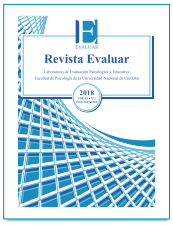
2018, Vol. 18, No 2.
ISSN 1667-4545
Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Laboratorio de Evaluación Psicológica y Educativa
Facultad de Psicología - Universidad Nacional de Córdoba



The Use of Measurement Invariance with Dichotomous Variables as Evidence of Validity

El uso de la invarianza de medición con variables dicotómicas como evidencia de validez

Luis Rojas * [1] , Guaner Rojas [1] , Armel Brizuela [1]

1 - Instituto de Investigaciones Psicológicas, Universidad de Costa Rica.

Recibido: 19/02/2018 **Revisado:** 15/03/2018 **Aceptado:** 18/03/2018

Resumen

En este artículo se explica y ejemplifica cómo se puede implementar un análisis de invarianza de medición con ítems dicotómicos con el fin de obtener evidencias de validez. El ejemplo fue desarrollado con datos de la Prueba de Aptitud Académica (PAA; PPPAA, 2014) de la Universidad de Costa Rica (UCR). El análisis fue desarrollado según género y según tipo de colegio (público y privado). El método usado fue un análisis factorial confirmatorio con dos factores correlacionados definidos por los dos componentes de la PAA (N = 11304). En cada análisis grupal se alcanzó la invarianza estricta. A partir de lo anterior, se concluyó que las puntuaciones del test presentan la misma unidad de medida en hombres y mujeres, al igual que en estudiantes de colegios privados y públicos. Finalmente, estos resultados constituyen una evidencia de validez de las puntuaciones de la PAA en términos de su estructura interna, desde la perspectiva de la equidad.

Palabras clave: *Validez, equidad, invarianza de medición, ítems dicotómicos, Prueba de Aptitud Académica*

Abstract

This paper explains and exemplifies how an analysis of the measurement invariance with dichotomous items may be implemented in order to obtain evidence of validity. The example was developed with data from the Academic Aptitude Test (PAA, for its initials in Spanish; PPPAA, 2014) at the University of Costa Rica (UCR). Measurement invariance was analyzed according to two classifications, by high school kind (public or private) and by sex. The method used was a confirmatory factor analysis with two correlated factors defined from the two components of the PAA (N = 11304). Each group analysis presented strict invariance. It was concluded that the test scores have the same unit of measurement for both men and women, and among students from public and private high schools. Finally, these results constitute evidence of validity of the PAA scores in terms of their internal structure from the perspective of fairness.

Keywords: *Validity, fairness, measurement invariance, dichotomous items, Academic Aptitude Test*

***Correspondencia a:** Luis Rojas, correo electrónico: luismiguel.rojas@ucr.ac.cr

Cómo citar este artículo: Rojas, L., Rojas, G., & Brizuela, A. (2018). The use of measurement invariance with dichotomous variables as evidence of validity. Revista Evaluar, 18(2), 45-58. Recuperado de https://revistas.unc.edu.ar/index.php/revaluar

Introduction

Fairness in the use of the results from the application of a test is one of the fundamental evidences of validity in the field of measurement. Fairness implies that test scores provide valid interpretations for specific uses regardless of the characteristics of individuals under examination and their evaluation context. Fairness in the specific domain of a test is achieved when the test evaluates the same construct for all the individuals examined; meanwhile, their scores have the same meaning for all individuals in the population of interest (AERA, APA, & NCME, 2014). Fairness in measurement can be studied by creating subgroups with common characteristics that ensure that a test evaluates the same constructs and meaning of scores in the established subgroups.

Neglecting equity may base the evaluation on biased tests and lead researchers to wrong conclusions. Biased tests present items that define different underlying constructs that are not comparable in population (Brown, 2006). The fact that constructs are not the same could be a consequence of different perceptions in the subgroups; this could also mean that the scale used to measure the construct doesn't work the same way on each subpopulation. The latter condition creates some systematic biases against a specific population. For instance, in a test biased against men, if a man and a woman have the same level in the evaluated construct, the men might display a lower score than the one obtained by the women.

The previous ideas reflect the need to generate evidence of fairness in terms of validity for the use of test scores. Evidence of validity can be obtained by analyzing measurement invariance, which will be explained later. For example, by testing measurement invariance, it was concluded that for the Scholastic Aptitude Test (SAT) there is no evidence that suggests that scores derived from this test have different interpretations for students with or without specific learning adaptations (Hartwig & Gregg, 2007). The analysis of invariance was also applied to confirm that the Graduate Record Examination (GRE) measures the same construct with the same units across different sex and ethnic populations such as white men, black men, white women and black women (Rock, Werts, & Grandy 1981). Moreover, some analyses implemented by Rock et al. (1981) have also generated evidence of validity related to equity, for several scales associated with psychological constructs (Cumsille, Martínez, Rodríguez, & Darling, 2014; Guedea, Ornelas, Rodríguez, & Gastélum, 2012; Piqueras, Olivares, Vera-Villarroel, Marzo, & Kuhne, 2012).

The purpose of this study is to demonstrate how a measurement invariance analysis can generate evidence of validity from the perspective of fairness and to explain how to implement this analysis with dichotomous items. The data from the Academic Aptitude Test (PAA, for its initials in Spanish) at the University of Costa Rica (UCR) are deployed in order to support this claim (Programa Permanente de la Prueba de Aptitud Académica [PPPAA], 2014). This test is used for the selection of new students at this university and the groups in which measurement invariance will be analyzed are laid down by sex and high school kind (public or private) where PAA test takers graduated.

Background

The Measurement Invariance

Measurement Invariance refers to the fact that scores obtained while assessing a variable must be unrelated to other characteristics that are not intended to be measured (Millsap,

2007). In this sense, the assessment refers to the scores collected in observed variables, and the mathematical definition of invariance establishes what properties should remain invariant.

The mathematical definition of invariance is represented by the following formula:

$$P(\boldsymbol{X_i}|\,\boldsymbol{\xi_i}, \boldsymbol{V_i}) = P(\boldsymbol{X_i}|\,\boldsymbol{\xi_i})$$

where $\boldsymbol{X}_i = (X_{ij})$ is a $q \times 1$ vector of variables that represents the scores in the observed variables (measurements) in the i^{th} person of the population, $\boldsymbol{\xi}_i = (\xi_{ij})$ is a $r \times 1$vector of factor scores for that person and $\boldsymbol{V}_i = (V_{ij})$ is a $s \times 1$ vector of variables observed for the same person that define characteristics of the population that should be irrelevant to $\boldsymbol{X}_i$, taking into consideration $\boldsymbol{\xi}_i$. Thus, measurement invariance for $\boldsymbol{X}$ in relation to $\boldsymbol{\xi}$ and $\boldsymbol{V}$ is equivalent to saying that for all $\boldsymbol{X}$, $\boldsymbol{\xi}$ and $\boldsymbol{V}$, the conditional probability of $\boldsymbol{X}$ given $\boldsymbol{\xi}$ and $\boldsymbol{V}$ is equal to the conditional probability of $\boldsymbol{X}$ given $\boldsymbol{\xi}$ only (Millsap, 2007).

For example, in a test designed to measure the latent variable $\boldsymbol{\xi}$, it is important to determine the item measurement invariance in relation to the latent variable $\boldsymbol{\xi}$ and the variable $\boldsymbol{V}$, as the latter represents measurement characteristics that are irrelevant to the variable $\boldsymbol{\xi}$ under assessment. Measurement invariance is achieved when all possible cases of scores in the items, levels of $\boldsymbol{\xi}$ and levels of $\boldsymbol{V}$ comply to the following rule: the probability of obtaining one specific score considering a certain level of $\boldsymbol{\xi}$ and a certain level of $\boldsymbol{V}$ equals the probability of obtaining the same score considering only the same certain level of $\boldsymbol{\xi}$.

Confirmatory factor analysis (CFA) allows researchers to implement an analysis of measurement invariance. It is known that under the CFA, the vector of observed variables $\boldsymbol{X}_i$ is modeled as follows:

$$\boldsymbol{X_i} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi_i} + \boldsymbol{\varepsilon_i},$$

in which the expected value and covariance matrix are represented respectively:

$$\mathrm{E}(\boldsymbol{X_i}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\kappa}$$
$$\mathrm{Var}(\boldsymbol{X_i}) = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_{ij})$ is the $q \times 1$ vector of error terms for the i^{th} person in the population; $\boldsymbol{\Lambda} = (\lambda_{jj'})$ is the $q \times$ r matrix of regression coefficients of $\boldsymbol{X}$ on $\boldsymbol{\xi}$; $\boldsymbol{\tau} = (\tau_j)$, the $q \times 1$ vector of intercepts; $\boldsymbol{\kappa} = (\kappa_j)$ is the $r \times 1$ vector of latent means; $\boldsymbol{\Phi} = (\varphi_{jj'})$ is the $r \times r$ covariance matrix of the latent variables, and $\Theta = (\theta_{jj'})$ is the $q \times q$ covariance matrix of errors terms.

Moreover, it is known that the estimation method most frequently used for the CFA is maximum likelihood, which assumes that the observed variables stem from a multinormal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, corresponding to the mean vector and covariance matrix of the observed variables (Kaplan, 2009). This implies that P $(\boldsymbol{X}_i = \boldsymbol{x} \mid \boldsymbol{\xi}_i)$ depends only on the mean vector and the covariance matrix of $\boldsymbol{X}_i$ given $\boldsymbol{\xi}_i$, which can be reduced to

$$\mathrm{E}(\boldsymbol{X_i}|\boldsymbol{\xi_i}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi_i}$$
$$\mathrm{Var}(\boldsymbol{X_i}|\boldsymbol{\xi_i}) = \boldsymbol{\Theta}$$

From the previous equation it is concluded that in order to obtain the invariance measurement on $\boldsymbol{V}$, it is necessary to guarantee that within the groups defined by $\boldsymbol{V}$, the regression coefficients ($\boldsymbol{\Lambda}$), the intercepts ($\boldsymbol{\tau}$) and the variances of error terms ($\boldsymbol{\Theta}$) remain constant; since for any value of $\boldsymbol{V}$ it should be true that:

$$\mathrm{E}(\boldsymbol{X_i}|\boldsymbol{\xi_i}, \boldsymbol{V_i}) = \mathrm{E}(\boldsymbol{X_i}|\boldsymbol{\xi_i}) = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi_i}$$
$$\mathrm{Var}(\boldsymbol{X_i}|\boldsymbol{\xi_i}, \boldsymbol{V_i}) = \mathrm{Var}(\boldsymbol{X_i}|\boldsymbol{\xi_i}) = \boldsymbol{\Theta}$$

Now, assuming that $\boldsymbol{V}$ defines K groups in the population, the conditions for invariance have the following formula

$$\mathrm{E}(\boldsymbol{X}_{ik}|\boldsymbol{\xi}_{ik}) = \boldsymbol{\tau}_{\boldsymbol{k}} + \boldsymbol{\Lambda}_{\boldsymbol{k}}\boldsymbol{\xi}_{\boldsymbol{ik}} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi}_{\boldsymbol{ik}}$$
$$\mathrm{Var}(\boldsymbol{X}_{ik}|\boldsymbol{\xi}_{\boldsymbol{ik}}) = \boldsymbol{\Theta}_{\boldsymbol{k}} = \boldsymbol{\Theta}$$

where $\boldsymbol{X}_{ik} = (X_{ijk})$ represents the vector of observed variables for the i^{th} individual in the k^{th} group defined by $\boldsymbol{V}$, with k in $\{1, 2,.., k\}$; $\boldsymbol{\xi}_{ik} = (\xi_{ijk})$ are the factor scores for the same individual in the kth group, and $\boldsymbol{\tau}_k = (\tau_{jk})$, $\Lambda_k = (\lambda_{jj'k})$ and $\boldsymbol{\Theta}_k = (\theta_{jj'k})$ are the intercepts vector, the vector of regression coefficients and the matrix of residual variance for the k^{th} group.

A sequence of nested models is analyzed in order to examine the equality of parameters involved in measurement invariance (Meredith, 1993). First, similarity of factorial structure among groups is assessed (configural invariance). Second, equality of regression coefficients between groups is assessed (weak invariance); subsequently, equal intercepts (strong invariance) and equal error variances (strict invariance) are evaluated.

Configural Invariance

Configural invariance implies the correct fit of the theoretical model in the established groups, regardless of the model fit coefficients for each group. This level of invariance indicates that the latent variables in each group are similar, but not identical (Widaman & Reise, 1997). In other words, members in different groups conceptualize constructs in the same way (Milfont & Fischer, 2010).

The previous references indicate that configural invariance evidences that similar latent variables were assessed in each group. It is worth mentioning that "To ensure that the same construct is being measured in different groups, measurement invariance measure is necessary but not sufficient" (Chen, 2007, p. 465). In order to find more evidence for the evaluation of the same construct, other criteria of validity should be applied. The failure to find configural invariance invalidates any possible comparison between groups regarding a specific construct. This result would show that the observed variables are indicators of different constructs within the established groups (Hortensius, 2012).

Weak or Metric Invariance

Weak invariance occurs when regression coefficients are equivalent between groups $(\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_2 = \ldots = \boldsymbol{\Lambda}_k)$. This level of invariance indicates that for any possible value of j, the latent variables $\xi_{j1}, \xi_{j2}, \ldots, \xi_{jk}$ have the same unit of measure. This means that given an indicator and a latent variable, their linear association is equal in every group (Chen, 2007; Hortensius, 2012). At this level, individual scores cannot be compared since, for some latent variables, the origin (starting value in the scale) may differ between groups. The only aspect weak invariance allows to conclude about is that the association between skill level for the assessed construct and scores on observed variables is not related to the group the individual belongs to.

Figure 1 shows the graph of equations of an observed variable X_j comprised in a test that evaluates the construct ξ_1, as this variable shows weak invariance between groups A ($k = 1$) and B ($k = 2$). The linear association between ξ_1 and the observed variable X_j is equivalent between groups. However, for a given level of ξ_1, the observed variable shows different scores observed for each

group; which is evidenced by the difference of the intercepts (τ_{j1} and τ_{j2}).

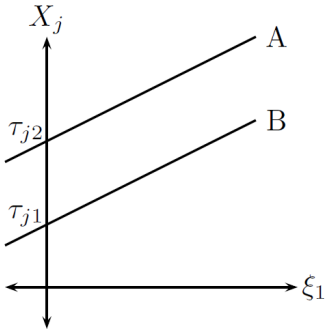


Figure 1
Example of Weak Factorial Invariance.

Strong or Scale Invariance

Strong Invariance is obtained when regression coefficients and intercepts are equal between groups ($\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \ldots = \mathbf{\Lambda}_k$ and $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \tau_k$). This level indicates that for any possible value of *j*, the latent variables ξ_{j1}, ξ_{j2}, …, ξ_{jk} have the same measurement scales between groups (same origin and unit of measure).

Thus, with strong invariance, individuals who would get the same scores on the latent constructs tend to get the same scores in the observed variables, regardless of the group they belong to (Milfont & Fischer, 2010). For this level of invariance, it is concluded that differences in the latent group means are reflected in the averages of the observed variables (Cheung & Rensvold, 2002).

The example previously analyzed in Figure 1 is shown in Figure 2, but now strong factorial invariance is assumed. In this case the linear equation for X_j is equal for both groups, indeed there appears only one τ_j intercept, since it is the same for A and B. In addition, the expected values in the latent variable are presented: κ_{11} and κ_{12}, and the observed variable is: μ_{j1} and μ_{j2}, for groups A and B, respectively. These values exemplify how the means' difference of the latent variables is reflected in the means of observed variables.

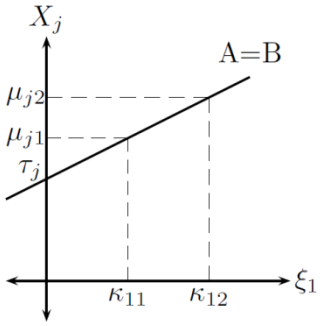


Figure 2
Example of strong factorial invariance.

This particular level of invariance is important in the construction of a test, since it indicates that the inferences made based on its scores are not biased by population groups. In addition, the presence of strong factorial invariance implies that the scores of individuals belonging to different established groups are comparable.

Strict or Residual Invariance

Residual Invariance is achieved when regression coefficients, intercepts and residual variance are equal between groups ($\mathbf{\Lambda}_1 = \mathbf{\Lambda}_2 = \ldots = \mathbf{\Lambda}_k$, $\boldsymbol{\tau}_1 = \boldsymbol{\tau}_2 = \ldots = \boldsymbol{\tau}_k$ and $\mathbf{\Theta}_1 = \mathbf{\Theta}_2 = \ldots = \mathbf{\Theta}_k$). This level of invariance indicates that the differences in observed variables exist due to differences in the latent variables (Chen, 2007). The evaluation of this invariance level presents several difficulties including that, generally, there is a variation of error variance according to the levels of evaluated constructs, contrary to the assumption of confirmatory factor analysis regarding the

independence between latent variables and errors (Kaplan, 2009; Widaman & Reise, 1997).

The step of evaluating this level of invariance can be omitted, since the information it provides is usually not necessary for the evaluation of hypotheses that led to the analysis of measurement invariance (Brown, 2006). For the most common uses of factorial invariance it is enough to achieve strong factorial invariance, since interpretations about unbiased observed scores and comparison of parameters associated with the latent variables can be performed without the need to reach the level of strict invariance (Chen, 2007).

Evaluation of Invariance

To evaluate configural invariance, a model that fits the data is required both individually for each group and in a multigroup analysis. Classical fit indices for CFA are used to evaluate these models (Brown, 2006). Next levels of invariance are evaluated by comparison of their adjustment indices to those of previous levels. Based on a simulation study, Chen (2007) recommends comparing the root mean square error of approximation (RMSEA; Hu & Bentler, 1999) and the comparative fit index (CFI; Bentler, 1990), being the last one among the most recommended in such studies.

Comparisons of fit indices consider the difference between the evaluated model indices and those of a previous evaluated invariance model (Δ index). The following cutoff values are recommended for large sample sizes (n > 300) as a criterion to accept the level of invariance in evaluation: ΔCFI > -.010 and ΔRMSEA < .015 (Chen, 2007).

Invariance in Models with Dichotomous Observed Variables

The basic equation for CFA with dichotomous latent variables is

$$\boldsymbol{X}_i^* = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i$$

where all the elements are defined similarly to the general formula of CFA. The additional element in this equation is the vector $\boldsymbol{X}_i^* = (X_{ij}^*)$, which represents scores underlying observed dichotomous variables for the i^{th} individual. $\boldsymbol{X}^*$ is a variable that is distributed as a multivariate normal with mean $\boldsymbol{\mu}^*$ and covariance matrix $\boldsymbol{\Sigma}^*$. The variable X_j^* represents an underlying variable to X_j by the relationship

$$X_{ij} = \begin{cases} 1, & \text{si } X_{ij}^* \geq \nu_j \\ 0, & \text{si } X_{ij}^* < \nu_j \end{cases}$$

where ν_j is a real number called *threshold*.

Models with dichotomous dependent variables tend to model the probability of success for each observed variable (usually $X_{ij} = 1$). In this case, the model assumptions permit for it to conclude the probability of success of X_{ij} given ξ_i

$$\mathrm{P}(X_{ij} = 1|\boldsymbol{\xi}_i) = \Phi(\tau_j + \boldsymbol{\Lambda}_j\boldsymbol{\xi}_i - \nu_j)$$

where $\boldsymbol{\Lambda}_j$ is the j^{th} row of $\boldsymbol{\Lambda}$ and Φ represents the probability function of the normal distribution. Then, by applying the inverse of Φ (probit function) in both sides of equality, one gets the probit of success probability, which is associated linearly with equation $\tau_j + \boldsymbol{\Lambda}_j\boldsymbol{\xi}_i - \nu_j$. In this way, a pragmatic approach is applied to the interpretation of coefficients model CFA with ordinal variables (Kosiol, 2010; Sideridis, Tsaousis, & Al-Harbi, 2015).

In a CFA in which the items are indicators

of a single factor, the conditions for identification are: that the variables underlying the observed variables be standard normal, and that in each latent variable, one of its indicators possess a regression coefficient constrained to 1. For multigroup models, conditions for identification are (Millsap & Yun-Tein, 2004):

1. That within group 1, latent variable means be constrained to 0 and the variances of the underlying variables be constrained to 1.

2. That for each latent variable, an indicator, whose regression coefficient will be 1 in all groups, be chosen. Additionally, the variance of variables underlying these indicators must be constrained to 1.

3. That all intercepts be 0 ($\tau_1 = \tau_2 = \ldots = \tau_k = 0$).

4. That thresholds be equal across the groups ($\nu_1 = \nu_2 = \ldots = \nu_k$).

Another way of identifying the multigroup model is given by theta parameterization (Muthén & Muthén, 1998), which involves changing the previous condition for the variances of underlying variables as presented in 1) and 2) for the restriction that the associated error variances are constrained to 1. Theta parametrization is highly recommended for dichotomic items (Pornprasertmanit, 2015).

Furthermore, in the analysis of strong invariance with categorical variables, it is usually the hypothesis of equal threshold that is evaluated, rather than that of intercepts (Kosiol, 2010). However, for dichotomous items it is not possible to evaluate the hypothesis of strong factorial invariance, since one of the restrictions established for the identification of a multigroup model is the constraining of thresholds; thus, in dichotomous items three levels of invariance models are evaluated: Configural, Weak and Strict.

Method

Participants

The data for this study were extracted from the PPA application corresponding to the academic year of 2015, which had the purpose of selecting applicants for the academic year of 2016. Four parallel formulas of PPA were distributed proportionally and randomly among all applicants, who were sitting in several classrooms. Considering this application design, it was inferred that data from a single form is sufficient for this study purposes; the selected formula was Formula 1.

Thus, the sample comprised all the applicants for admission at UCR in 2015 who responded to the items of Formula 1 of the PAA. This sample consisted mostly of secondary education seniors. Moreover, this sample represents most of the geographical areas of Costa Rica, due to the efforts made by the UCR in terms of access to the admission system.

The total sample consisted of 11304 individuals (45% men, n = 5091; 55% women, n = 6213). 81% (n = 9175) came from public schools, while 19% (n = 2129) came from private schools.

Instrument

The PAA is a high-stake standardized test that has been applied since 1960. Its construction, analysis, application and psychometric quality are based on modern standards (AERA, APA, & NCME, 2014). As mentioned above, the PAA is a test designed to select new students at UCR. This use of its scores is supported by several evidences of validity (Jiménez & Morales, 2010; Montero-Rojas, Villalobos-Palma, & Valverde-Bermúdez, 2014; Rojas-Torres, 2013).

This test consists of two sections: reasoning in mathematical context (RMC) and reasoning in verbal context (RVC). The sections use selected response items with five options. The time granted for the resolution of the test was three hours. The participants responded to 50 items of RVC and 35 items of RMC. Fifteen of those items (10 for RVC and 5 for RMC) had not been previously used, so it was decided to leave them out of the analysis, as well as 2 items pertaining RMC that did not meet the psychometric qualities necessary to stay on the PAA item bank. Finally, 40 items from RVC and 28 items from RMC were effectively used for this study, Cronbach's alpha coefficients for RVC and RMC were .89 and .84, respectively.

Procedure

The CFA model to be estimated is shown in Figure 3, this factor structure is based on Rojas-Torres's (2014) work, which concluded that the PAA fits a factorial structure defined by two highly correlated latent variables: *reasoning in verbal context* and *reasoning in mathematical context*. This model was estimated, first in samples defined by the sex of the examinee and then in both groups simultaneously (configural invariance model). Later, the model was estimated considering the restrictions in order to assess weak invariance and, subsequently, strict invariance.

The analysis, according to high school kind, followed the same procedure as the adopted in the analysis of invariance by sex. Now, Bollen (1989) indicates that it is recommended, for the analysis of invariance, that contrasting groups possess similar sample sizes, a requirement which is not met in this division of sample, since most of examinees come from public schools. For this reason, an additional analysis was conducted using a sample of examinees from public schools of the same size that the private schools sample.

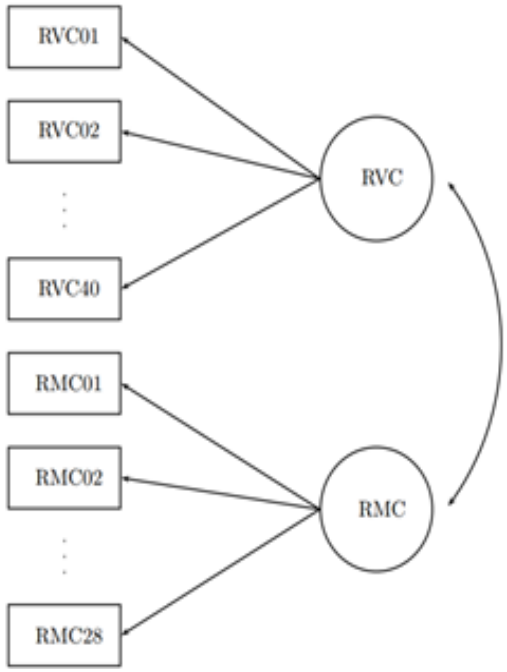


Figure 3
CFA Model.

Data Analysis

Model estimation was based on a matrix of tetrachoric correlations since the items have a dichotomous nature. The estimating method for CFA was weighted least squares adjusted for the mean and variance (WLSMV), recommended for use with dichotomous items (Millsap & Yun-Tein, 2004).

Analyses were performed with the software Lavaan 0.5 -18 (Rosseel, 2012), designed within R (R Core Team, 2014). Conditions stated by Millsap and Yun-Tein (2004) were used to identify the proposed model, while introducing the variant of theta parametrization as suggested by Múthen and Múthen (1998).

For the analysis of the configural invariance, model fit was evaluated by the following criteria: CFI (Bentler, 1990) greater than or equal to .95,

and a RMSEA (Hu & Bentler, 1999) less than or equal to .05. The subsequent levels of invariance (weak and strict) were analyzed based on the criteria suggested by Chen (2007).

Results

Table 1 shows indices associated to the analysis of invariance by sex. It can be seen that the model has an acceptable fit for both men (RMSEA = .010, CFI = .996) and women (RMSEA = .010, CFI = .996). In each of the models, the regression coefficients were statistically different from zero ($p < .05$) and their completely standardized value was greater than .30. This suggests that the indicators are associated with the corresponding latent variables. The configural invariance model also presented an acceptable fit (RMSEA = .010, CFI = .996); consequently, it can be concluded that this level of invariance is achieved in comparison by sex.

In addition, fit contrast indices between the weak invariance model and the configural invariance model indicated that weak invariance is achieved (ΔRMSEA = .005, ΔCFI = -.005). Then, strict invariance was evaluated with positive results (ΔRMSEA = .001, ΔCFI = -.001).

The number of degrees of freedom of the models estimated in specific groups is 2276, which correspond to 2415 not redundant coefficients in the matrix of tetrachoric correlations of observed variables (69 * 70/2) minus 139 estimated parameters (67 regression coefficients, 69 thresholds, 2 latent variances and one latent covariance. The intercepts, latent means and variances of error are constrained according to criteria mentioned above.). For the configural invariance model, as it considers two groups, there are no redundant pieces of information about the correlation matrix (2415 * 2 = 4830). Consequently, the parameters to be estimated in the first group are 139 (67 regression coefficients, 69 thresholds, 2 latent variances and 1 latent covariance), while in the second, they are 139 (67 regression coefficients, 2 latent variances, 1 latent covariance, 2 latent means and 67 error variances. The thresholds are not considered because they are constrained to those in the first group). Therefore, the operation 4830 - 139 * 2 provides 4552 degrees of freedom. In the weak invariance model, 67 regression coefficients are constrained, so the degrees of freedom increase to 4619. In the strict invariance model, 67 error variances are constrained, which is the reason why the degrees of freedom increase to 4686.

Table 1
Assessment of measurement invariance in the PAA according to sex.

	χ^2	*df*	$\Delta\chi^2$	Δdf	*RMSEA*	Δ*RMSEA*	*CFI*	Δ*CFI*
Fit in specific groups								
Men (n = 5091)	3489.86**	2276			.010		.996	
Women (n = 6213)	3812.38**	2276			.010		.996	
Fit in multi-group models and contrast to the previous model								
Configural Invariance	7302.24**	4552			.010		.996	
Weak Invariance	10717.46**	4619	3415.22**	67	.015	.005	.991	-.005
Strict Invariance	11520.78**	4686	803.32**	67	.016	.001	.990	-.001

Note. ** $p < .05$

Table 2 shows the results associated with the analysis of invariance, according to kind of high school (public or private). The models that were estimated with specific groups in their original sample sizes showed acceptable fit to the data (public: RMSEA = .011, CFI = .994; private: RMSEA = .010, CFI = .996). Similarly, for a public school sample with an adapted sample size, the model showed an acceptable fit (RMSEA = .009, CFI = .997). These three models presented significant regression coefficients ($p < .05$) and completely standardized coefficients were higher than .30.

Table 2
Evaluation of measurement invariance of the PAA according to the high school kind.

	χ^2	*df*	$\Delta\chi^2$	Δdf	*RMSEA*	Δ*RMSEA*	*CFI*	Δ*CFI*
Fit in specific groups								
Private (n = 2129)	2782.24**	2276			.010		.996	
Public (n = 9175)	4795.07**	2276			.011		.994	
Public* (n = 2129)	2564.42**	2276			.008		.997	
Fit in multi-group and contrast models with the previous model (different sizes)								
Configural Invariance	7523.32**	4552			.011		.995	
Weak Invariance	10073.74**	4619	2550.42 **	67	.014	.003	.990	-.005
Strict Invariance	10615.13**	4686	541.39 **	67	.015	.001	.989	-.001
Fit in multi-group models and contrast to the previous model (equal sizes)								
Configural Invariance	5292.66**	4552			.009		.997	
Weak Invariance	7095.27**	4619	1802.61 **	67	.016	.007	.988	-.009
Strict Invariance	7445.19**	4686	349.92 **	67	.017	.001	.987	-.001

Note. *Sampled population ** $p < .05$

Subsequent levels of invariance were firstly estimated with original sample sizes. Strict invariance was tested through RMSEA and CFI bearing good/acceptable outcomes (for RMSEA: configural invariance: RMSEA = .011; weak invariance: ΔRMSEA = .003, strict invariance: ΔRMSEA = .001; for CFI: configural invariance: CFI = .995; weak invariance: ΔCFI = -.005, strict invariance: ΔCFI = -.001). Similarly, by analyzing invariance with equal sample sizes, it was concluded that the PAA presents strict invariance according to kind of high school (configural invariance: RMSEA = .009, CFI = .997; weak invariance: ΔRMSEA = .007, ΔCFI = -.009; strict invariance: ΔRMSEA = .001, ΔCFI = -.001).

Discussion

The results obtained in the previous section imply that the PAA presents strict invariance by sex and high school kind.

Two different analyses led to the same conclusion, one of these was carried out with the original sample sizes while the other consisted of equivalent sample sizes. To analyze the strict invariance outcomes for the PAA, it is necessary

to analyze previous invariance levels (configural and weak). Invariance by sex and high school kind came to the same results, which is reason enough to present this first part of the discussion using a single criterion for defining groups (sex), and understanding that the conclusions are similar to those of the groups defined by other criteria (high school kind).

Configural invariance leads to the conclusion that the PAA evaluated similar constructs in men and women. Thus, invariance analysis interpretation provides evidence to conclude that RMC is conceptualized by men similarly to the conceptualization made by women; it happens in the same way with RVC (Milton & Fischer, 2010; Widaman & Reise, 1997). In addition, the goodness of fit of the model for each group independently indicates that the factorial structure presented in Rojas-Torres (2014) is reproduced in the same way in men and women.

Meanwhile, weak invariance results lead to the equivalence of measurement units of the evaluated constructs between sex (Chen, 2007; Hortensius, 2012). To clarify the implications of the outcomes of this study, we should consider that X1 is an indicator positively associated with RVC. Weak invariance implies that an increase of one unit in the ability RVC, generates an increase in the probits of right response in item X1, which is independent of the sex of the examinee. Generally, an increase in the skill level of RVC is associated with an increase in probit of success in each of the items of RVC, regardless of the sex of the examinee. This is analogous for items of RMC.

This result does not indicate that a man and a woman with the same skill level of RVC will get the same probit of success in the item X1. To generate this conclusion, it is necessary to obtain evidence of strong invariance, which cannot be evaluated with dichotomous items since the assumptions for strong invariance are used to identify multi-group models. Therefore, the only evidence that can be generated is that, for both groups, the evaluated constructs have the same metrics.

Then, analysis of strict invariance incorporates the condition that error variability for variables underlying the indicators is equivalent among both sexes. Thus, strict invariance indicates that the variables underlying the items show measurement invariance by sex, which means that the probit of success for the items is independent of the sex of the examinee. Up to this level, equality of thresholds, which would imply strong invariance, has indeed not been evaluated, but it has been established as a condition to constrain the metric of latent variables.

These findings generate new evidence about the validity of inferences based on PAA scores from the perspective of fairness. At the same time, these results indicate that this test evaluates the same constructs in both men and women and in students from both public and private educational institutions. It was thus concluded that the meaning of the scores in the PAA does not depend on these groups. These results also suggest that the PAA is an unbiased test for the variables of sex and kind of high school (Brown, 2006). It is noteworthy that measurement invariance in the PAA might be associated with the efforts of tests developers and researchers who have worked to ensure that the items are accessible to all people regardless of sex or kind of high school.

On the other hand, the methodology used in this work presents two limitations: the impossibility of analyzing the strong invariance and the absence of studies associated with the criteria for determining measurement invariance in ordinal variables. The major consequence of the first limitation is the absence of a strong invariance test, which bears important relevance

in the evidence of validity from the standpoint of fairness. The analysis of differential item functioning (DIF) could be a supplementary analysis to the analysis of strong invariance under the Item Response Theory framework. This model allows for the assessment of the equality of parameters of discrimination and difficulty between groups, which are related to the regression coefficients and thresholds of confirmatory factor analysis (Brown, 2006).

The second limitation is mentioned in Desa (2014), who states that models for measurement invariance with dichotomous items are adaptations of models built for another kind of variables. It is necessary to build models for dichotomous variables considering the nature of these variables. Therefore, this problem is considered to be a line of research to be developed within psychometrics.

Finally, one methodological aspect to be analyzed is the analysis of invariance with equal sample sizes. In this study, this analysis generated the same results as the one with original sizes. This result was to be expected due to the good fit of model previously evaluated in particular groups. If the model had not showed similar adjustments in the groups with original sample sizes, the analysis would have probably yielded different results (Brown, 2006).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246. doi: 10.1037//0033-2909.107.2.238

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley. doi: 10.1002/9781118619179

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. doi: 10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(2), 233-255. doi: 10.1207/s15328007sem0902_5

Cumsille, P., Martínez, M. L., Rodríguez, V., & Darling, N. (2014). Análisis psicométrico de la Escala Parental Breve (EPB): Invarianza demográfica y longitudinal en adolescentes chilenos. *Psykhe, 23*(2), 1-14. doi: 10.7764/psykhe.23.2.665

Desa, D. (2014). Evaluating measurement invariance of TALIS 2013 Complex Scales: Comparison between continuous and categorical multiple-group confirmatory factor analyses. OECD *Education Working Papers, 103*, 2-39. Paris: OECD Publishing. doi: 10.1787/5jz2kbbvlb7k-en

Guedea, J. C., Ornelas, M., Rodríguez, J. M., & G astélum, G. (2012). Invarianza factorial de la Escala de Ansiedad Asociada a la Imagen Corporal en estudiantes universitarios de educación física y ciencias de la salud. *Formación Universitaria, 5*(6), 39-50. doi: 10.4067/s0718-50062012000600005

Hartwig, J., & Gregg, N. (2007). The role of extended time on the SAT reasoning test for students with disabilities and/or Attention-Deficit/Hyperactivity Disorder. *Learning Disabilities Research & Practice, 22*(2), 85-95. doi: 10.1111/j.1540-5826.2007.00233.x

Hortensius, L. (2012). Project for Introduction to Multivariate Statistics: Measurement Invariance. Retrieved from https://pdfs.semanticscholar.org/6d69/f889d6acd9a58706e8e230e4e43ab4ae3bda.pdf

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes

in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidsciplinary Journal, 6*(1), 1-55. doi: 10.1080/10705519909540118

Jiménez, K., & Morales, E. (2010). Validez predictiva del promedio de admisión de la Universidad de Costa Rica y sus componentes. *Actualidades en Psicología, 23*(110), 21-55. doi: 10.15517/ap.v23i110.11

Kaplan, D. (2009). *Structural Equation Modeling: Foundations and Extensions* (2nd ed.). Thousand Oaks, CA: SAGE Publications. doi: 10.4135/9781452226576

Kosiol, N. A. (2010). *Evaluating measurement invariance with censored ordinal data: A Monte Carlo comparison of alternative model estimators and scales of measurement* (Master's thesis). Retrieved from https://digitalcommons.unl.edu

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*(4), 525-543. doi: 10.1007/bf02294825

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1),111-121. doi: 10.21500/20112084.857

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. doi: 10.1007/s11336-007-9039-7

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Montero-Rojas, E., Villalobos-Palma, J., & Valverde-Bermúdez, A. (2014). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: Un análisis multinivel. *Relieve, 13*(2), 215-234. doi: 10.7203/relieve.13.2.4208

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. Los Angeles, CA: Author.

Piqueras, J. A., Olivares, J., Vera-Villarroel, P., Marzo, J. C., & Kuhne, W. (2012). Invarianza factorial de la Escala para la Detección de Ansiedad Social (EDAS) en adolescentes españoles y chilenos. *Anales de Psicología, 28*(1), 203-214. Retrieved from http://www.redalyc.org/articulo.oa?id=16723161023

Pornprasertmanit, S. (2015). *Package semTools*. Retrieved in 2015 from http://cran.rproject.org/web/packages/semTools/semTools.pdf

PPPAA (2014). Prueba de Aptitud Académica. In Smith, V. (Ed.), *Compendio de Instrumentos de Medición del IIP-2014*.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Rock, D. A., Werts, C., & Grandy, J. E. (1981) *Construct validity of the GRE Aptitude Test across populations-An empirical confirmatory study* (Report 81-57). ETS. doi: 10.1002/j.2333-8504.1981.tb01284.x

Rojas-Torres, L. (2013). Validez predictiva de los componentes del promedio de admisión a la Universidad de Costa Rica utilizando el género y el tipo de sexo como variables control. *Actualidades Investigativas en Educación, 13*(1), 1-24. doi: 10.15517/aie.v13i1.11707

Rojas-Torres, L. (2014). Evidencias de validez de la Prueba de Aptitud Académica de la Universidad de Costa Rica basadas en su estructura interna. *Actualidades en Psicología, 28*(116), 15-26. doi: 10.15517/ap.v28i116.14889

Rosseel, Y. (2012). Lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. doi: 10.18637/jss.v048.i02

Sideridis, G. D., Tsaousis, I., & Al-harbi, K. A. (2015). Multi-population invariance with dichotomous measures: Combining multi-group and MIMIC methodologies in evaluating the General Aptitude Test in the Arabic language. *Journal of Psychoeducational Assessment, 33*(6), 568-584. doi: 10.1177/0734282914567871

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West

(Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association. doi: 10.1037/10222-009