Revista Interamericana de Psicología/Interamerican Journal of Psychology - 2013, Vol. 47, Num. 3, pp. 423-432

Differential Item Functioning and Educational Risk Factors in Guatemalan Reading Assessment

Alvaro M. Fortin Morales[1]
Universidad del Valle de Guatemala, Guatemala and Tilburg University, Netherlands
Fons J. R. van de Vijver
Tilburg University, the Netherlands and North-West University, South Africa
Ype H. Poortinga
Tilburg University, Netherlands

Abstract

We examined Differential Item Functioning (DIF) indicators for four variables that repeatedly have been demonstrated to constitute risk factors in primary school achievement in Guatemala. These factors are over-age of enrollment, urban/rural area of residence, ethnicity, and gender. We used scores from national reading assessments in third-grade for this study. Given the instability often reported in DIF literature, we employed three different approaches (chi-square, Rasch, and logistic regression) and checked for their consistency with data from three calendar years. We found substantial evidence of DIF. However, removal of DIF items did not influence differences in test scores between groups. Findings suggest that educational risk factors act in concert in this Guatemalan population and, that that at least to some degree, they interact to create bias. We conclude that DIF analysis and test-writing would benefit from taking into account multiple background risk variables simultaneously.
Keywords: differential item functioning, reading assessment, multiple risk factors, Guatemala.

Factores de riesgo educativo y funcionamiento diferencial de ítems en la evaluación de la lectura en Guatemala
Resumen

Examinamos indicadores de Funcionamiento Diferencial de Ítemes (FDI) asociados a cuatro variables que han demostrado de manera repetida ser factores de riesgo para el logro escolar. Estos factores son el sobre-edad para el grado de matriculación, área de residencia urbana/rural, etnia y género. Para este estudio utilizamos los datos de las evaluaciones nacionales del tercer grado. Dado que en la literatura se reporta con frecuencia que los indicadores de FDI son inestables, utilizamos tres diferentes métodos para estimarlo (chi-cuadrado, Rasch, regresión logística) y evaluamos su consistencia en datos de tres diferentes años de evaluaciones. Encontramos evidencia de FDI. Sin embargo, la eliminación de ítemes con FDI no cambió las diferencias entre grupos que se encontraron en las puntuaciones de las evaluaciones. Los hallazgos sugieren que los factores de riesgo educativo actúan de manera conjunta en esta población guatemalteca y que hay alguna interacción entre estos factores de riesgo para generar sesgo. Concluimos que será de beneficio tomar en cuenta múltiples variables de contexto asociadas al riesgo educativo de forma simultanea al analizar FDI y al desarrollar evaluaciones.
Palabras claves: funcionamiento diferencial del ítem, evaluación de lectura, factores de riesgo múltiples, Guatemala.

Guatemala is an ethnically diverse, multilingual society. Throughout the history of the country, ethnic diversity has been associated with various social disparities. Since the 1980s the increase in school enrollment has highlighted the government's belief that education could potentially contribute to offsetting these conditions. Efforts to monitor these initiatives have included the collection of data on enrollment and school efficiency and, more recently, educational assessment.

Four factors, namely ethnicity, gender, urban or rural area of residence and school location, and being over the age for the school grade (over-age) have been documented as risk factors for poor educational achieve-

[1] Correspondence about this article should be addressed to Tilburg University, the Netherlands. Email: alvarofortin@gmail.com. We wish to acknowledge the contribution of the Ministry of Education and the Center for Educational Research of Universidad del Valle de Guatemala for providing the necessary data to conduct this study.

ment in several countries, with these factors often acting simultaneously and jointly (Deater-Deckard, Dodge, Bates & Pettit, 1998; Gerard & Buehler, 1999; Rutter, 2001; Rutter, 1979, 1988; Sameroff, Bartko, Baldwin, Baldwin, & Seifer, 1998). Educational assessment results have consistently indicated that these are also variables associated to lower performance of pupils in Guatemala (de Baessa, 1999, 2000; Moreno-Grajeda, Gálvez-Sobral, Bedregal, & Roldán, 2008).

There is also extensive evidence that the aforementioned factors can potentially create bias in assessment. This is a reason for concern when implementing educational policies based on assessment data, as interventions might reflect inaccurate considerations of the pupils' actual potentials and accomplishments. In this paper we study the case of Guatemala, where these four risk factors are known to be present (de Baessa, 1999, 2000; Moreno-Grajeda, Gálvez-Sobral, Bedregal, & Roldán, 2008) and presumably are interrelated (Esquivel Villegas, 2006). More concretely, we examine the relationship of the four factors in producing item bias or Differential Item Functioning (DIF) in reading tests of third graders.

Educational Risk Factors in Guatemala

Guatemala's population is multiethnic and multilingual. In addition to the "Mestizo" or "Ladino" and immigrant groups, about 40% of the population belongs to one of some 24 indigenous ethnicities (21 different Mayan groups, the Xinca, and the Garinagu, also known as Garífuna) (Richards, 2003; World Factbook, 2011). There are nearly as many languages as ethnic groups, in addition to Spanish that is the official language of the country (Richards, 2003). Guatemala is a "mid-development" country as measured by the Human Development Index with uneven distribution of wealth as measured by the Gini Index (55.9) (United Nations Development Programme, 2011). The disparities are accentuated in rural areas, where the poorest segments of the population and most Mayans are living (Antillón Milla, 1997).

School enrollment statistics show significant proportions of over-aged pupils (being 1.5 years older than the expected age for the grade level of enrollment), particularly in rural areas (Ministry of Education of Guatemala, 2011). The number of students enrolled in school has increased over the last decade, but educational indicators show that urban areas, males and non-indigenous populations have benefited most (Álvarez & Schiefelbein, 2007; Esquivel Villegas, 2006). Although global enrollment indicators for boys and girls are similar, more detailed analysis shows that women still have less access to school in scarcely populated and predominantly Mayan areas (Ministry of Education, 2011). In summary, being an older pupil, being female, being Mayan and attending a rural school are usually risk factors. These characteristics constitute risk factors in that they are strongly associated to the access pupils have to good quality of education.

Differential Item Functioning

To compare groups on test scores, there must be sufficient evidence that the scores of different groups can be interpreted in the same way. Given the differences in contexts for various segments of the population in Guatemala, the assessment practitioner should analyze data with a view to identify various forms of bias or inequivalence (van de Vijver & Tanzer, 2004). Differential Item Functioning (DIF) detection methods are the most relevant and the most frequent bias

An item shows DIF when test-takers of different groups who have the same ability have different probabilities of obtaining a correct answer (in this study, for example, an item would show gender DIF when a boy who obtained the same overall score as a girl is more likely to get the answer to that particular item correct for reasons not related to their ability) (Angoff, 1993; Dorans & Holland, 1993; Ellis, 1990; Finch & French, 2008; Uiterwijk & Vallen, 2003; van de Vijver & Tanzer, 2004; van den Noortgate & de Boeck, 2005; Zenisky et al., 2003). DIF is "uniform" when it favors the same group across all ability levels and "non-uniform" when the size or direction of the bias effect varies across ability levels (Jodoin & Gierl, 2001; Welkenhuysen-Gybels, 2003).

DIF often shows poor coherence across computational procedures (Bond, 1993; Bond & Fox, 2007; Camilli, 2006; Dodeen, 2004; Linn, 1993; Longford, Holland, & Thayer, 1993; O'Neill & McPeek, 1993; Wiberg, 2007). These variations in statistical outcomes may result from different data assumptions or methodological variations (Angoff, 1993; Robitzsch & Rupp, 2008; Jodoin & Gierl, 2001; Linn, 1993; O'Neill & McPeek, 1993; Scheuneman, 1987; Scheuneman & Gerritz, 1990; Wiberg, 2007). In the present study we attempted to overcome this by employing three different, but widely used procedures: chi-square techniques, Rasch method, and logistic regression.

The chi-square or contingency table techniques compare the proportion of examinees per test score level responding correctly to the item across groups (Crocker & Algina, 1986). The Mantel-Haenszel statistic and the Breslow-Day test of trend in odds ratio heterogeneity are chi-square techniques; the former mainly to detect uniform DIF and the latter non-uniform DIF (Angoff, 1993; Bertrand & Boiteau, 2003; Dorans & Holland, 1993; Fidalgo & Madeira, 2008; Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Narayanan

& Swaminathan, 1994; Penfield, 2003). Item Response Theory (IRT) methods assume that the probability of solving an item correctly is a function of the total test score, and that this function follows a logistic curve (Jin-Shei, Teresi, & Gershon, 2005). The curve is defined by one, two, or three parameters (discrimination, proficiency level, and pseudo-guessing) (Angoff, 1993). DIF is detected when there is a significant difference between the populations in one or more parameters of the item characteristic curves (Angoff, 1993; Crocker & Algina, 1986; Thissen, Steinberg, & Wainer, 1993). The Rasch model is a one-parameter IRT method (Bond & Fox, 2007) that usually produces results consistent with other multiparameter IRT procedures (Thissen, Steinberg, & Wainer, 1993). Uniform and non-uniform DIF can also be detected by estimating a logistic regression where the right/wrong answer on the item is predicted by group membership and performance level (Jodoin & Gierl, 2001; Swanson, Clauser, Case, Nungester, & Featherman, 2002).

DIF research has not been successful in consistently identifying item characteristics that generate bias. Perhaps the most consistent result has been that highly discriminating and more difficult items are also more likely to exhibit DIF towards the non-risk group (Linn, 1993; Scherbaum & Goldstein, 2008). Other than these, good predictions are difficult as item characteristics can interact to create DIF in some conditions and not in others (O'Neill & McPeek, 1993). These findings do not lead to practical guidelines for item writing. DIF identification can also be used to "purify" tests by removing the biased items (French & Maller, 2007). The desirable outcome is the removal of items that have a large impact on the results of the tests.

The Present Study

We employed the Mantel-Haenszel chi-square statistic the Breslow-Day chi-square statistic, Rasch and logistic regression to detect DIF items in Spanish reading tests used in 1999, 2000, and 2004 in an attempt to establish whether there is a general pattern of educational risk in Guatemala. To accomplish this we examined the convergence of DIF indicators across risk factors. To make sure the choice of DIF detection method is not a factor in determining the convergence, we compared the results across computational procedures. To understand the impact of DIF in the assessment results, we explored the influence of DIF removal on the size and direction of group differences.

Method

Participants

We analyzed data provided to us by the Ministry of Education from the Spanish national reading tests administered on nationally representative samples in 1999, 2000, and 2004 among third grade students of public elementary schools (DIGEDUCA, 2008). Table 1 shows the number of students in each group for each year. Age and gender were reported on the test forms, and were verified by test administrators and teachers. Age was dichotomized according to appropriateness for the grade level: Those of age 11 or older were considered over-aged for the third grade. Area of residence was dichotomized into urban and rural according to the Ministry's official classification of the schools (each school serves students whose residence is within a three kilometer radius). A large number of cases in the data sets were missing; individual registries on ethnicity and the ones available were not always consistent with the main languages spoken in the schools. Therefore, a proxy was used, based on the density of Mayan students enrolled in public institutions per Department (political division of the country). Students from those Departments that according to the registries of the Ministry of Education of Guatemala (2009) had a Mayan enrollment of 90% or more were classified as Mayan. Students coming from Departments where Mayan enrollment was 10% or less were classified as non-Mayan. Cases that did not belong to either of these groups were removed from the data set to estimate DIF by ethnicity. The Department of Guatemala, where the capital city is located and would have fallen under

Table 1
Number of Cases per Risk Factor and Year

Risk Factor	1999	2000	2004
Age			
Appropriate	3999	5502	2851
Over-age	3022	3446	1733
Area			
Urban	3433	4519	1493
Rural	3588	4429	3091
Ethnicity			
Non-Mayan	1791	2486	1363
Mayan	1841	1699	974
Gender			
Male	3646	4659	2271
Female	3375	4289	2156
TOTAL	7021	8948	4584

Note. Totals of categories across risk factors do not always add up to total sample size, due to missing scores.

the non-Mayan categorization, was also removed due to the high likelihood of mixed enrollment in schools.

Instruments and Data Sets

Each test consisted of 40 multiple-choice items. Each item had four response options, which were converted into a dichotomous correct / incorrect variable for this study. Original assessment documentation reported Cronbach's alpha values of .80 or higher in all data sets and the tests were generally considered difficult for the target population (de Baessa, 1999, 2000; Moreno-Grajeda et al., 2008). Items were written according to a common set of specifications and attempting to replicate the same types of distractors and contexts. Only items that had been employed in both rural and urban areas and for which information on the other risk factors was available, were selected for the analysis. This data set included 80 items (20 for 1999, 20 for 2000, and 40 for 2004).

Procedure and Analysis

We first conducted item-related analyses to detect DIF and compute the effect size of the DIF indicators using the computational procedures already described. Then we investigated the convergence of the indicators across risk factors and computation methods. Finally, we conducted a "purification" analysis, removing items that showed evidence of bias. Each of these analysis is further described below.

Estimation of the statistical significance of DIF. We used three computational methods to estimate the statistical significance of DIF. Firstly, we estimated the Mantel-Haenszel chi-square and the Breslow-Day chi-square using the software Differential Item Functioning Analysis System (DIFAS 4.0). Following the suggestion contained in the manual, an item was flagged for DIF when either of these two indicators was significant at a Type I error rate of .025 (Penfield, 2007a). Also, we estimated Rasch indices using Winsteps 3.65 (Linacre, 2006); an item was considered biased when the t statistic for the difference in logits between groups was significant ($p < .05$). Finally, we estimated logistic regression coefficients where the (dichotomous) focal split for each risk factor was entered as a categorical covariate. The total score and the interaction between the classificatory variable and the total score were then entered as covariates; here a significant model for an item implies DIF. We synthesized the results from these three analyzes by flagging an item as showing DIF for a risk factor when the indicators for all three DIF identification procedures were statistically significant for that factor.

Estimation of the effect size of DIF. We estimated the effect size of the item bias to acquire a measure of the magnitude of the biasing effect. For the chi-square procedures we used the absolute value of the standardized measure of the log-odds ratio (Camilli, 2006) as it is provided by the Differential Item Functioning Analysis System (DIFAS 4.0; Penfield, 2007b). For the Rasch procedure we used the average impact on the person parameter by estimating the absolute value of the quotient obtained by dividing the difference in logits of the dominant to non-dominant groups by the number of items in the test (Linacre, 2006). In the case of the logistic regression method, which does not provide a direct measure of the variance explained by the predictor variables (R^2), we used the Nagelkerke R^2 and compared the values of the full model and the model explained only by the total score to estimate ΔR^2 (Hidalgo & López-Pina, 2004; Jin-Shei et al., 2005; Jodoin & Gierl, 2001; Swanson et al., 2002).

Convergence of DIF indicators. Once individual items had been analyzed for the statistical significance of the three DIF indicators, we explored their convergence. Using the φ statistic, we computed the correlations for item sets between pairs of DIF outcomes and between pairs of risk factors. Positive correlations across the risk factors would indicate that there is a tendency for items to behave in similar manner for these factors. Positive correlations across methods would indicate that these methods show consistent outcomes with regard to DIF/non-DIF classifications. We also calculated the corresponding Pearson correlations between effect size measures. As described above, we did this for the three computation procedures with the same risk factor and for the four risk factors with the same effect size computation procedure.

Assessing the impact of bias. To evaluate the substantive effect of DIF on the conclusions drawn from the assessment scores, we compared results for the full test with results for a purified version from which biased items had been removed. Mean scores before and after removal of items were compared using a t test for independent samples. We did this once using the dichotomous classification based on statistical significance, and once again using the criterion based on the effect size of the items. In both cases an item was only removed if flagged for all three computational procedures (chi-square, Rasch, and logistic regression). As criterion for statistical significance, an alpha level of .05 was used. The criterion to flag an item as biased due to effect size was that the item's measure be located in the top 25% among the items for the particular risk factor and method under consideration. The decision whether or not an item was to be removed was taken separately for each risk factor. For example, if an item was flagged as showing DIF for gender but not for age, it was removed when comparing the non-DIF versus

DIF gender split, but was not removed for the age split. To estimate the effect size of these comparisons we used Cohen's *d*.

Results

We found a high frequency of DIF when flagging items based on statistical significance (see Table 2). This was true for all four risk factors using any of the three computation methods. When the statistical significance criterion was used, we found a fair degree of convergence of DIF indicators between risk factors using the same computation procedure. The associations were usually stronger between the effect sizes of the DIF computations (see Table 3). There were two exceptions, the area/ethnicity correlation using logistic regression effect sizes and the age/gender pair using chi-square effect size. This is suggestive of risk factors that act in concert.

Table 2
Percentages of Items Flagged for DIF by Risk Factor and Year

Risk Factor	DIF detection method	1999 (of 20 items)	2000 (of 20 items)	2004 (of 40 items)
Age	Chi-square	55	75	30
	Rasch	85	70	48
	Logistic regression	60	65	43
Area	chi-square	80	80	30
	Rasch	80	85	58
	Logistic regression	55	75	45
Ethnicity	Chi-square	55	45	50
	Rasch	80	75	68
	Logistic regression	50	65	53
Gender	Chi-square	45	50	53
	Rasch	55	55	60
	Logistic regression	35	15	40

Table 3
Correlations of Effect Sizes between Risk Factors for Each DIF Detection Method

	Chi-square		Rasch		Logistic Regression	
Risk factors	Statistical significance	Effect size	Statistical significance	Effect size	Statistical significance	Effect size
Age – Area	.46*	.53*	.39*	.89*	.45*	.47*
Age – Ethnicity	.05	-.06	.28*	.44*	.20	.51
Age – Gender	.25*	.08	-.25	.38*	.18	.18
Area – Ethnicity	-.10	-.13	.02	.33*	.29*	.06
Area – Gender	.21	.04	-.29*	.43*	-.02	-.07
Ethnicity – Gender	.30*	.49*	.02	.43*	.20	.38*

*$p < .01$ (one-tailed)

We also found a fair degree of convergence between detection methods for the same risk factor. This was true when the statistical significance criterion was used to dichotomously classify items as DIF (or not DIF), although the convergence became even stronger when the effect size estimates were used instead (see Table 4). This suggests that using effect size to assess DIF improves the agreement between methods and further reinforces the idea that risk factors act in concert.

Table 4
Correlations of Effect Sizes between DIF Detection Methods for Each Risk Factor

		Rasch – Logistic regression	Chi-square – Rasch	Logistic regression – Chi-square
Age	Statistical significance	.14	.32*	.30*
	Effect size	.40*	.60*	.78*
Area	Statistical significance	.01	.40*	.09
	Effect size	.49*	.73*	.66*
Ethnicity	Statistical significance	.12	.34*	.35*
	Effect size	.52*	.60*	.80*
Gender	Statistical significance	-.16	-.25*	.36*
	Effect size	-.06	.55*	.55*

$p < .05$ (one-tailed)

We checked for the substantive impact of the bias by conducting a "purification" analysis; i.e., removing items with DIF and checking the consistency of results (see Table 5). The number of items removed using the statistical significance criterion was usually larger than the number of items removed using the effect size criterion. We estimated Cohen's *d* to compare the magnitude of the comparison of mean differences before and after item deletion. As many as 60% of the items required removal (12 items removed for area in 2000 using the statistical significance criterion) and as few as 5% (1 item removed for gender in 2000 using the significance criterion and 1 item for age and ethnicity in 2000 using the effect size criterion). The changes in Cohen's d as a consequence of removal of items were always small.

Discussion

In this study we investigated the congruence of DIF indicators for four background variables that in the Guatemalan context are risk factors for school performance in reading. This research is relevant to Guatemalan educational policy development for several reasons. Firstly, it provides information useful for further developing an assessment system that is sensitive to the needs of heterogeneous populations. Secondly, the findings support the research on cross-cultural comparisons regularly undertaken in the country, but that oftentimes lack evidence of bias-control. Thirdly, and in terms of the wider literature, it provides some support for the need to continue research on the possible impact DIF may have on assessment and how to better measure it (i.e., effect size as opposed to statistical significance).

Since DIF indicators have been found to show low consistency, we conducted our analysis using three different DIF detection procedures (chi-square, Rasch, and logistic regression), using both a statistical significance based and effect size criterion. We found a large percentage of items to be flagged for each risk factor using any of the three DIF detection procedures when the statistical significance criterion was used (see Table 2). We found some consistency across risk factors of indicators drawn with a single method and across methods of the indicators drawn for a single risk factor. In both cases the degree of congruence increased when the effect size was used instead of the statistical significance criterion. Yet, we failed to find any consequential impact from the removal of flagged items (see Table 5), either when the statistical significance criterion was used to delete items or when the deleted items were those with the largest effect sizes.

The latter finding is consistent with previous research where eliminating biased items hardly had an impact on the effect size of observed group differences (Meiring, Rothmann & Barrick, 2005; Te Nijenhuis & Van der Flier, 2009; Van de Vijver, 2011). For example,

Table 5
Cohen's d Before and After Removal of DIF item per Risk Factor

Year	1999 Number of items = 20				2000 Number of items = 20				2004 Number of items = 40			
	Age	Area	Ethnic	Gender	Age	Area	Ethnic	Gender	Age	Area	Ethnic	Gender
# items removed for statistically significant DIF	7	8	7	3	10	12	7	1	5	2	11	3
# items removed for effect size	2	2	2	3	1	2	1	3	5	2	4	6
p value of *t* test before DIF removal	< .01	< .01	< .01	0.04	< .01	< .01	< .01	0.45	< .01	< .01	< .01	0.77
p value of *t* test after DIF removal based on statistical significance selection	< .01	< .01	< .01	< .01	< .01	< .01	< .01	0.91	< .01	< .01	< .01	0.73
p value of *t* test after DIF removal based on effect size selection	< .01	< .01	< .01	< .01	< .01	< .01	< .01	0.24	< .01	< .01	< .01	0.164
Cohen's d before DIF removal	0.55	0.54	0.71	0.05	0.53	0.52	0.73	-0.02	0.41	0.67	0.96	-0.01
Change in *Cohen's d* after removal (statistical significance)	-0.05	0.01	-0.24	0.03	-0.03	-0.04	-0.04	0.01	-0.02	0.00	0.06	0.02
Change in *Cohen's d* after DIF removal (effect size)	-0.00	0.05	-0.08	0.01	0.02	0.04	-0.05	-0.01	-0.01	0.00	-0.03	-0.03

in a study conducted with South African personnel selection instruments, around 50% of the items in the cognitive tests were flagged for statistically significant DIF indicators (Meiring et al., 2005); numbers were much lower when bias was defined as medium or large effect sizes. When effect sizes were used as the classification criterion for removing items, group differences remained unaltered. However, these findings do not preempt the possibility of substantial effects of the removal of biased items in the assessment of other constructs or other cultures.

These findings suggest a paradox. On the one hand, we found that many items were biased and that, as measured by bias effect sizes, there was considerable convergence across methods. This convergence seems to provide a firm basis for item removal. However, we also found that removing biased items did not change the patterning of score differences across risk factors. Although this purification might have shown a greater impact at an earlier phase of development of the tests, in their current form we found that the removal of DIF items did not improve the adequacy of the assessment. We argue that these results point to a defining characteristic of education in Guatemala, namely differential access and opportunities of pupils. As a consequence, groups exposed to more and better education will do better on educational achievement tests. The differences in performance between the "privileged" and "underprivileged" cannot be reduced to item bias. Differences in educational gains of groups of children are so pervasive that removing items will only have a limited effect on group size differences. The poor performance of "underprivileged" children in reading achievement cannot be "recovered" by fine-tuning

items. Their poor performance is a valid reflection of their low levels of reading skills. On the other hand, our results do not suggest that adaptation of reading achievement tests for use in multicultural settings is superfluous. A poorly adapted test may well overestimate the performance differences; however, a properly designed test does not imply automatically that DIF will not occur or that performance differences between children with a different standing on risk factors will not be found.

Our findings also highlight the relevance of approaching the analysis of item bias in terms of statistical significance and effect size, and ultimately in terms of impact on group differences in score distributions. Effect sizes demonstrate a greater convergence of indicators, more clearly demonstrating the extent of the biasing effect on the assessment resulting from the interaction between risk factors and particular items, and providing a picture of the contribution of items that have not been flagged for statistical significance. However, even when effect size is used to identify the items with the greater biasing effect, item removal might not have a relevant impact on test results. Impact should be the ultimate criterion. When negotiating between the diminished biasing effects that the removal of an item might have on the test, and the loss of construct representativeness, item removal would make sense only if this purification brings about substantive changes for the interpretation of test score distributions.

Our findings lead us to believe that from a theoretical perspective it remains relevant to address multiple potential sources of bias simultaneously when developing assessment tools. In the context of Guatemala, risk factors seem to act in concert and might compound each other. As a result, they must not be addressed in isolation. The lack of substantive impact of the removal of the DIF items in this study suggests that the effect of the differential acces to educational opportunities for at risk groups is a better explanation than differences in the tests' representation of the skill domains.

Two caveats to our study are needed. First, the study centered on Guatemalan populations, and thus the findings we present are applicable to these particular groups. Second, in this study the divide for the ethnic risk factor was based on a geographical classification according to the predominant population in an area. In future studies an improved classification should be explored. This would be relevant as the diverse ethnic groups in the country assert themselves as distinct cultural groups with particular needs.

Analyzing piloted items at different stages of development would highlight more pointedly sources of DIF and the impact of the bias across risk populations. Testing the convergence of DIF across risk factors in items of tests designed to assess different curricular areas would also improve the diagnosis of the efficacy of the educational system (Teddlie & Reynolds, 2003). Furthermore, although this study spans three years following-up on items that share common specifications, not all items were identical across the years. Studies where the same set of items is analyzed across years in similar populations would provide further evidence of the stability of DIF. Lastly, the size of the tests in 1999 and 2000 were relatively small and longer tests would have provided a better picture of DIF behavior. Extension of the information could also contribute to determine how DIF varies across time and which factors increase or decrease as risks to bias.

Despite these limitations, we believe that our study provides important insights into the interaction of bias and different risk factors. From a practical standpoint our study has provided evidence to support two suggestions for test developers. First, it is important to consider multiple background variables. Risk factors seem to converge and their impact probably compounds. Therefore, analyzing bias for isolated pupil characteristics might fail to identify all relevant DIF sources. Second, using the effect size of DIF indicators provides more practical measures than their statistical significance. Effect sizes show more convergence across methods and risk factors, thus probably making DIF detection more accurate. Moreover, effect sizes are expressed on an interval or ratio scale, permitting the detection of degrees of bias for more refined analysis.

We set out to explore the consistency of DIF across risk indicators in Guatemalan reading assessment. Along the way our findings highlighted the complexity of this issue. Rather than finding a single cause for bias across measures, we found hints of ways to improve the assessment of DIF by estimating effect sizes and considering multiple sources of bias in an item. All these efforts become relevant for gaining better assessment scores through DIF analysis and purification. We believe that taking into account the aforementioned issues can help test developers to construct better assessment instruments, even in a context where differential educational opportunities create more score disparities than can be accounted for by item bias.

References

Álvarez, H., & Schiefelbein, E. (2007, December). *Informe integrado del sector educación: Informe final [Consolidated report of the education sector: Final report]*. (Report financed by the Interamerican Development Bank and the Swedish Agency for International Development as input for the strategy on Policy Harmonization and Alignment). Guatemala: Ministry of Education of Guatemala (BID / MINEDUC / ASDI).

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-23). Mahwah, NJ: Erlbaum.

Antillón Milla, J. (1997). La educación [Education]. In Asociación Amigos del País [Association of Friends of the Country] (Ed.) *Historia General de Guatemala: Vol. VI. Época contemporánea, de 1945 a la actualidad* (pp. 591-612). Guatemala: Asociación de Amigos del País, Fundación para la Cultura y el Desarrollo.

Bertrand, R., & Boiteau, N. (2003). *Comparing the stability of IRT-based and non IRT-based DIF methods in different cultural context using TIMSS data*. (EDRS Reports – Research -143-, ED 476 924, TM 034 975). Quebec, Canada: NA. (ERIC Document Reproduction Service No. ED476924).

Bond, L. (1993). Comments on the O'Neill & McPeek paper. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 277-279). Hillsdale, NJ: Erlbaum.

Bond, T. G., & Fox, Ch. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 221-256). Westport, CT: American Council on Education and Praeger Publishers.

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Orlando, FL: Harcourt Jovanovich College Publishers.

de Baessa, Y. (1999). *Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: Año 1999 [Report of results from the National Program for School Achievement Evaluation: Year 1999]*. (Report from Universidad del Valle de Guatemala of evaluation for 1999 under contract with the Ministry of Education of Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministry of Education of Guatemala.

de Baessa, Y. (2000). *Informe de Resultados del Programa Nacional de Evaluación del Rendimiento Escolar: Año 2000 [Report of results from the National Program for School Achievement Evaluation: Year 2000]*. (Report from Universidad del Valle de Guatemala of evaluation for 2000 under contract with the Ministry of Education of Guatemala) Guatemala: Universidad del Valle de Guatemala / Ministry of Education of Guatemala.

Deater-Deckard, K., Dodge, K. A., Bates, J. E., & Pettit, G. S. (1998). Multiple risk factors in the development of externalizing behavior problems: Group and individual differences. *Development and Psychopathology, 10*, 469-493. doi:10.1017/S0954579498001709

DIGEDUCA. (2008). *Bases de datos de evaluaciones de estudiantes [Data sets of student evaluations]*. Data file handed in personally. Code book retrieved from http://www.mineduc.gob.gt/digeduca/

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *The Journal of Experimental Education, 72*, 181-193. doi:10.3200/JEXE.72.3.181-193.

Dorans, N. J., & Holland, P. W. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Erlbaum.

Ellis, B. B. (1990). Assessing intelligence cross-nationally: A case for differential item functioning detection. *Intelligence, 14*, 61-78. doi:10.1016/0160-2896(90)90014-K.

Fidalgo, Á. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement, 68*, 940-958. doi:10.1177/0013164408315265.

Finch, W. H., & French, B. F. (2008). Anomalous Type I Error rates for identifying one type of differential item functioning in the presence of the other. *Educational and Psychological Measurement, 68*, 742-759. doi:10.1177/0013164407313370.

French, W. H., & Maller, S. J. (2007). Interactive purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393. doi:1177/0013164406294781.

Gerard, J. M., & Buehler, C. (1999). Multiple risk factors in the family environment and youth problem behaviors. *Journal of Marriage and Family, 61*, 343-361. doi:10.2307/353753

Hidalgo, M., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*, 903-915. doi:10.1177/0013164403261769.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

Jin-Shei, L., Teresi, J., & Gershon, R. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation & the Health Professions, 28*, 283-294. doi:10.1177/0163278705278276.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*, 329-349. doi:10.1207/S15324818AME1404_2.

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response items. *Educational and Psychological Measurement, 65*, 935-953. doi:10.1177/0013164405275668.

Linacre, J. M. (2006). WINSTEPS (Version 3.65) [Computer software]. Chicago, IL: Winsteps.com

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 349-364). Hillsdale, NJ: Erlbaum.

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 171-196). Hillsdale, NJ: Erlbaum.

Meiring, D., van de Vijver, F. J. R., Rothmann, S., & Barrick, M. R. (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *SA Journal of Industrial Psychology, 31*, 1-8.

Ministry of Education of Guatemala (2009). *Yearly National Guatemalan educational statistics from the Ministry of Education: Year 2008*. Retrieved from http://www.mineduc.gob.gt/estadistica/2008/main.html

Ministry of Education of Guatemala (2011). *Yearly National Guatemalan educational statistics from the Ministry of Education: Year 2011 final version*. Retrieved from http://www.mineduc.gob.gt/estadistica2011/

Moreno-Grajeda, M. R., Gálvez-Sobral, J. A., Bedregal, S., & Roldán, K. (2008, April). Informe de Resultados, Evaluación de Primaria 2006, Tercer Grado. *[Report of Results, Primary Level Evaluation 2006, Third Grade]*. (Report from the Unit for Statistical Analysis of DIGEDUCA of Ministry of Education of Guatemala).

Narayanan, P., & Swaminathan, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement, 18*, 315-328. doi:10.1177/014662169401800403.

O'Neill, K. A., & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 255-276). Hillsdale, NJ: Erlbaum.

Penfield, R. D. (2003). Applying the Breslow-Day test of trend in odds ratio heterogeneity to the analysis of non-uniform DIF. *Alberta Journal of Educational Research, 49*, 99-112.

Penfield, R. D. (2007a). *Differential item functioning analysis system –DIFAS 4.0-: User's manual*. Gainesville, FL.

Penfield, R. D. (2007b). Differential item functioning analysis system –DIFAS 4.0- [Computer software]. Gainesville, FL.

Richards, M. (2003). *Atlas lingüístico de Guatemala [Linguistic atlas for Guatemala]*. Guatemala: SEPAZ, UVG, URL, USAID.

Robitzsch, A., & Rupp, A. (2009). Impact of missing data on the detection of differential item functioning. *Educational and Psychological Measurement, 69*, 18-34. doi:10.1177/0013164408318756.

Rutter, M. (1979). Protective factors in children's responses to stress and disadvantage. In M. W. Kent & J. E. Rolf (Eds.), *Primary prevention of psychopathology: Social competence in children* (Vol. 3, pp. 49-74). Hanover, NH: University Press of New England.

Rutter, M. (1988). Longitudinal data in the study of causal processes: Some uses and some pitfalls. In M. Rutter (Ed.). *Studies of psychosocial risk: The power of longitudinal data*. (pp. 1-28). Cambridge: Cambridge University Press.

Rutter, M. (2001) Psychosocial adversity: Risk, resilience and recovery. In. J. M. Richman & M. W. Fraser (Eds.), *The context of youth violence. Resilience, risk and protection* (pp. 13-42). Westport, CT: Praeger Publishers.

Sameroff, A. J., Bartko, W. T., Baldwin, A., Baldwin, C., & Seifer, R. (1998). Family and social influences on the development of child competence. In M. Wei & C. Fairing (Eds.), *Families, risk, and competence* (pp. 161-183). Mahwah, NJ: Erlbaum.

Scherbaum, C. A., & Goldstein, H. W. (2008). Examining the relationship between race-based differential item functioning and item difficulty. *Educational and Psychological Measurement, 68*, 537-553. doi:10.1177/0013164407310129.

Scheuneman, J. D. (1987). An experimental, exploratory study of causes of bias in test items. *Journal of Educational Measurement, 24*, 97-118. doi:10.1111/j.1745-3984.1987.tb00267.x

Scheuneman, J. D., & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*, 109-131. doi:10.1111/j.1745-3984.1990.tb00737.x

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*, 53-75. doi:10.3102/10769986027001053

te Nijenhuis, J., & van der Flier (2009). Bias research in the Netherlands: Review and implications. *European Journal of Psychological Assessment, 15*, 165-175. doi:10.1027//1015-5759.15.2.165

Teddlie, Ch., & Reynolds, D. (2003). T*he international handbook of school effectiveness research*. London, United Kingdom: Falmer Press.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Uiterwijk, H., & Vallen, T. (2003). Test bias and differential item functioning: A study on the suitability of the CITO primary education final test for second generation immigrant students in the Netherlands. *Studies in Educational Evaluation, 29*, 129-143. doi:10.1016/S0191-491X(03)00019-1

United Nations Development Programme (2011). Human development report 2011; *Sustainability and equity: A better future for all*. New York, NY: United Nations Development Programme.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54*, 119-135. doi:10.1016/j.erap.2003.12.004

van de Vijver, F. J. R. (2011). Bias and real differences in cross-cultural differences: Neither friends nor foes. In F. J. R. van de Vijver, A. Chasiotis & S. M. Breugelmans (Eds.), *Fundamental questions in cross-cultural psychology* (pp. 235-257). New York, NY: Cambridge University Press.

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*, 443-464. doi:10.3102/10769986030004443

Welkenhuysen-Gybels, J. (2003). *The detection of differential item functioning in Likert score items*. Unpublished manuscript. Leuven, Belgium: Catholic University of Leuven.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods* (EM No. 60, ISSN 1103-2685). Umeå, Sweden: Umeå Universitet.

World factbook. (2011). USA: CIA. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/gt.html

Zenisky, A. L., Hambleton, R. K., & Robin, F. (2003). Detection of differential item functioning in large-scale state assessment: A study evaluating a two-stage approach. *Educational and Psychological Measurement, 63*, 51-64. doi:10.1177/0013164402239316

Received:11/10/2012
Accepted:01/01/2014

Alvaro M. Fortin Morales. Universidad del Valle de Guatemala, Guatemala and Tilburg University, Netherlands
Fons J. R. van de Vijver. Tilburg University, Netherlands and North-West University, South Africa
Ype H. Poortinga. Tilburg University, Netherlands