Psicothema

Psicothema 2019, Vol. 31, No. 2, 194-203
doi: 10.7334/psicothema2018.303

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

An ecological view of measurement: Focus on multilevel model explanation of differential item functioning

Pamela Woitschach[1], Bruno D. Zumbo[1], and Rubén Fernández-Alonso[2]
[1] University of British Columbia and [2] University of Oviedo

Abstract

Background: The 3rd Regional Comparative and Explanatory Study reports, analyses and compares academic results in mathematics, sciences, and reading for 15 Latin American countries. Validity is the foundation of a testing procedure, and the process of validation is important to the overall success of educational assessment as a whole. This methodological study deals specifically with an ecological point of view which includes and situates the person, process, context, and time of the testing situation. These descriptions indicate specific events showing how and what variables at the individual, school, or country level can give a deep understanding of the response process. The aims were to investigate ecological explanations of differential item functioning. **Method:** The study analysed the science test given in 2013 to 6th grade students and the data pool consisted of 12,657 students from 2,609 schools and 15 countries. A progressive inclusion of the variance distribution in different Bernoulli logistic regression models was carried out. **Results:** The analyses showed the presence of differential item functioning in 32% of the science items. **Conclusions:** The main source of differential item functioning was related to the human development level of the participating countries.

Keywords: Large-scale educational testing, differential item functioning, validity, hierarchical linear model.

Resumen

Un enfoque ecológico de la medición: explicación con modelos multinivel del funcionamiento diferencial de los ítems. **Antecedentes**: el tercer Estudio Regional Comparativo y Explicativo informa, analiza y compara los resultados académicos en matemáticas, ciencias y lectura de 15 países latinoamericanos. La validez es el fundamento del procedimiento de prueba y el proceso de validación es clave para el éxito de la evaluación educativa en general. Este estudio metodológico se enfoca desde un punto de vista ecológico que sitúa a la persona, el proceso, el contexto y el tiempo donde se desarrolla la prueba. Estas descripciones señalan eventos específicos cómo y qué variables a nivel individual, escuela o país pueden dar un entendimiento profundo del proceso de respuesta. El objetivo fue investigar el funcionamiento diferencial del ítem desde un marco ecológico. **Método:** se analizó la prueba de ciencias aplicada en 2013 a los alumnos de 6º grado, los datos abarcan a 12.657 alumnos, 2.609 escuelas y 15 países. Se realizó una inclusión progresiva de niveles de distribución de la varianza en diferentes modelos de regresión logística Bernoulli. **Resultados:** los análisis mostraron la presencia de funcionamiento diferencial del ítem en el 32% de la prueba de ciencias. **Conclusión:** la principal fuente de funcionamiento diferencial del ítem se ve asociado al nivel de desarrollo humano de los países participantes.

Palabras clave: evaluación educativa a gran escala, funcionamiento diferencial del ítem, validez, modelo jerárquico lineal.

It is fundamental that there is robust evidence of validity that supports test score interpretations and uses in educational assessments. The greater the impact of test score social consequences, the higher the level of validity evidence is required to support the interpretations and uses. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) states that the major purposes for educational testing are to inform decisions about test takers, as well as to make inferences about their results and the teaching-learning process. The current uses of educational testing results, however, go beyond those purposes, especially in terms of their global significance. In addition to comparisons between countries, the results from these international educational assessments are mostly used for supervision, intervention, innovation or changes in all levels of educational policies. Zumbo (2014) made the case that although the Standards reflect a consensus about test standards and practices based in the United States of America (USA), they can be seen to play a key role in the test and assessment community globally.

Throughout the last century, the conceptualization of validity and validation have evolved through the theories and the strategies to discover and support the inferences, and through the policy implications of the evaluation process. The last version of the Standards refers to validity as the degree to which evidence and theory support the interpretations of test scores for proposed uses of the test. Meanwhile, validation is defined as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use (AERA et al., 2014, p. 11). According to AERA et al. (2014), validity is viewed as a holistic or integrated concept that includes

Received: November 2, 2018 • Accepted: February 27, 2019
Corresponding author: Pamela Woitschach
University of British Columbia
V6T1Z4 Vancouver, British Columbia
Canada
e-mail: pamelawoitschach@gmail.com

evidence from the test content, the response processes, the internal structure, the relations among and with other variables, and the social consequence of testing. In conjunction, these sources of validity evidence are synthesized on three different sets of standard procedures such as establishing intended uses and interpretations, the uses regarding samples and setting used in validation, and finally the specific forms of validity evidence.

The Standards state that differential item functioning (hereafter referred to as DIF) occurs when diverse groups of test takers with similar overall ability, or similar status on an appropriate criterion, have, on average, systematically different responses to a particular item (AERA et al., 2014, p. 16). Over the last quarter century, DIF has become a useful methodology to inform evidence of: (i) fairness and equity in testing (e.g., Elosua & Mujika-Lizaso, 2013; Cheema, 2017), (ii) internal and construct validity (e.g., Gadermann, Chen, Emerson, & Zumbo, 2018; Villegas, González, Sánchez-García, Sánchez-Barba, & Galindo, 2018), (iii) the comparability between groups and test forms (e.g., Gómez-Benito, Balluerka, González, Widaman, & Padilla, 2017), (iv) measurement invariance (e.g., Byrne & van de Vijver, 2017), and (vi) item response processes (e.g., Zumbo et al., 2015; Chen & Zumbo, 2017). Recently, Gómez Benito, Sireci, Padilla, Hidalgo, and Benítez (2018) proposed a conceptual strategy situated within the Standards that transforms DIF in an integrated validation study for all sources of evidence (instead of only evidence of validity in the internal structure). Although Gómez Benito et al., (2018) pointed out that their DIF validation proposal can be extended to educational testing; the mixed methods framework proposed did not address the complete scenario of the testing situation factors.

An alternative theory of DIF that is informed by an explanation-focused view of test validity (Zumbo, 2007a), and hence an explanation-focused view of DIF, has been developed over the last nearly 15 years. Beginning as early as 2005, Zumbo and Gelin (2005) recognized the intrinsic value of the contextual contribution to the overall response process. Positioned from Zumbo's (2007b) description of the third generation of the method, DIF is an integrated and ecological view of testing procedures in which the person does not exist as an isolated unit, and DIF analysis is focused more on the sources of contextual and holistic explanations rather than on individuals, per se. In the presence of DIF, the inferences that are made on the basis of the scale scores are not equally appropriate, useful, or meaningful across different subgroups of the target population (Zumbo, 2007a). As such, DIF methods may also aid in the investigation of the item response processes that inform test validity (Zumbo, 2007b; Zumbo et al., 2015; Zumbo & Hubley, 2017). Zumbo and Gelin's conceptual framework is the precursor to the ecological model of the item responding (Zumbo et al., 2015), which in educational assessments can include items and test characteristics, individual, classroom or school characteristics, and country factors. More recently, evidence of the impact of country characteristics can be seen in Chen and Zumbo (2017) using two-level logistic regression model with PISA data. For the discussion of multilevel logistic regression DIF involving country characteristics with psychological measures and for steps beyond DIF detection see also Gadermann et al., (2018).

Up until now, the evidence of DIF from a holistic point of view that is based on multilevel analysis includes the information of the students at the individual level and item characteristics at the nested level (Balluerka, Gorostiaga, Gómez-Benito, & Hidalgo, 2010; Balluerka, Plewis, Gorostiaga, & Padilla, 2014; Swanson, Clauser, Case, Nungester, & Featherman, 2002; van den Noortgate, & de Boeck, 2005). Given that DIF usually occurs in the context of observational rather than experimental studies, especially in educational assessments, the practice of including contextual information can address not only the sources of DIF evidence but also move towards an ecological, and even a more scientific, explanation of the item response process. Multilevel regression models can therefore expand the knowledge of DIF causes, specifying a DIF parameter that varies randomly over items and testing hypotheses on sources of DIF shared by the school and country bundles. Thus, the objective of this research is to identify the underlying explanations of differential item functioning in international assessments using multilevel regression models.

Generalized Linear Mixed Model

Generalized linear mixed model (GLMM) or hierarchical generalized linear mixed model (HGLMM) belongs to a general family of mixed effects models, which can be used for continuous, binary, ordinal, categorical, nominal, categorical variables and may include both random and fixed effect in the analysis. When the variable of interest is binary, where usually zero means an incorrect answer and one is equal to a correct answer, the distribution must be considered from a binomial view. Given the predicted value of the outcome, the level 1 random effect can take on only one of two values, and therefore cannot be normally distributed. Thus, the level 1 random effect cannot have homogeneous variance. Instead, the variance of this random effect depends on the predicted value as specified below (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2011, p. 104).

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right)$$

In other words, η_{ij} is the log of the odds of success. Thus, if the probability of success, ϕ_{ij}, is 0.5, the odds of success is 1.0 and the log-odds or logit is zero. When the probability of success is less than 0.5, the odds are less than one and the logit is negative; when the probability is greater than 0.5, the odds are greater than unity and the logit is positive. Thus, while ϕ_{ij} is constrained to be in the interval (0,1), η_{ij} can take on any real value. The level 1 model can be expressed by the next equation (1):

$$\text{logit}[\text{Prob}(Y_{ijk}=1)] = \pi_{0jk} + \pi_{1jk}*\text{ability} + \pi_{2jk}*\text{grouping} \qquad (1)$$

where Y_{ijk} is the binary response/probability of success for a test taker i, from the school j and country k. The level-2 intercept expresses π_{0jk} as a function of random intercept at level-2 $\beta 00k$ plus the level-1 residual error term r_{0jk} and the random intercept at level-3 γ_{000} plus the level-2 residual error term $u_{00k.}$ The level -1 intercept is a function of the grand mean units at level-2 and level-3. If the clustered structure is omitted or not taken into account, then the data may lead to misleading results and incorrect conclusions. The linear mixed regression model allows a random intercept (i.e., each cluster has a different intercept), and a random slope (i.e., each cluster has a different slope).

$$\text{Prob}\,(IT1\_19_{ijk}=1|\pi_{jk}) = \phi_{ijk}$$
$$\log\,[\phi_{ijk}/\,(1-\phi_{ijk})] = \eta_{ijk}$$
$$\eta_{ijk} = \pi_{0jk} + \pi_{1jk}*(\text{ability}_{jk}) + \pi_{2jk}*(\text{grouping}_{ijk})$$

Level-2 model

$$\pi_{0jk} = \beta_{00k} + r_{0jk}$$
$$\pi_{1jk} = \beta_{10k} + r_{1jk}$$
$$\pi_{2jk} = \beta_{20k} + r_{2jk}$$

Level-3 model

$$\beta_{00k} = \gamma_{000} + u_{00k}$$
$$\beta_{10k} = \gamma_{100} + u_{10k}$$
$$\beta_{20k} = \gamma_{200} + u_{20k}$$

Combined equation

$$\eta_{ijk} = \gamma_{000} + \gamma_{100}*\text{ability} + \gamma_{200}*\text{grouping}_{ijk} + r_{0jk} + r_{1jk}*\text{ability}_{ijk} + r_{2jk}*\text{grouping}_{ijk} + u_{00k} + u_{10k}*\text{ability}_{ijk} + u_{20k}*\text{grouping}_{ijk} \quad (2)$$

Above is the mixed model in which the first right side of the equation is the fixed effect and the left second part of the equation is the random term (Equation 2). Random effects are represented as random variables in an LMM; therefore, a random effect has a distribution with an error term, which allows one to generalize the results to a population with a defined probability distribution. The β_{00k} are the means of the level-1 regression coefficients, r_{2jk} are random variables that represent unexplained variability across schools, γ_{000} are the means of the level-1 regression coefficients and u_{20k} are the random variables that represent unexplained variability across countries. Random intercepts represent random deviations for a given cluster or subject from the overall fixed intercept. Random slopes represent random deviation for a given cluster or a subject from the overall fixed effects (slopes). Random effects are random values associated with random factors, contain measurement errors, and vary from sample to sample.

Method

Participants

The data pool used was from the science test which consists of 12,657 students (49.3% students identified as a girl and 50.7% identified as a boy) and 2,609 schools participating in the Third Regional Comparative and Explanatory Study (TERCE) conducted in 2013 in 15 countries in Latin America. The objective of TERCE was to evaluate the knowledge of 6th-grade students. The sample design has been stratified by conglomerates, with a random and systematic selection in two stages. In these designs, the sampling units (schools, classrooms and students) are selected in two or more stages, and these sample units do not have the same probability of being chosen, please see Table 1.

Instruments

Description of the 6th-grade science test

TERCE evaluates three cognitive processes (recognition of information and concepts, understanding and application of concepts, and scientific thinking and problem solving), and five domains of knowledge (health, living beings, environment, the earth and the solar system, and matter and energy). The items were composed of multiple response options and constructed responses; the final data set has the responses coded as binary items (UNESCO-OREALC, 2016). Moreover, the science test is composed of 92 items that were distributed in six blocks or clusters. These blocks were distributed in six different booklet models by an incomplete block design. Each booklet is made up of two blocks or clusters of items between 26 and 30 items totally, and each cluster appears twice throughout the collection of

Table 1
Students and Schools Sample Distributions of the Weighted Sample

	Sample distributions				Predictors means values					
	Student sample	School sample	Gender		Science ability		School physical resources	School SES	GII	HDI
Countries			girls	boys	Z values	Mean of 5 PV				
Argentina	875	165	412	462	-0.05	699.61	0.39	0.79	0.36	0.83
Brazil	855	194	444	411	-0.06	698.46	0.73	0.61	0.41	0.75
Chile	846	168	410	435	0.83	779.01	1.20	0.84	0.32	0.85
Colombia	815	200	446	370	0.34	734.58	0.78	0.29	0.39	0.73
Costa Rica	838	164	406	432	0.59	757.13	0.39	0.55	0.31	0.78
Dominican Republic	833	178	393	441	-0.69	641.20	0.06	0.04	0.39	0.74
Ecuador	822	158	379	443	0.05	708.39	-0.02	-0.21	0.47	0.72
Guatemala	825	182	438	386	-0.22	683.96	-0.50	-0.34	0.49	0.64
Honduras	856	177	421	435	-0.39	668.76	-0.83	-0.94	0.46	0.63
Mexico	821	180	399	422	0.33	733.90	-0.22	0.06	0.35	0.76
Nicaragua	876	171	427	450	-0.49	659.72	-1.02	-1.01	0.46	0.65
Panama	857	188	442	415	-0.37	670.05	-0.02	-0.23	0.46	0.79
Paraguay	835	171	426	409	-0.60	649.99	-0.50	-0.13	0.46	0.69
Peru	833	146	401	432	0.04	707.04	-0.22	-0.50	0.39	0.74
Uruguay	870	168	461	409	0.21	722.91	0.62	1.19	0.28	0.80
Weighted Total	12,657	2,609	6,305	6,352						

booklets, once at the beginning and once at the second position of the booklet. Not-missing values were reported as TERCE provided complete data sets

The variable focus of the analysis in this study was extracted from booklet number one of the 6th-grade science test as follows:

(a) Dependent variable: Item 19 from the science booklet. It is important to denote that TERCE items were presented to the students in a multiple-choice format, but that information is not available to researchers due to TERCE has recoded responses in binary format in the open access dataset. In the current item 19, the coding 0 represents an incorrect answer and 1 represent a correct answer (mean 0.54 and SD = .498). The distribution of the sample by gender is described in Table 2.

The predictors included at the student level were extracted from TERCE student data set.

(a) Sciences ability: In order to remain consistent with the TERCE reporting and analytic methodology, the mean of the five plausible values for every student was computed for the science test. Thus, the complete data set presents a mean of 700.795 (SD = 90.41). For a better comparison, that variable was standardized to the region in a normal distribution with mean 0 and standard deviation 1.
(b) Gender recoded as a 0 for girls and 1 for boys.

The predictors included at the school level were extracted from TERCE school principal and family data set.

(a) School SES (SCH-SES): Index of socioeconomic and cultural status standardized to the region, which is a continuous variable with a mean of 0.28 (SD = 1.05), with a minimum value of -2.41 and maximum of 3.27. The index includes information from 17 items about mother education and house services, resources, and infrastructure (Alpha de Cronbach ranged around .80 between countries).
(b) School physical resources (INFRASTR): Index of the school infrastructure standardized to the region, which is a continuous variable with mean 0.29 (SD = 1.03), with a minimum of -2.37 and a maximum of 2.86. The index includes information from 19 items about services, resources, and school physical infrastructure (Alpha de Cronbach ranged around .70 between countries).

The predictors included at the country level were extracted from the Human Development Report 2013 of the United Nations Development Programme (UNDP, 2013).

Table 2
Gender Distributions on item 19 (weighted sample)

		Gender		
		Girls	**Boys**	**Total**
IT1_19	0	3,166 (25%)	2,814 (22.2%	5,980 (47.2%)
	1	3,140 (24.8%)	3,537 (27.9%)	6,677 (52.8%)
Total		6,306 (49.8%)	6,351 (50.2%)	12,657 (100%)

Note: The percentages reported reflect the percentage of the total sample

(a) Gender inequality index (GII): GII is an index measuring gender disparity. It ranges from 0, which indicates that women and men perform equally, to 1, which indicates that women have the poorest opportunities in all measured dimensions.
(a) Human development index (HDI): HDI is a composite index of life expectancy, education, and average income. It ranges from 0 to 1. A nation scores higher on HDI when its population has a longer life expectancy at birth, longer period of education, and higher average income.

Procedure

The tests were administered by experts from each country in two consecutive days. The first day for reading and writing, and the second day for mathematics and science. Each subject was tested during 45 to 60 min, with a 30 min break in between of each test. The student context questionnaires took about 45 minutes to complete. The family, school and teacher´s questionnaires were distributed on the first day and collected at the end of the second administration day. The study was carried out following the UNESCO ethical guidelines, and the families were informed by the government and school's administrations.

Data analysis

Given the multilevel nature of the TERCE data, a gradual inclusion of the variance distribution in different Bernoulli logistic regression models was carried out. First, we processed the analysis including a two-level model (student-school; student-country), next we tested a three-level model analysis including the student, school, and country information. Penalized quasi-likelihood estimation was the type of estimation applied, which involves the use of a standard HLM model with the introduction of appropriate weighting at level 1. However, after this standard HLM analysis has converged, the linearized dependent variable and the weights must be recomputed. Then, the standard HLM analysis is recomputed. This iterative process of analyses and re-computing weights and linearized dependent variable continues until estimates converge (Raudenbush et al., 2011). All variables in two level models were centered at the group mean (Enders & Tofighi, 2007) and in the case of three level models all variables were centered following Brincks et al.'s (2017) strategy; which implies the use of grand mean centered in order to preserve two sources of variability: within-country, between -school variability and between country variability. The study included senatorial weights from students and school in all the analysis carried out (UNESCO-OREALC, 2016).

Based on the research goals, the analysis were carried out in a out a consecutive order of steps. First, we identify gender DIF using two- and three-level binary (Bernoulli) logistic regression models for every item of the booklet. Equation one was used including only level 1 predictors (ability and gender). There were two goals to be accomplished in this step: the first was to identify the gender DIF in the country average, and the second was to discover significant variability in the random gender slope, which exemplifies not only the presence of gender DIF but also the variability across countries.

The second step in the analysis was to run a Bernoulli logistic regression model, treating the data in two- and three-level

hierarchical modelling. Each analysis included in level 1 the same variables as equation 1 and controlling their effects by adding different predictors in each subsequent level. All the variables at the student level were left constant in all models, and each predictor at level 2 and 3 was included separately based on the complexity of the model and to avoid the collinearity (considering the sample size at a higher level of fifteen countries). Given this same information, Browne and Draper (2006) were able to obtain unbiased variance components with REML estimation with only 6 units at the highest level for a simple model.

1. Two-level Bernoulli logistic regression models including the student level at level-1 and school grouping at level-2.
   Predictors included at level-1: Ability in sciences and gender. Both variables have been centred around the group mean.
   Predictors included at the random slope of gender in level-2: School SES and school infrastructure index.
2. Two level Bernoulli logistic regression models including the student level at level-1 and country grouping at level-2.
   Predictors included at level 1: Ability in sciences and gender. Both variables have been centred around the group mean.
   Predictors included at the random slope of gender in level-2: Gender inequality index and human development index.
3. Three level Bernoulli logistic regression models including the student level at level-1 and school grouping at level-2, and 15 countries at level-3.
   Predictors included at level-1: Ability in sciences and gender. Both variables have been centred around the grand mean.
   Predictors included at the random slope of gender in level-2: School SES and school infrastructure index.
   Predictors included at the random slope of gender in level-3: Gender inequality index and human development index.

Results

An exhaustive analysis of every item in the booklet one was carried out. Nine items were flagged with significant ($p < .05$) coefficients for gender DIF in science booklet number one – which corresponds to 32% of this booklet. Given that DIF distribution in those nine items, girls are more likely to endorse a correct answer in four items, and boys in five of the items. Four of nine items with DIF were flagged with a significant coefficient in gender DIF as well as a significant variability between countries. In broad terms, our first approach has shown the presence of gender DIF in at least 32% of the binary items in booklet number one. Moreover, in consideration of DIF notably, that presence is homogeneous between countries in around five of the items, regardless of some variations between countries in four items flagged with DIF.

Considering our research goals, the item that presented a significant variability in the random slope for the gender coefficient was selected for demonstration purposes of the psychometric methodology. The next step further analyzed the association between the presence of gender DIF and other predictors. Consequently, for the following steps, the item number 19 was included in all the models, taking into consideration the complexity of the models and the methodological goal of this research. Firstly, from the perspective of a non-nested structure, a Chi-Square test was applied to discover the association between the responses' distribution of item 19 and gender, showing a significant association between those variables (χ^2=27.166, p<*.000)*. Secondly, taking into account a nested structure of the data, different models were performed. Even though all the models will be explained in the subsequent pages, a brief description of our model zero (gender DIF) for all the levels analyzed is presented in Table 3.

The model zero (M0) is based in equation 1, and it has the aim of detecting not only if an average gender DIF effect exists, but also if this DIF effect has shown variability across groups (schools or countries). Comparing a holistic visualization of the gender DIF coefficients in all models (column 3 of Table 3), we were able to detect a positive coefficient. Based on our gender codification in the data set, girls equal zero and boys equal one. This result shows that even though when omitting or including variability across levels, boys are more like to endorse (answer correctly) on item number 19 than girls. It is important to note that the results in model zero (M0) are not controlling for contextual variables. That result, or phenomenon is variant across countries but is constant across schools (column 9, table 3).

Progressively, we drew in more of the nested structure information in our analysis. The next step included the variation of school level. We analyzed two-level Bernoulli logistic regression models including the student variables at level 1 (ability and gender) and controlling the random slope of gender by school characteristics (school SES and school infrastructure) at level two. Each variable was included separately in the analysis in contemplation of the estimation complexity and to avoid the multicollinearity due to the high correlation (r = .783, n = 2663, p = .000). With the intention of discovering predictors that can explain the relationship between items responses and gender in different levels, we included variables that characterized the school profile. Table 4 displays all the models analyzed; it clearly shows

Table 3
Summary of the DIF Results Presented by Multilevel Models

Item 19		Fixed Effect Gender (gamma γ_{20})		Random Effect Gender				
		Coefficients	odds ratio		SD	Variance component	Chi-square (df)	p-value
Two-level student/country	M0 Gender-DIF	0.311****	1.365	u_2	0.214	0.04592	41.137 (14)	<0.001
Two-level student/school	M0 Gender-DIF	0.288****	1.334	u_2	0.123	0.01526	1442.365 (1618)	>0.500
Three-level student/school/country	M0 Gender-DIF	0.199****	1.220	r_2	1.677	2.81446	1425.808 (1606)	>0.500
				u_{20}	0.327	0.10704	39.14620 (14)	<0.001

Note: Gender was codified by 0 for girls and 1 for boys. **p<.05, ***p<.01, ****p<.001, SD: Standard deviation, df: degrees of freedom

the presence of DIF favouring boys in all the models (column 2) but not a significant variability between schools, which represent a similar profile of DIF across schools (column 10).

The coefficients of all the variables included at the school level can be seen in Table 5, column two. The coefficients of gender DIF are significant and positive in all the models. Given our variable codification, the intercept of the model is zero for urban school in M5 and zero for public schools in M6, but both of those variables are a non-significant predictor of gender slope (p = .962 and p = .950). In the same line, school climate (M2) and teacher strategy (M3) present a negative coefficient as well as non-significant values (p = .289 and p = .388). However, in Table 5, two variables associated with school and family resources were positive but not significant predictors of gender slope at the school level (columns 4-5 and 10-11). As a result, none of the variables (such as school climate, type of professor strategy used, and rural or private school) are significant predictors of the relationship between gender and item responses.

Focusing on our principal purpose—the impact of country predictors—the two-level Bernoulli logistic regression model was run. Student characteristics were included at level 1 (ability and gender), while random gender slope was controlled by country characteristic (GII and HDI). Taking into consideration that the correlation between GII and HDI is -.703, which implies a high correlation between those two indexes, every variable in the model was included separately.

Similar to the results in Table 3, the results on Table 6 shows that gender DIF for item 19 is favouring boys even after it is controlled by the country gender inequality index. It is important to observe, however, the gender DIF switches to favouring girls when controlled by the country level of human development (column 2). It is noteworthy that girls are four times more likely to endorse that item correctly when the country increases the amount in their human development index (Table 7).

Considering the strength of the multilevel approach, we carried out an analysis that allowed for the insertion of the variability

Table 4
Two level Models: Student and School

	Fixed Effect Gender (gamma γ_{20})					Random Effect gender u_2			
Item 19	**Coefficients**	**p-value**	**odds ratio**	**Confidence Interval**	**Who is more likely to endorse**	**Standard deviation**	**Variance component**	**Chi-square (df)**	**p-value**
M0 Gender_DIF	0.288502	<0.001	1.334	(1.166-1.528)	boys	0.12355	0.01526	1442.365 (1618)	>0.500
M1_SCH_SES	0.266529	<0.001	1.305	(1.129-1.510)	boys	0.13506	0.01824	1440.718 (1617)	>0.500
M2_CLIMATE	0.277729	<0.001	1.320	(1.151-1.514)	boys	0.12426	0.01544	1440.048 (1617)	>0.500
M3_STRATEGY	0.274064	<0.001	1.315	(1.144-1.513)	boys	0.12589	0.01585	1442.621 (1617)	>0.500
M4_INFRASTR	0.245505	<0.001	1.278	(1.105-1.479)	boys	0.13948	0.01945	1439.905 (1617)	>0.500
M5_RURAL	0.290627	<0.001	1.337	(1.129-1.584)	boys	0.12389	0.01535	1442.414 (1617)	>0.500
M6_TYPE_SCH	0.290875	<0.001	1.337	(1.134-1.577)	boys	0.12278	0.01507	1442.435 (1617)	>0.500

Note: Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom

Table 5
Two level Models: Student and School

	Fixed Effect Gender (gamma γ_{20})		Model 1		Model 2		Model 3		Model 4		Model 5		Model 6	
Item 19	**Coefficients Gender (gamma γ_{20})**	**p-value**	**SCH_SES (gamma γ_{21})**	**p-value**	**SCH_ CLIMATE (gamma γ_{21})**	**p-value**	**TEACH_ STRAT (gamma γ_{21})**	**p-value**	**SCHO_ INFAEST (gamma γ_{21})**	**p-value**	**RURAL (gamma γ_{21})**	**p-value**	**TYPE_SCH (gamma γ_{21})**	**p-value**
M0 Gender_DIF	0.288502	<0.001	–	–	–	–	–	–	–	–	–	–	–	–
M1_SCH.SES	0.266529	<0.001	0.060	0.378	–	–	–	–	–	–	–	–	–	–
M2_CLIMATE	0.277729	<0.001	–	–	-0.073	0.289	–	–	–	–	–	–	–	–
M3_STRATEGY	0.274064	<0.001	–	–	–	–	-0.075	0.388	–	–	–	–	–	–
M4_INFRASTR	0.245505	<0.001	–	–	–	–	–	–	0.110	0.102	–	–	–	–
M5_RURAL	0.290627	<0.001	–	–	–	–	–	–	–	–	-0.007	0.962	–	–
M6_TYPE_SCH	0.290875	<0.001	–	–	–	–	–	–	–	–	–	–	-0.009	0.950

Note: Gender was codified by 0 for girls and 1 for boys

Table 6
Two level Models: Student and Country

	Fixed Effect Gender (gamma γ_{20})					Random Effect gender u_2			
Item 19	**Coefficients**	**p-value**	**odds ratio**	**Confidence Interval**	**Who is more likely to endorse**	**Standard deviation**	**Variance component**	**Chi-square (df)**	**p-value**
M0 Gender-DIF	0.311543	0.002	1.365531	(1.144-1.630)	boys	0.21430	0.04592	41.13742 (14)	>0.001
M1 DIF controlled by GII	1.437591	0.009	4.210539	(1.541-11.506)	boys	0.10718	0.01149	20.83019 (13)	0.076
M2 DIF controlled by HDI	-1.896979	0.022	0.150021	(0.031-0.721)	girls	0.09330	0.00871	19.29655 (13)	0.114

Note: Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom

Table 7
Two level Models: Student and Country

	Fixed Effect Gender (gamma γ_{20})			Model 1 controlled by GII		Model 2 controlled by HDI	
Item 19	**Coefficients Gender (gamma γ_{20})**	**p-value**	**Who is more likely to endorse**	**GII (gamma γ_{21})**	**p-value**	**HDI (gamma γ_{21})**	**p-value**
M0 Gender-DIF	0.311543	0.002	boys	–	–	–	–
M1 Gender inequality index	1.437591	0.009	boys	-2.819408	0.026	–	–
M2 Human development index	-1.896979	0.022	girls	–	–	2.984254	0.010

Note: Gender was codified by 0 for girls and 1 for boys

Table 8
Three Level Models: Student, School and Country

	(M1)		(M2)		(M3)		(M4)	
Fixed effects	**Coefficients**	**p-value**	**Coefficients**	**p-value**	**Coefficients**	**p-value**	**Coefficients**	**p-value**
Gender γ_{200}	1.198166	0.344	-3.067581	0.148	1.095772	0.392	-1.111502	0.093
SCH_SES γ_{210}	0.132051	0.387	0.080799	0.596	–	–	–	–
SCH_INFRA γ_{210}	–	–	–	–	0.199960	0.180	0.066322	0.185
GII γ_{201}	-2.546304	0.407	-	-	-2.321548	0.456	–	–
HDI γ_{201}	–	–	4.400968	0.124	–	–	1.829206	0.048
Random effects	**Variance component**	**p-value**	**Variance component**	**p-value**	**Variance component**	**p-value**	**Variance component**	**p-value**
Gender r_2	2.82194	>0.500	2.82088	>0.500	2.85952	>0.500	0.06129	>0.500
Gender u_{20}	0.07641	0.004	0.04202	0.039	0.06211	0.011	0.10653	0.225

Note: Gender was codified by 0 for girls and 1 for boys

between schools and countries. For that purpose, different models were performed including the three-level Bernoulli logistic regression models. Four different models were analyzed. For all the models, the variables at level 1 were constant (ability and gender) and the slope of gender was controlled by one variable separately at each time in every level 2 and 3 (Table 8). In the first model (M1) implemented (column 2-3), gender DIF was controlled by the school's socioeconomic status at level 2 and country index of gender inequality (GII) at level 3.

In addition to the variables at level 1 (ability and gender), the second model (M2) included the school SES at level 2 and the index of human development at the country level. In the third model (M3), the variation in gender coefficient was controlled by the school infrastructure index (level 2) and the gender inequality index at the country level (level 3). In the last model (M4), the coefficient of gender was controlled by the school infrastructure index (level 2) and the human development index at the country level (level 3). After controlling for level 2 and level 3 variables, the principal result is that the coefficients of gender DIF are not significant in all models. Notwithstanding, the inclusion of the country human development index switches the sign of gender coefficient. Hence, this results in favouring girls over boys (Table 8).

After running a series of analysis including variables at both the school and country level, as well as bearing in mind the previous non-remarkable results for two-level analysis, we decided to allow the inclusion of the natural variability for the school level. As well,

due to the model complexity we omit the inclusion of predictors at level 2 (Table 9). The data is then presented in a holistic visualization, which includes the variability or the impact of the student characteristics, school's effects, and country properties. We found that, even after controlling for different conditions, gender uniform DIF is still present. However, the association between gender and item responses changes to favouring girls when the variable human development index is included at the country level (Table 9 and 10, column 2). Considering the negative relationship between gender DIF and GII, this result implies that a medium size probability of gender DIF is associated with lower inequality. Taking into consideration the relationship between gender and items response controlled by HDI, girls are four times more likely to give a correct answer than boys. That relationship suggests that with higher levels in HDI, it is more likely to favour girls than boys in most of the nations participating in TERCE.

Discussion

It is important to keep in mind that, although some psychometric theorists certainly recognize and acknowledge that contextual effects are worthy of consideration, conventional validation practices and theorizing do not pay much attention to contextual effects as part of validation. That is, although conventional validation practice would not disagree with the generic role of context in assessment, it does not pay much attention to it. Conventional validation practices place the contextual effects in the background while individual differences between test takers are in the foreground (Zumbo & Forer, 2011). This is particularly important given the well-known large education inequality in Latin America that are related to contextual factors (UNESCO-OREALC, 2016a).

This research has aimed to provide a holistic explanation about why DIF was occurring and how that situational factors can bias the results obtained in educational assessments in Latin America contexts. The validity of the inferences one makes from test scores is bounded by place, time, and use of the score resulting from a measurement operation (Zumbo, 2007a). In our case, DIF was explained by various factors from an ecological view, including the information about the schools and countries characteristics. Even though TERCE states that they performed a gender DIF analysis, the technical report indicates that no item has shown to be a significant gender DIF. The results obtained for TERCE are not available for methodological analysis. Additionally, the technical report states that gender DIF is not a criterion for item elimination (UNESCO-OREALC, 2016b, p. 252). However, the absence of significant gender DIF results can be explained not only by the technique used (in this case, Mantel-Haenszel analysis) but also due to the omission of the information from the nested structure of the data.

The data reveals to us, in a holistic visualization of the results, that even if the model includes or omits the variability or the impact of the students' characteristics, schools' effects and countries' properties, that gender DIF is still present. However, the association between gender and item responses changes to favoring girls when the human development index is included at the country level. A further dilemma arises for the particular process of DIF validity studies as the nested nature of the data cannot be underestimated and test takers have to be viewed in their complete life circumstances. A compounding variable in testing is the fact that although a great deal of the work is done in isolation, it is nevertheless influenced by contextual factors, such as the class environment, the school resources, country politics,

Table 9
Three Level Models: Student, School and Country

	Fixed Effect Gender (gamma γ_{200})				Random Effect Gender				
Item 19	**Coefficients**	**p-value**	**odds ratio**	**Confidence Interval**		**Standard deviation**	**Variance component**	**Chi-square (df)**	**p-value**
M0 Gender-DIF	0.199494	0.234	1.220785	(0.865-1.722)	r_2	1.67764	2.81446	1425.80876 (1606)	>0.500
					u_{20}	0.32717	0.10704	39.14620 (14)	>0.001
M1 DIF controlled by GII	1.715366	0.167	5.558710	(0.442-69.851)	r_2	1.67919	2.81969	1425.87216 (1606)	>0.500
					u_{20}	0.29750	0.08851	33.54459 (13)	0.002
M2 DIF controlled by HDI	-3.544203	0.095	0.028892	(0.000-2.043)	r_2	1.67939	2.82036	1426.13816 (1606)	>0.500
					u_{20}	0.19896	0.03959	22.71597(13)	0.045

Note: Gender was codified by 0 for girls and 1 for boys, df: degrees of freedom

Table 10
Three Level Models: Student, School and Country

	Fixed Effect Gender (gamma γ_{20})			M1 controlled by GII		M2 controlled by HDI	
Item 19	**Coefficients Gender (gamma γ_{20})**	**p-value**	**Who is more likely to endorse**	**GII (gamma γ_{201})**	**p-value**	**HDI (gamma γ_{21})**	**p-value**
M0 Gender-DIF	0.199494	0.234	boys	–	–	–	–
M1 Gender inequality index	1.715366	0.167	boys	-3.785609	0.215	–	–
M2 Human development index	-3.544203	0.095	girls	–	–	5.063227	0.075

Note: Gender was codified by 0 for girls and 1 for boys

and socioeconomic reality. The inclusion of the environmental information into the educational assessment methodology is not necessarily a new approach (i.e. computation of plausible values and sampling). Although, if we read carefully the psychometric chapter in UNESCO's Technical Report most of the item decision criteria are based in psychometric analysis performed without including the contextual information (i.e. classic item difficulty, IRT difficulty, item discrimination, and reliability). The nested structure in psychometrics can be used in the invariance analysis (Balluerka et al., 2010, 2014; Byrne & van de Vijver, 2017; Chen & Zumbo, 2017; Gadermann et al., 2018; Swanson et al., 2002; van den Noortgate & de Boeck, 2005), and also in reliability estimation (Nezlek, 2017).

Most large-scale data sets are not constructed with explanatory modeling in mind. Therefore, a limitation of modeling extant data sets is that the explanatory variables that one can use in their models are limited to those that the initial survey designers included. We encourage assessment specialists to consider explanatory models from the initial planning of a study so that competing explanatory item response theories can be empirically tested. This, we believe, moves psychometrics directly in to the scientific worldview where theory building and theory-testing (in our case of item responses and test scores, in the tradition of explanatory psyhometrics advocated by Zumbo, 2007a) is the core of the activities of a psychometric science.

The basis of the objectives and results of this paper was to understand that "contextual" measurement determines not only the opportunities to learn that students are exposed to, but also the way the students understand and respond to test items. The study was performed using a novel analytical strategy and theory that allowed the inclusion of many of the variables which describe the educational environment. The contribution of those results may be in their application at both the methodological and educational policy level. They stand as evidence of the validity of TERCE measures, in the evaluation of the test construct and the analysis of the test response process. Validity is the foundation of a testing procedure, and the process of validating is key to the overall success of the educational assessment as a whole. This study deals specifically with the position of an ecological point of view which includes and situates the person, process, context, and time of the testing situation. These descriptions pinpointed specific incidents of how and what variables at the individual, school, or country level can give a deep understanding of the response process in Latin America countries.

Acknowledgments

The authors wish to thank Professor José Muñiz and Professor Yan Liu for their insights and feedback on earlier versions of this paper.

References

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

Balluerka, N., Gorostiaga, A., Gómez-Benito, J., & Hidalgo, M. D. (2010). Use of multilevel logistic regression to identify the causes of differential item functioning. *Psicothema*, *22*(4), 1018-1025.

Balluerka, N., Plewis, I., Gorostiaga, A., & Padilla, J.L. (2014). Examining sources of DIF in psychological and educational assessment using multilevel logistic regression. *Methodology*, *10*(2), 71-79. doi:10.1027/1614-2241/a000076

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research*, *52*(2), 149-163. doi:10.1080/00273171.2016.1256753

Browne, W., & Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*(3), 473-514. doi:10.1080/00273171.2016.1256753

Byrne, B.M., & van de Vijver, F.J.R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: A paradigmatic cross-cultural application. *Psicothema, 29*(4), 539-551. doi:10.7334/psicothema2017.178

Cheema, J. (2017): Cross-country gender DIF in PISA science literacy items. *European Journal of Developmental Psychology*, *16*(2), 152-166. doi: 10.1080/17405629.2017.1358607

Chen, M.Y., & Zumbo, B.D. (2017). Ecological Framework of Item Responding as Validity Evidence: An application of Multilevel DIF Modeling using PISA Data. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and Investigating Response Processes in Validation Research* (pp. 53-68). New York, NY: Springer International Publishing.

Elosua, P., & Mujika-Lizaso, J. (2013). Invariance levels across language versions of the PISA 2009 reading comprehension test in Spain. *Psicothema, 25*(3), 390-395. doi:10.7334/psicothema2013.46

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121-138. doi:10.1037/1082-989X.12.2.121

Gadermann, A.M., Chen, Y., Emerson, S.D., & Zumbo, B.D. (2018). Examining validity evidence of self-report measures using differential item functioning: An illustration of three methods. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 14*(4), 164-175. http://dx.doi.org/10.1027/1614-2241/a000156

Gómez Benito, J., Balluerka, N., González, A., Widaman, K., & Padilla, J. L. (2017). Detecting differential item functioning in behavioral indicators across parallel forms. *Psicothema*, *29*(1), 91-95. doi: 10.7334/psicothema2015.112

Gómez Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109. doi:10.7334/psicothema2017.183

Nezlek, J. B. (2017). A practical guide to understanding reliability in studies of within-person variability. *Journal of Research in Personality*, *69*, 149-155. http://dx.doi.org/10.1016/j.jrp.2016.06.020

Raudenbush, S., Bryk, A., Cheong, Y. K., Congdon, R., & du Toit, M. (2011). *HLM 7 Hierarchical Linear and Nonlinear Modeling*. Lincolnwood,IL: SSI Scientific Software International.

Swanson, D.B., Clauser, B.E., Case, S.M., Nungester, R.J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53-75. doi:10.3102/10769986027001053

United Nations Development Programme [UNDP] (2013). *Human Development Report 2013. The rise of the South: Human Progress in a Diverse World*. New York: UNDP.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & la Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC] (2016a). *Recomendaciones de políticas educativas en América Latina en base al TERCE* [Recommendations for educational policies in Latin America based on TERCE]. Santiago de Chile: UNESCO.

Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura & la Oficina Regional de Educación para América Latina y el Caribe [UNESCO-OREALC] (2016b). *Reporte técnico tercer estudio regional comparativo y explicativo*. *TERCE* [Technical report third comparative and explanatory regional study. TERCE]. Santiago de Chile: UNESCO.

van den Noortgate, W., & de Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*(4), 443-464.

Villegas, G., González, N., Sánchez-García, A., Sánchez-Barba, M., & Galindo, M. (2018). Seven methods to determine the dimensionality of test: Application to general self-efficacy scale in twenty-six countries. *Psicothema*, *30*(4), 442-448. doi: 10.7334/psicothema2018.113

Zumbo, B.D. (2007a). Validity: Foundational Issues and Statistical Methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, Vol. 26: *Psychometrics* (pp. 45-79). The Netherlands: Elsevier Science.

Zumbo, B.D. (2007b). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223-233.

Zumbo, B.D. (2014). What role does, and should, the test standards play outside of the United States of America? *Educational Measurement: Issues and Practice, 33*, 31-33.

Zumbo, B. D., & Forer, B. (2011). Testing and Measurement from a Multilevel View: Psychometrics and Validation. In J.A. Bovaird, K.F. Geisinger & C.W. Buckendahl (Editors). *High Stakes Testing in Education - Science and Practice in K-12 Settings* (pp. 177-190). Washington, D.C.: American Psychological Association Press.

Zumbo, B.D., & Gelin, M.N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies*, *5*, 1-23.

Zumbo, B.D., & Hubley, A. M. (Eds.) (2017). *Understanding and Investigating Response Processes in Validation Research*. New York, NY: Springer.

Zumbo, B.D., Liu, Y., Wu, A.D., Shear, B.R., Astivia, O.L.O., & Ark, T.K. (2015). A methodology for Zumbo's Third Generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*, 136-151.