as shown in equation 7. Table 9 shows the sample estimates for the factor variances. For the data simulated under model A the sample estimates of the factor variance got closer to the population value of .5 as the sample size increased.

Under model B the sample estimates of the factor variance were larger than the population value; as the population differences in the *Z* latent intercepts increased, the estimate of the factor variance also increased. As the sample size

increased, the factor variance decreased, however, the sample estimates remained larger than the population value.

Table 9

Average estimates of Φ when p=1. Population value Φ=.5

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.707 (1.16)	.621 (.749)	.547 (.34)	.511 (.108)	.505 (.07)	.501 (.03)	.5 (.015)
	1	.705 (1.197)	.614 (.689)	.547 (.353)	.514 (.158)	.505 (.069)	.501 (.03)	.501 (.015)
	1.5	.724 (1.31)	.618 (.79)	.544 (.347)	.512 (.108)	.505 (.07)	.501 (.03)	.5 (.015)
Model B	.5 vs .7	1.225 (1.98)	1.094 (1.355)	.944 (.65)	.865 (.19)	.846 (.117)	.836 (.049)	.834 (.024)
	.5 vs 1	2.002 (3.01)	1.747 (2.011)	1.545 (.97)	1.389 (.323)	1.359 (.194)	1.337 (.082)	1.334 (.04)
	.5 vs 1.5	3.21 (4.282)	2.953 (3.328)	2.473 (1.446)	2.253 (.527)	2.209 (.332)	2.177 (.14)	2.168 (.069)

Note. Standard errors are in parenthesis.

The inflation in the estimates of the factor variance under model B can be explained as a consequence of the underestimation of the *Y* loading. The covariance between *Y* and *Z* is determined by the *Y* loading, the *Z* loading, and the factor variance. The *Y* loading was underestimated as explained before, and since the *Z* loading was fixed to one in both groups the only parameter that could

compensate for the small values of the *Y* loading was the factor variance. As a consequence, the sample estimates of the factor variance increased as the population differences in the *Z* latent intercepts increased.

Y unique variance θ_y. The estimated values of θ_y are shown in Table 10. For the data generated under model A the values of the *Y* unique variance got closer to the population value of .18, and the standard errors became smaller as the sample size increased. In contrast, for the data generated under model B the values of the *Y* unique variance became farther away from the population value as the population differences in the *Z* latent intercepts increased.

The large values of the *Y* unique variance are due to the underestimation of the *Y* loading. From equation 7 it can be shown that the parameters that determine the variance of *Y* are the squared of the *Y* loading, the factor variance, and the *Y* unique variance. The estimated values of the *Y* loading became even smaller after being squared, so the inflation of the factor variance was not enough to compensate for the underestimation of the *Y* loading. To compensate for the small values of the squared *Y* loading, the estimates of the *Y* unique variance were also inflated.

Table 10

Average estimates of θ_y when p=1. Population value θ_y=.18

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.128 (.219)	.158 (.136)	.172 (.069)	.177 (.036)	.179 (.025)	.18 (.011)	.18 (.005)
	1	.127 (.221)	.159 (.14)	.173 (.066)	.178 (.036)	.179 (.024)	.18 (.011)	.18 (.005)
	1.5	.128 (.216)	.158 (.13)	.172 (.073)	.178 (.036)	.179 (.025)	.18 (.011)	.18 (.005)
Model B	.5 vs .7	.238 (.073)	.248 (.05)	.251 (.036)	.252 (.023)	.252 (.016)	.252 (.007)	.252 (.004)
	.5 vs 1	.283 (.054)	.289 (.04)	.292 (.029)	.292 (.018)	.292 (.013)	.292 (.006)	.292 (.003)
	.5 vs 1.5	.310 (.05)	.315 (.036)	.317 (.026)	.318 (.016)	.318 (.012)	.316 (.005)	.318 (.003)

Note. Standard errors are in parenthesis.

Z unique variances. In Table 11 it can be observed that no matter if the data were generated under model A or B, on average all the *Z* unique variances were negative for a sample size of 50. However, under model A as the sample size increased the mean of the *Z* unique variances rapidly increased; with sample sizes of 100 it became positive, and with a sample size of 500 the estimates were very close to the population value of .21.

Under model B the mean of the *Z* unique variances became more negative as the population difference in the *Z* latent intercepts increased. Although the *Z*

unique variances increased with sample size, the mean remained negative in all conditions under model B.

Table 11

Average estimates of θ_z when p=1. Population value θ_z=.21

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	-.013 (1.157)	.082 (.747)	.16 (.338)	.198 (.106)	.204 (.069)	.209 (.029)	.21 (.015)
	1	-.011 (1.193)	.088 (.685)	.16 (.352)	.195 (.146)	.204 (.067)	.209 (.029)	.21 (.015)
	1.5	-.029 (1.308)	.084 (.787)	.163 (.345)	.196 (.105)	.204 (.069)	.209 (.029)	.21 (.015)
Model B	.5 vs .7	-.529 (1.974)	-.391 (1.35)	-.239 (.648)	-.156 (.187)	-.137 (.113)	-.126 (.047)	-.124 (.024)
	.5 vs 1	-1.306 (3.004)	-1.045 (2.006)	-.837 (.964)	-.68 (.318)	-.649 (.191)	-.627 (.08)	-.624 (.039)
	.5 vs 1.5	-2.514 (4.274)	-2.25 (3.322)	-1.766 (1.442)	-1.544 (.521)	-1.5 (.329)	-1.467 (.138)	-1.458 (.068)

Note. Standard errors are in parenthesis.

The explanation of the negative values of the *Z* unique variance under model B is concerned with the inflation in the estimates of the factor variance. It should be noted that the variance of *Z* is determined by the squared *Z* loading, the factor variance and the *Z* unique variance. Since the *Z* loading is fixed to one in both groups, the only parameter that can compensate for the inflation in the factor

variance was the Z unique variance, so its estimated values became smaller than the population values.

The negative values in the *Z* unique variances under model A are due to sampling error. In Tables 7, 8, and 9 it can be observed that at small sample sizes the sample estimates have large standard errors. As a consequence of the high variability in the estimates, in some samples the values were far from the true population value, originating the chain reaction explained for model B. For example, Table 9 shows that at small sample sizes the standard errors of the factor variance are large, thus, in some samples the values of the factor variance are noticeably inflated, which produced the underestimation of the *Z* unique variance.

Constraint in Z unique variance

The results from the one predictor case suggest that the low statistical power of the model is related to the negative unique variance of *Z*. To investigate this relationship, a set of simulations using the same population parameters as in the previous simulations were run with the difference that the *Z* unique variance was constrained to be a positive value.

The results for the Type I errors are shown in Table 12. It can be observed that at small sample sizes the Type I errors are slightly higher than expected, but as the sample size increase the Type I errors became closer to 5%. It should also be noted that the Type I errors are very similar to the ones reported when the unique variance was not constrained (Table 3).

Table 12

Type I errors with Z unique variance constrained to be positive when p=1

N	$\tau_z = .5$	$\tau_z = 1$	$\tau_z = 1.5$
50	6.6	6.5	6.9
100	5.8	5.9	5.8
200	5.6	5.5	5.2
500	5.2	5.1	4.9
1000	5	4.8	5
5000	4.9	4.8	4.7
20000	4.8	4.9	4.8

Regarding statistical power, Table 13 indicates that when the *Z* unique variances are positive the chi-square fit statistic can detect violations to the measurement invariant model. The power is high when the differences in the latent intercepts are large (.5 vs 1.5): with only a sample size of 100 the power is .97. However, when the difference in latent intercepts is only .2 it is necessary to have a sample size of 5,000 to have statistical power of .8.

Table 13

Statistical power with Z unique variance constrained to be positive when p=1

N	τ_z .5 vs .7	τ_z .5 vs 1	τ_z .5 vs 1.5
50	10.4	30.1	76.9
100	11.4	46.7	97.1
200	12.8	75.6	100
500	17.7	98.9	100
1000	25	100	100
5000	79.8	100	100
20000	100	100	100

From these results it is possible to say that the lack of power in the previous simulations and the negative *Z* unique variances are closely related. When the unique variance is not constrained, it adopts negative values making the model fit the data; in contrast, when the unique variance is forced to be positive the model no longer fits the data and the chi square fit statistic is able to detect it. In other words, the lack of fit of the model can be detected with the chi-square fit statistic or by the presence of *Z* negative unique variances. However, it is important to note that if the negative unique variance is used as a way to detect violations to the measurement invariant model, it provides higher power than the chi square fit statistic. When comparing Table 5 and Table 13 it can be observed

that higher power can be achieved when using negative unique variances as indicators of violations to factorial invariance.

Case 2, p=2

Type I errors

Table 14 shows that the results of Type I errors in the case of p=2 closely match the results obtained in the case of p=1. In about 5% of the samples the chi-square incorrectly rejected the hypothesis of factorial invariance. In other words, the Type I errors are what would be expected at an alpha level of .05.

Table 14

Percentage of samples with $p < .05$ when p=2 under Model A (Type I errors)

N	$\tau_z = .5$	$\tau_z = 1$	$\tau_z = 1.5$
50	6.4	6.4	6.6
100	5.4	5.3	5.0
200	5.6	5.3	5.0
500	4.9	5.3	5.0
1000	4.8	5.2	5.1
5000	5.2	5.1	5.1
20000	5.3	5.1	5.1

Statistical Power

The percentage of samples in each condition that correctly rejected the hypothesis of factorial invariance is shown in Table 15. The results indicate that, in contrast to the case of p=1, the model has enough power to detect violations of factorial invariance. It can be observed that as the population differences in the Z_1 latent intercepts increased and the sample size increased the power also increased. When the difference in the Z_1 latent intercepts between the two populations was .2 the sample size needed to have power of .80 was 500. When the population differences in the Z_1 latent intercepts increased to .5 a sample size of 100 was enough to have power of .90; and a sample size of only 50 was enough to achieve a power of .90 when the population differences in the Z_1 latent intercept was 1.

Table 15

Percentage of samples with $p < .05$ when p=2 under Model B (Statistical power)

N	$\tau_{z\,.}$5 vs .7	$\tau_{z\,.}$5 vs 1	$\tau_{z\,.}$5 vs 1.5
50	14.4	56.1	89.7
100	23.3	89.9	99.8
200	44.4	99.8	100
500	89.3	100	100
1000	99.8	100	100
5000	100	100	100
20000	100	100	100

Unique variance estimates for Z_1

As shown in Table 16, negative Z_1 unique variances were also observed in the case of two predictors, but in a smaller proportion than in the case of one predictor.

As in the case of p=1, a larger percent of Z_1 negative unique variances was obtained under model B than under model A. Actually, the percentage of samples with negative Z_1 unique variances was nearly zero under model A even in small sample sizes.

Table 16

Percentage of negative unique variances for Z_1 when p=2

N	Data generated for model A			Data generated for model B		
	$\tau_{z.}$= .5	$\tau_{z.}$= 1	$\tau_{z.}$= 1.5	$\tau_{z.}$.5 vs .7	$\tau_{z.}$.5 vs 1	$\tau_{z.}$.5 vs 1.5
50	0.90	1.0	0.8	3.2	32.7	96.2
100	0.10	0.02	0.1	0.5	26.3	99.7
200	0	0	0	0	17.8	100
500	0	0	0	0	6.9	100
1000	0	0	0	0	1.9	100
5000	0	0	0	0	0	100
20000	0	0	0	0	0	100

Another similarity with the case of p=1 is that in model B the percentage of negative unique variances in Z_1 increased as the difference in the latent intercepts increased. However, the effect of sample size was not the same in all conditions in model B. When the population difference in the Z_1 latent intercept was 1 the percentage of negative Z_1 unique variances increased as the sample size increased, thus, replicating the results of the case of p=1. But when the population differences in the Z_1 latent intercepts were .2 and .5, the percentage of negative unique variances decreased with sample size. In fact, when the difference in the Z_1 latent intercepts was .2, all the samples had positive Z_1 unique variances at a sample size of 200; when the difference in the latent intercepts was .5, most of the samples had positive Z_1 unique variances by a sample size of 1000.

Origin of the negative Z_1 unique variance: sample estimates

In order to explain the large percentage of samples with negative Z_1 unique variances under model B, the distribution of the sample estimates was examined. As in the case of p=1, a chain reaction that started with the population differences in the Z_1 latent intercepts produced a series of distortions in the sample estimates that led to the Z_1 negative unique variances.

The difference in the Z_1 latent intercepts first affected the sample estimates related to the expected values: the Z_1 latent intercept in group 2 was underestimated, causing inflated values of the factor mean in group 2, which in turn caused an underestimation in the *Y* loading. Because of the underestimation

in the *Y* loading, the *Y* unique variance and the factor variance were inflated, producing the underestimation of the Z_1 unique variances. Figure 3 shows the chain reaction for the case of p=2.

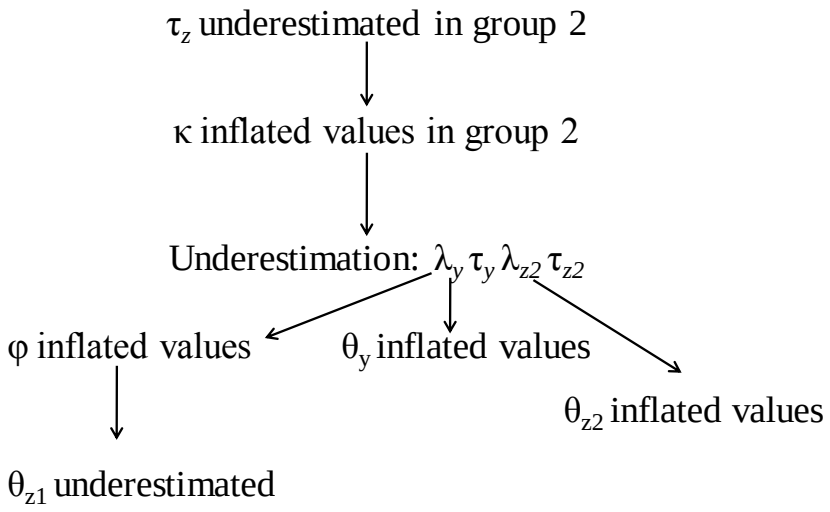


Figure 3. Chain reaction of the alterations in the sample estimates originated by the differences in the *Z* latent intercepts when p=2.

In order to explain in detail the mechanism that led to the Z_1 unique variances, the sample estimates affected are presented next.

Z latent intercepts τ_{z1}. In the case of p=2 the latent intercepts in Z_1 were manipulated to have three different values under model A: .5, 1 and 1.5. In Table 17 it is shown that the estimated values of the Z_1 latent intercepts under model A closely match the population values.

Under model B the groups were generated to have differences in the Z_1 latent intercepts. When fitting the invariant factor model to model B, the Z_1 latent

intercepts were forced to be the same in both groups. As a result, the sample estimates of the Z_1 latent intercepts were close to the population value of group 1, as shown in Table 17.

Table 17

Average estimates of τ_{z1} for each condition when p=2.

Values of τ_{z1}		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.497 (.115)	.501 (.082)	.5 (.058)	.5 (.037)	.5 (.026)	.5 (.012)	.5 (.006)
	1	1 (.116)	1 (.082)	1 (.057)	1 (.037)	1 (.026)	1 (.012)	1 (.006)
	1.5	1.5 (.116)	1.5 (.082)	1.5 (.057)	1.501 (.037)	1.5 (.026)	1.5 (.012)	1.5 (.006)
Model B	.5 vs .7	.533 (.123)	.533 (.087)	.533 (.061)	.534 (.039)	.534 (.028)	.534 (.012)	.534 (.006)
	.5 vs 1	.53 (.144)	.528 (.103)	.525 (.072)	.524 (.045)	.523 (.032)	.523 (.015)	.523 (.007)
	.5 vs 1.5	.47 (.135)	.461 (.084)	.46 (.058)	.46 (.036)	.46 (.026)	.460 (.011)	.46 (.006)

Note. Standard errors are in parenthesis.

The fact that the sample estimates of the Z_1 latent intercepts in group 2 were the parameter values of group 1, has the same explanation given in the case of p=1. In order to identify the model in the CFA, the factor mean was fixed to zero and the Z_1 loading was fixed to one in the first group, so the only parameter that was free to estimate in group 1 that affects the expected value of Z_1 was the

Z_1 latent intercept. Since the factor mean was fixed to the population value of zero, the sample estimate of the Z_1 latent intercept in group 1 was the population value of .5. The value of the latent intercept of Z_1 in group 2 was the same as in group 1 because of invariance constraints.

Factor mean κ_2. As in the case of p=1 the factor mean was fixed to zero in group 1 for identification purposes, and was freely estimated in group 2. The sample estimates of the factor mean in group 2 are shown in Table 18.

The results closely resembled the findings of the case of p=1. The sample estimates of the factor mean in group 2 under model A correspond to the population value of .3. In contrast, the sample estimates under model B increased as the population differences in the Z_1 latent intercepts increased; when the population differences in the Z_1 latent intercepts were 1, the factor mean reached values of 1.34, which is approximately 4 times larger than the population value.

The inflated values of the factor mean of group 2 under model B can be explained as a consequence of the underestimation of the Z_1 latent intercept in group 2. Under model B, the Z_1 means were different in each group due to the differences in the Z_1 latent intercepts and in the factor means. However, because the Z_1 latent intercepts were constrained to be the same in both groups the Z_1 mean differences had to be attributed to the factors means. In other words, the factor means increased to compensate for the underestimation of the Z_1 latent intercept in group 2.

Table 18

Average estimates of κ_j when p=2. Population value κ_j = .3

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.302 (.159)	.299 (.112)	.299 (.079)	.301 (.05)	.3 (.035)	.3 (.016)	.3 (.008)
	1	.302 (.161)	.299 (.112)	.3 (.079)	.3 (.05)	.3 (.035)	.3 (.016)	.3 (.008)
	1.5	.303 (.161)	.301 (.112)	.3 (.079)	.3 (.05)	.3 (.035)	.3 (.016)	.3 (.008)
Model B	.5 vs .7	.435 (.178)	.434 (.125)	.434 (.089)	.433 (.057)	.432 (.04)	.432 (.018)	.432 (.009)
	.5 vs 1	.738 (.232)	.744 (.167)	.749 (.118)	.753 (.074)	.754 (.053)	.754 (.024)	.754 (.012)
	.5 vs 1.5	1.36 (.211)	1.377 (.117)	1.379 (.078)	1.38 (.049)	1.38 (.035)	1.379 (.015)	1.379 (.008)

Note. Standard errors are in parenthesis.

Y loading λ_y. In Table 19 it can be observed that the sample estimates of the *Y* loading matched the population value of .6 under model A. But, under model B the sample estimates decreased as the population differences in the Z_1 latent intercepts increased. It is important to note that the underestimation of the *Y* loading was not as large as in the case of p=1.

From equation 6 it can be shown that the population differences in the expected values of *Y* are due only to group differences in the factor means

because there are no differences in the *Y* latent intercepts or *Y* loadings. However, under model B the sample value of the factor mean in group 2 is inflated as a consequence of the underestimation of the Z_1 latent intercept in group 2. To compensate for the large values in the factor mean in group 2, the sample estimates of the *Y* loading decreased.

Table 19

Average estimates of λ_y when p=2. Population value λ_y=.6

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.604 (.101)	.601 (.069)	.601 (.048)	.601 (.031)	.6 (.021)	.6 (.01)	.6 (.005)
	1	.604 (.102)	.603 (.069)	.601 (.048)	.601 (.031)	.6 (.021)	.6 (.01)	.6 (.005)
	1.5	.605 (.102)	.603 (.069)	.602 (.049)	.601 (.03)	.6 (.021)	.6 (.01)	.6 (.005)
Model B	.5 vs .7	.559 (.103)	.556 (.07)	.555 (.049)	.556 (.031)	.555 (.022)	.555 (.01)	.555 (.005)
	.5 vs 1	.438 (.121)	.433 (.084)	.432 (.06)	.431 (.038)	.430 (.027)	.43 (.012)	.43 (.006)
	.5 vs 1.5	.23 (.089)	.225 (.049)	.225 (.032)	.225 (.021)	.225 (.014)	.226 (.007)	.226 (.003)

Note. Standard errors are in parenthesis.

Z_2 loading λ_{z2}. Table 20 shows that while under model A the sample estimates of the Z_2 loading were the population value, the values were

underestimated under model B. The sample estimates of the Z_2 loading in model B decreased as the population difference in the Z_1 latent intercepts increased.

Table 20

Average estimates of λ_{z2} when p=2. Population value λ_{z2}=.8

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.806 (.134)	.801 (.091)	.801 (.063)	.801 (.04)	.8 (.028)	.8 (.013)	.8 (.006)
	1	.805 (.136)	.803 (.092)	.802 (.064)	.8 (.04)	.8 (.028)	.8 (.013)	.8 (.006)
	1.5	.81 (.132)	.803 (.09)	.803 (.065)	.8 (.04)	.8 (.028)	.8 (.013)	.8 (.006)
Model B	.5 vs .7	.744 (.134)	.741 (.092)	.74 (.065)	.74 (.041)	.74 (.029)	.740 (.013)	.74 (.006)
	.5 vs 1	.583 (.159)	.578 (.111)	.576 (.08)	.574 (.05)	.573 (.035)	.573 (.016)	.573 (.008)
	.5 vs 1.5	.309 (.118)	.302 (.065)	.3 (.042)	.302 (.027)	.302 (.019)	.302 (.008)	.302 (.004)

Note. Standard errors are in parenthesis.

The underestimation of the Z_2 loading in model B has the same explanation given above for the underestimation of the *Y* loading. Equation 6 shows that the population differences in the expected values of Z_2 are only due to differences in the factor means since there are no population differences in the Z_2 loading and in the Z_2 latent intercepts. However, the sample estimates of the factor

mean of group 2 were inflated under model B; to compensate for the large values of the factor mean the values of the Z_2 latent intercepts were underestimated.

Y latent intercept τ_y. In Table 21 it can be observed that under model A the estimates of the *Y* latent intercept were .3, which corresponds to the population value. In contrast, under model B the estimates became slightly lower than .3 as the difference in the Z_1 latent intercepts increased.

Table 21

Average estimates of τ_y when p=2. Population value τ_y=.3

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.299 (.077)	.302 (.055)	.3 (.038)	.3 (.024)	.3 (.017)	.3 (.008)	.3 (.004)
	1	.3 (.077)	.301 (.055)	.3 (.038)	.3 (.025)	.3 (.017)	.3 (.008)	.3 (.004)
	1.5	.3 (.078)	.3 (.054)	.3 (.038)	.3 (.024)	.3 (.017)	.3 (.008)	.3 (.004)
Model B	.5 vs .7	.273 (.076)	.271 (.054)	.27 (.038)	.27 (.024)	.27 (.017)	.27 (.008)	.27 (.004)
	.5 vs 1	.238 (.075)	.233 (.052)	.23 (.036)	.229 (.023)	.228 (.016)	.228 (.007)	.228 (.004)
	.5 vs 1.5	.237 (.079)	.235 (.056)	.235 (.039)	.235 (.025)	.235 (.017)	.235 (.008)	.234 (.004)

Note. Standard errors are in parenthesis.

From equation 6 it can be observed that the *Y* latent intercept along with the *Y* loading and the factor mean affects the expected values of *Y*. In model B, to compensate for the large values in the factor mean in group 2 there was a slight decrease in the sample estimates of *Y* latent intercept.

Z_2 latent intercept τ_{z2}. The sample estimates of the Z_2 latent intercept are shown in Table 22. For the data simulated under model A the mean of the sample estimates was the population value of .6. Under model B, as the population differences in the Z_1 latent intercept increased, the sample estimates of the Z_2 latent intercept decreased to .5.

The same explanation provided for the decrease in the estimated values of the *Y* latent intercept applies for the decrease in the Z_2 latent intercept. In equation 6 it can be observed that along with the Z_2 loading, another way to compensate for the large value of the factor mean is through the Z_2 latent intercept. The decrease in the sample estimates of the Z_2 latent intercept is a consequence of the inflated values of the factor mean of group 2 under model B.

Table 22

Average estimates of τ_{z2} when p=2. Population value τ_{z2} =.6

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.598 (.101)	.602 (.072)	.6 (.051)	.6 (.032)	.6 (.023)	.6 (.01)	.6 (.005)
	1	.6 (.102)	.601 (.072)	.6 (.051)	.6 (.032)	.6 (.023)	.6 (.01)	.6 (.005)
	1.5	.6 (.102)	.6 (.072)	.6 (.051)	.6 (.032)	.6 (.023)	.6 (.01)	.6 (.005)
Model B	.5 vs .7	.562 (.1)	.562 (.07)	.56 (.05)	.56 (.032)	.56 (.022)	.56 (.01)	.56 (.005)
	.5 vs 1	.516 (.098)	.512 (.069)	.507 (.048)	.505 (.03)	.505 (.021)	.504 (.009)	.504 (.005)
	.5 vs 1.5	.515 (.103)	.512 (.073)	.513 (.051)	.512 (.032)	.511 (.023)	.512 (.01)	.512 (.005)

Note. Standard errors are in parenthesis.

Factor variance Φ. Table 23 shows the sample estimates of the factor variances. Under model A the sample estimates of the factor variance were .5 which corresponds to the population value. However, under model B the sample estimates of the factor variance increased as the population difference in the Z_1 latent intercepts increased.

As in the case of p=1, the large values of the factor variance under model B are explained as a consequence of the underestimation of the *Y* loading. The covariance between the criterion and the predictors is determined by the loadings

of the criterion, the loadings of the predictors, and the factor variance. As explained before, the Z_1 loading was fixed to 1 in both groups for identification purposes, and the Y and Z_2 loadings were underestimated; in order to compensate for the underestimation of the loadings, the estimates of the factor variance were inflated.

Table 23

Average estimates of Φ when p=2. Population value Φ = .5

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.497 (.115)	.499 (.082)	.499 (.057)	.5 (.036)	.501 (.025)	.5 (.011)	.5 (.006)
	1	.495 (.116)	.497 (.081)	.499 (.057)	.5 (.036)	.5 (.026)	.5 (.012)	.5 (.006)
	1.5	.494 (.115)	.5 (.081)	.5 (.058)	.5 (.036)	.5 (.025)	.5 (.012)	.5 (.006)
Model B	.5 vs .7	.535 (.122)	.536 (.083)	.538 (.058)	.538 (.037)	.538 (.026)	.539 (.011)	.539 (.006)
	.5 vs 1	.658 (.167)	.651 (.104)	.649 (.071)	.649 (.044)	.648 (.031)	.648 (.014)	.648 (.007)
	.5 vs 1.5	1.115 (.528)	1.063 (.209)	1.05 (.134)	1.041 (.081)	1.039 (.056)	1.036 (.025)	1.036 (.012)

Note. Standard errors are in parenthesis.

Y unique variance θ_y. As shown in Table 24, the sample estimates of the *Y* unique variance under model A were .18, which corresponds to the parameter

value. In contrast, under model B as the difference in the Z_1 latent intercept increased the *Y* unique variance got slightly larger reaching values of .29.

Equation 7 shows that the variance of *Y* is determined by the squared of the *Y* loading, the factor variance, and the *Y* unique variance. Under model B, the sample estimates of the factor variance and the *Y* unique variance increased to compensate for small values of the *Y* loading.

Table 24

Average estimates of θ_y when p=2. Population value θ_y=.18

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.176 (.035)	.178 (.025)	.179 (.017)	.18 (.011)	.18 (.008)	.18 (.003)	.18 (.002)
	1	.176 (.035)	.178 (.024)	.179 (.017)	.18 (.011)	.18 (.008)	.18 (.003)	.18 (.002)
	1.5	.176 (.035)	.178 (.025)	.179 (.017)	.18 (.011)	.18 (.008)	.18 (.004)	.18 (.002)
Model B	.5 vs .7	.184 (.036)	.186 (.026)	.187 (.018)	.187 (.011)	.187 (.008)	.188 (.004)	.188 (.002)
	.5 vs 1	.215 (.046)	.219 (.033)	.220 (.024)	.221 (.015)	.222 (.011)	.222 (.005)	.222 (.002)
	.5 vs 1.5	.282 (.047)	.288 (.032)	.29 (.022)	.291 (.014)	.291 (.01)	.291 (.004)	.291 (.002)

Note. Standard errors are in parenthesis.

Z_2 unique variance θ_{z2}. In Table 25 the sample estimates of the unique variances of Z_2 are shown. It can be observed that under model A the estimates

were equal to the population value of .3. However, under model B the estimates became larger as the difference in the Z_1 latent intercepts increased; the estimated values of the Z_2 unique variance increased up to .5.

Table 25

Average estimates of θ_{z2} when p=2. Population value θ_{z2}=.3

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.293 (.06)	.297 (.042)	.3 (.03)	.299 (.019)	.3 (.013)	.3 (.006)	.3 (.003)
	1	.293 (.06)	.296 (.042)	.3 (.03)	.3 (.019)	.3 (.013)	.3 (.006)	.3 (.003)
	1.5	.293 (.06)	.3 (.042)	.3 (.03)	.3 (.02)	.3 (.013)	.3 (.006)	.3 (.003)
Model B	.5 vs .7	.307 (.063)	.311 (.044)	.312 (.031)	.313 (.02)	.314 (.014)	.314 (.006)	.314 (.003)
	.5 vs 1	.363 (.079)	.369 (.057)	.372 (.042)	.374 (.026)	.375 (.019)	.375 (.008)	.375 (.004)
	.5 vs 1.5	.481 (.081)	.491 (.053)	.494 (.033)	.496 (.024)	.496 (.016)	.496 (.008)	.496 (.004)

Note. Standard errors are in parenthesis.

The increase in the Z_2 unique variance has the same explanation as the increase on the *Y* unique variance. The variance of Z_2 is determined by the Z_2 loading squared, the factor variance, and the Z_2 unique variance. As previously explained, the estimates of the Z_2 loading under model B are smaller than the

population values; to compensate for this underestimation, the estimates of the factor variance and the Z_2 unique variance increased.

Z_1 unique variances. As observed in Table 26, under model A the sample estimates of the Z_1 unique variances were the population value of .21. In contrast, under model B the sample estimates decreased as the population difference in the Z_1 latent intercept increased, reaching negative values when the difference between populations in the Z_1 latent intercepts was 1.

The parameters involved in the calculation of the Z_1 variance are the Z_1 loading, the factor variance and the Z_1 unique variance. It is important to note that the Z_1 loading is fixed to one in both groups, so it cannot shrink to compensate for the inflation of the factor variance. As a consequence, the only parameter that could compensate for the large values of the factor variance was the Z_1 unique variance. The estimates of the Z_1 unique variance decreased to the point of getting negative values.

It is important to note that standard errors of the estimates of Z_1 unique variance decreased as the sample sizes increased. In the conditions in which the population differences in the Z_1 latent intercept were .2 and .5, the average of the Z_1 unique variance across samples was positive but with high variability in small sample sizes, and as a consequence some samples had negative Z_1 unique variances. As the sample size increased, the variability of the estimated values across samples decreased so all the samples had positive values in the Z_1 unique

variance. This explains the results in Table 14 showing that the percentage of samples with negative values decreased as the sample size increased for these conditions in model B.

Table 26

Average estimates of θ_{z1} when p=2. Population value θ_{z1}=.21

Values of τ_z		Sample Size						
		50	100	200	500	1000	5000	20000
Model A	.5	.199 (.075)	.205 (.051)	.208 (.036)	.209 (.022)	.209 (.016)	.21 (.007)	.21 (.004)
	1	.206 (.076)	.206 (.051)	.208 (.036)	.209 (.023)	.21 (.016)	.21 (.007)	.21 (.004)
	1.5	.201 (.075)	.206 (.051)	.208 (.036)	.209 (.023)	.21 (.016)	.21 (.007)	.21 (.004)
Model B	.5 vs .7	.175 (.093)	.181 (.061)	.184 (.042)	.186 (.027)	.187 (.019)	.187 (.008)	.187 (.004)
	.5 vs 1	.056 (.069)	.069 (.118)	.075 (.084)	.078 (.052)	.079 (.037)	.08 (.016)	.08 (.008)
	.5 vs 1.5	-.464 (.546)	-.412 (.214)	-.397 (.135)	-.386 (.082)	-.382 (.056)	-.38 (.025)	-.379 (.013)

Note. Standard errors are in parenthesis.

When the difference in the Z_1 latent intercepts was 1, the estimated values of the Z_1 unique variance were negative with high variability in small sample sizes, thus some samples had a positive estimate. However, as the sample size increased, there was less variability in the estimates and as a consequence all the

values were negative. This explains that the percentage of samples with Z_1 negative unique variances increased as the sample size increased when the difference in the Z_1 latent intercepts was 1, as shown in Table 16.

Chapter 5

Discussion and Conclusions

Given the importance of measurement and predictive invariance in psychological testing, some studies have been conducted to examine the relationship between them (Millsap, 1995, 1997, 1998). The present research focuses on the relationship between both forms of invariance in the presence of group differences in the regression intercepts.

The common interpretation of group differences in regression intercepts is that they represent differences in the population and do not represent measurement bias (Humphreys, 1986; Linn, 1984). This interpretation is supported by the fact that group differences in the regression intercepts can exist under factorial invariance as shown by Birnbaum (1979). However, Millsap (1998) showed that this interpretation is true only under some restricted conditions, and proposed a method to test them. The method consists of fitting a factorial invariant model with invariant factor variances to the data; if the model fits the data then there is no evidence that the differences between groups are reflecting measurement bias, and the regression intercept differences can be explained as a consequence of having fallible measures in the predictor and criterion, or as true population differences. In this case, the traditional interpretation of the regression intercepts would be supported. However, if the model fails to fit then, the differences between groups are due to measurement bias.

Since it cannot be assumed that group differences in the regression intercepts are only reflecting population differences, and that the conditions for this statement to hold have to be tested, it is important to determine the power to detect violations to the model. The purpose of the present research was to study Type I errors and the statistical power of the method proposed by Millsap (1998) under different sample sizes when 1 or 2 predictors are available.

The results indicate that while Type I errors are within appropriate values when p=1 and p=2, the statistical power of the model in the one predictor case was almost non-existent regardless of the sample size. An interesting result when fitting an invariant model to data created with group differences in the case of p=1 was the presence of a large percentage of negative unique variances in the predictor that increased with sample size.

Based on the results of this research, it is proposed that with sample sizes of at least 500 the negative unique variances in the predictor can be used as an indication of violations to the invariant model with a power larger than .8 and Type I errors of .05. This proposal is supported by the fact that when the unique variance of the predictor is constrained to be a positive value, the power of the chi-square fit statistic increases but it requires higher sample sizes to detect violations to the invariant model than when using the negative unique variances as indicators of bias. As a consequence, the negative unique variances are considered a better way to detect violations to the invariant model.

In order to explain the negative unique variances the distribution of all the sample estimates were examined. It was found that the differences in the latent intercepts started a chain reaction that affected the parameters in the mean and the covariance structures, eventually leading to the negative unique variances in the predictor.

In the two predictor case the statistical power of the model was highly improved. With a sample size of 100 the power to detect violations to the invariant model is .9 when the difference in the latent intercepts is .5. The negative unique variance estimate in the predictor in this case is not as good an indicator of measurement bias as the chi-square statistic test. For example, with a difference of .2 in the latent intercepts and a sample size of 100 the negative unique variance indicate violations to the model in only 5% of the samples while the chi-square detect them in 89% of the samples.

The examples reported in Millsap (1998) are consistent with the findings of the present research. In both of the examples with one predictor in which the factorial invariant model was rejected, the model failed to achieve convergence after 1000 iterations and negative unique variances were obtained for the predictor. As expected from the results of the present research, with the large sample sizes used these examples, 9748 and 68,766 individuals, the negative unique variances are a good indicator of the violations to the factorial invariant model. Also, when the reverse regression was conducted in these examples, that is, when the roles of the criterion and the predictor were reversed, an inconsistent

pattern was found: the group with the higher intercept in the forward regression did not have the higher intercept in the reverse regression, indicating violations to the factorial invariant model. These examples show that researchers have to be aware of different indications of violations to invariance, and not focus exclusively in the chi-square fit statistic.

It is important to note the limitations of the present study. The first limitation is that the data were simulated under a normal distribution. The behavior of the model with data that are not normal is still to be studied. A second limitation is that the data were simulated with equal sample sizes in both groups. Since it is often the case that the groups studied have different sample sizes, future research could study its impact in the Type I errors and power of the model; if having groups with different sample sizes changes the results of the present study it would be important to determine how large the difference in the sample sizes between the groups must be to start detecting changes in the results reported in the present study.

One more limitation is that only a specific set of parameters values were studied. The parameters values were picked to reflect values usually found in literature, however, it would be worthwhile to study values that reflect other set of communalities or reliabilities in the predictors and criterion.

An area of future research will deal with better ways of detecting violations to measurement bias in the presence of differences in the regression intercepts when there is only one predictor of interest and with sample sizes

smaller than 500. Before such a method is found, the statistical power of the model can be improved by increasing either the sample size or the number of predictors. If the availability of individuals is limited, for example by the number of employees in a company, it is still possible to increase the number of predictors. Millsap (1998) explained that if the predictor is a multi-item summative scale it is possible to disaggregate the tests into parcels of items, and to use the score of each parcel as a different indicator. The model would then be tested at the parcel level but conclusions can be drawn to the overall test. If the test shows violations to the invariant model, these violations will show up at the parcel level. In the same sense, if the model holds at the parcel level then the model must also hold at the test level. However, it should be noted that if measurement bias is found at the parcels level, it can be argued that the bias can be canceled out if the whole test is analyzed. In this case, finding evidence of violations to measurement invariance at the parcels level would lead to the wrong conclusion that the whole test is biased, so special caution is advised when conducting the test at the parcels level. The advantage of having more than one predictor is that it is possible to detect the size and the source of measurement bias. With only one predictor it is not possible to free parameters to identify the source of the bias because the model would not be identified.

The results of the present study show that researchers have to be aware of the importance of testing for measurement invariance before assuming that

differences in regression intercepts are due to population differences. To ensure that no measurement bias is present both the chi-square statistic and the unique variances have to be examined. If negative unique variances are detected or if the chi square statistic indicates that the factorial invariant model is rejected, measurement bias is present in the data. The bias could be due to differences in the latent intercepts, unique variances, or factor loadings; if the differences are in the factor loadings then violations to the invariant regression slopes will be present. On the other hand, if the model fits the data it can be concluded that there is no evidence of measurement bias. The difference in the regression intercepts can be interpreted as latent mean differences, or as a consequence of having fallible measures for the criterion and predictors.

REFERENCES

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

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.

Borsboom, D. (2004). When does measurement invariance matters? *Medical care, 44*, 176-181.

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. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.

Gottfredson, L. S. (1988). Reconsidering fairness: A matter of social and ethical priorities. *Journal of Vocational Behavior, 33*, 293-319.

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

Humphreys, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal of Applied Psychology, 71*, 327-333.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes job knowledge, and job performance. *Journal of vocational behavior, 29*, 340-362.

Jensen, A. R. (1992). Spearman's hypothesis: Methodology and evidence. *Multivariate Behavioral Research, 27*, 225-233.

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

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149.

Mellenbergh, G. J. (1989). Item bias and Item Response Theory. *International Journal of Educational Research, 13*, 127-143.

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. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461-473.

Muchinsky, P. (1993). Validation of intelligence and mechanical aptitude tests in selecting employees for manufacturing jobs. *Journal of Business and Psychology*, *7*, 373-382.

Muthén, L.K., & Muthén, B.O. (2009). M*plus* (Version 5.21). Los Angeles: Muthén & Muthén.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79*, 518-524.

Sackett, P. R., Schmitt, N., Ellingson, J. E., Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. *American Psychologist*, *56*, 302-318.

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

Stark, S., Chernyshenko, O. S. & Drasgow F. (2006). Detecting Differential Item Functioning with Confirmatory Factor Analysis and Item Response Theory: Toward a unified strategy. *Journal of Applied Psychology, 91*, 1292-1306.

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

This document was generated using the Graduate College Format Advising tool. Please turn a copy of this page in when you submit your document to Graduate College format advising. You may discard this page once you have printed your final document. DO NOT TURN THIS PAGE IN WITH YOUR FINAL DOCUMENT!

68

68