Equivalence of Testing Instruments in Canada: Studying Item Bias in a Cross-Cultural Assessment for Preschoolers

Luana Marotta
Stanford University

Lucia Tramonte
University of New Brunswick

J. Douglas Willms
University of New Brunswick

Abstract

Item bias, which occurs when items function differently for different groups of respondents, is of particular concern to cross-cultural assessments. It threatens measurement equivalence and causes intergroup comparisons to be invalid. This study assessed item bias among francophone, anglophone, and Aboriginal preschoolers in New Brunswick, Canada. We used data from the Early Years Evaluation-Direct Assessment (EYE-DA), an assessment tool that measures children's early educational development. The analytical approach used to investigate item bias is called differential item

Canadian Journal of Education / Revue canadienne de l'éducation 38:3 (2015)

www.cje-rce.ca

functioning (DIF). This study offers an application of DIF analysis that combines statistical testing and graphical representation of DIF. Analyses yielded consistent results revealing that linguistic and cultural differences between francophone and anglophone children are more challenging to achieve transferability than cultural differences between Aboriginal and anglophone examinees.

Keywords: measurement equivalence, cross-cultural testing, item bias, differential item functioning, early childhood assessment

Résumé

Le biais d'item, qui a lieu quand des items fonctionnent différemment selon les caractéristiques des groupes de répondants à un test, pose de sérieux problèmes aux évaluations interculturelles. Le biais d'item affecte l'équivalence des mesures et rend les comparaisons entre groupes invalides. Cet article offre un exemple d'étude de biais d'item dans le contexte d'un test administré aux enfants francophones, anglophones, et aborigènes dans la province du Nouveau Brunswick, au Canada. Notre étude utilise des données provenant de l'Évaluation de la petite enfance-Appréciation directe (ÉPE-AD), un instrument d'évaluation qui mesure les habiletés des enfants de 3 à 6 ans dans leur transition à l'école. La méthode d'analyse utilisée pour évaluer le biais d'item s'appelle fonctionnement différentiel d'items (FDI). Cet article présente une application de l'analyse du FDI qui combine une approche statistique et une représentation graphique du FDI. Les résultats de notre étude montrent que les différences linguistiques et culturelles entre enfants anglophones et francophones sont plus remarquables que celles entre enfants aborigènes et enfants anglophones. Ces différences affectent la possibilité de transférer directement l'instrument d'évaluation d'une langue à l'autre.

Mots-clés : équivalence des mesures, test interculturel, biais d'item, fonctionnement différentiel d'items, évaluation de la petite enfance

Introduction

Measurement equivalence is of special concern when studies involve salient groupings with diverse cultural and linguistic backgrounds. Measurement tools are not "context free" (Zumbo, 2009, p. 76): social measures often refer to objects, events, or situations in social life whose meaning and significance are not culturally stable. Instruments may also have their meaning affected after being translated to multiple languages. Greater measurement variance is likely to occur when the differences among groups are large; cross-national researchers, for example, often encounter methodological problems of measurement equivalence, as shown by Davidov, Schmidt, and Schwartz (2008) and Kankaras and Moors (2009). If survey instruments do not measure the same thing across different contexts, results are not comparable and inference about group differences is misleading.

Bias is used as a general term to represent the lack of correspondence between measures applied to different cultural groups (Van de Vijver & Poortinga, 1997). Measurement equivalence is threatened by different forms of bias such as construct bias (when the meaning of the studied construct varies among cultural groups), sample bias (when samples involving different cultural groups are not equally representative of the populations under study), instrument bias (when characteristics of instruments induce measurement error for particular groups of respondents), and administration bias (when measurement error stems from the interaction between survey administrators and respondents of particular groups) (Van de Vijver, 2000). This study focuses on a specific form of bias called *item bias*, which refers to differences in item-level responses that occur when items function differently for certain groups of respondents.

Item bias stems from contextual and communication factors. In educational assessments, for example, test items are biased when their content is not equally familiar to groups of students exposed to different curricula (an example of contextual source of bias). When students who speak different languages are evaluated, items may work differently if they are poorly translated (a communication source of bias).

This article investigates item bias in an assessment tool that is widely administered to Canadian preschoolers. Few studies have been done about instrument equivalence among young children (He & Wolfe, 2010). Item bias may, however, be more likely to occur in early childhood assessments that involve children with different linguistic and cultural backgrounds. Poor translation often increases word difficulty, which can be

particularly challenging for young pupils. Items with culturally specific content, another common source of item bias, may also be more difficult for young children than teenagers or adults, who are more likely to have learned elements from diverse cultures.

The assessment tool under study is the Early Years Evaluation-Direct Assessment (EYE-DA). The EYE-DA is an individually administered direct measure of the developmental skills and school preparedness of preschoolers aged three to five years old. The EYE-DA has been used by school districts across Canada and in a number of research studies. This study analyzed data from New Brunswick, Canada. The EYE-DA samples under consideration comprise over 13,000 children from anglophone, francophone, and Aboriginal communities[1]. The study demonstrates empirically whether the instrument's properties are affected when the EYE-DA evaluation is used in different cultural and linguistic contexts.

Quantitative assessments of Aboriginal students are controversial. Over the past years, psychologists and educators have claimed that standardized tests are invalid for students of minority cultures, including Aboriginals (Gopaul-McNichol & Armour-Thomas, 2002; Philpott, Nesbit, Cahill, & Jeffery, 2004). According to Gopaul-McNichol and Armour-Thomas (2002), standardized tests are consistent to the values of the dominant culture and do not reflect learning experiences of minority children, who are unfairly penalized.

Assessments, however, have become a crucial tool in education reform. They are expected to evaluate the needs of *all* students. In the context of cultural pluralism in a democratic society such as Canada, educational equity for culturally diverse groups is a critical goal. Of the three groups assessed by the EYE-DA, Aboriginals and francophones are among New Brunswick's minorities—studies reveal that both groups have performed below the national average in reading and mathematics (Bussiere et al., 2001; Carr-Stewart, 2006). An important issue concerns the extent to which differences in test performance among groups are attributable to actual differences in reading and mathematics achievement versus bias associated with the instruments. Analysis of item bias in the EYE-DA is crucial to having a better understanding of measurement variance in educational tests involving Canadian children.

1 Aboriginal preschoolers assessed by the EYE-DA were enrolled in band-operated schools, locally controlled by First Nations (Labercane & McEachern, 1995).

The analytical approach used to examine item bias is called differential item functioning (DIF; Holland & Wainer, 1993). DIF investigates whether those of different social groups who hold like values with respect to the construct being measured provide different response patterns to the same question. Most analytical approaches consist of statistical tests that assess whether there is DIF and if so, measure the magnitude of its effect—what Scrams and McLeod (2000) call a "global" approach to DIF. Some of the most popular approaches include IRT (Thissen, Steinberg, & Wainer, 1993), the Mantel-Haenzel (Holland & Thayer, 1988), the standardization procedure (Dorans & Kulick, 1986), and the logistic regression method (Swaminathan & Rogers, 1990). Graphical DIF approaches have been used occasionally in the literature (Pashley, 1992; Douglas, Stout, & DiBello, 1996; Scrams & McLeod, 2000; Bolt & Gierl, 2006; Willms, Tramonte, & Chin, 2007). This study offers an approach to DIF analysis that combines statistical testing with the graphical representation of DIF.

A more extensive discussion of item bias follows. Next, we describe methodological issues involved in the empirical investigation. Research findings are offered in the third section and the article closes with the overall conclusions and implications of the study.

Measurement Equivalence at the Item Level

Survey or test items are biased when they systematically fail to measure what they are intended to measure, which is of particular concern when diverse social groups are involved. Item bias violates the assumption of measurement invariance, which holds that measurement properties should not be affected by the context (Zumbo, 2009). In the literature, item bias has been examined in "latent variable analysis," in which multiple variables are combined into a smaller number of latent factors. Item bias in latent analysis threatens "scalar equivalence" (Hui & Triandis, 1985; Van de Vijver & Tanzer, 2004). Scalar equivalence presumes that a value on a scale or index refers to the same degree, intensity, or magnitude of the construct for every respondent regardless of their group affiliation (Hui & Triandis, 1985). Scaling of items presupposes, therefore, that item characteristics are similar for all individuals under study (Schaeffer, 1988).

Item bias occurs when elements not relevant to test purposes (e.g., poor translation and curriculum differences) affect the probability of a correct answer from examinees with comparable ability. Item bias is different from item impact, which occurs when groups of examinees have different probabilities of responding correctly to an item because these groups in fact differ in the ability measured by the item.

Item bias can result from contextual or communication issues. The former refers to circumstances in which attributes of the social structure affect the significance of item content so that the item does not measure what it is expected to measure. In other words, contextual differences influence the item's capacity to equally reflect the underlying construct being measured in different populations. Communication problems occur where ambiguities of meanings are concerned; that is, when interpretation by respondents is affected by group affiliation[2].

When bias results from contextual issues, the item content becomes inappropriate with respect to the construct being measured after the item is adapted to a different context, as shown in Figure 1. When bias results from communication issues, the item content is still appropriate for the survey purposes after the item is adapted to another context; yet, the way in which the item content is communicated is incorrect, causing the item to be invalid, as shown in Figure 2.

2 He and Wolfe (2010) classify sources of item bias into linguistic and cultural features. According to the authors, cultural features refer to sociocultural differences among different population. "Contextual issues" described herein refer to broader structural features affecting instrument transferability. Suppose, for example, that in a cross-national survey people are asked questions about their financial worth. Imagine that among the items associated with home possessions there is one that asks respondents whether or not they own a dryer. North American researchers, used to the presence of dryers in their homes, might expect that individuals with high economic status are more likely to own a dryer than individuals with low economic status. Dryers, however, are not customarily in domestic use in sunny countries and may not be an indicator of socio-economic status. In this case, differences in the weather, not sociocultural differences, are the main source of item bias.

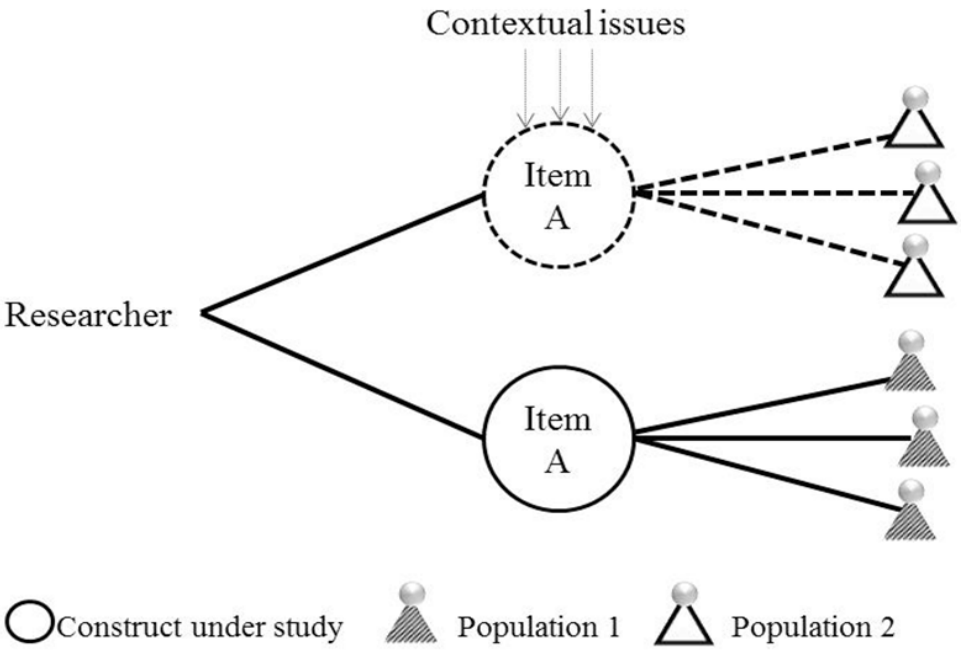

Figure 1. Item bias stemming from contextual issues.

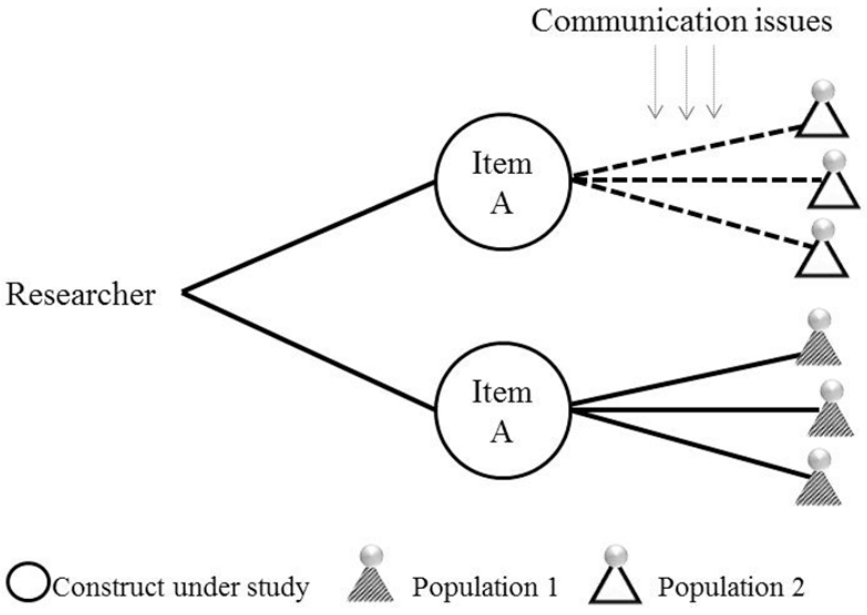

Figure 2. Item bias stemming from communication issues.

An example of item bias stemming from contextual differences is provided by Batista-Foguet, Fortiana, Currie, & Villalbí (2004), who demonstrate that some items included in the Family Affluence Scale (FAS) are not equally weighted across countries. They show, for example, that the item "number of holidays spent by the family" has a different impact on the construct "family wealth" when FAS is measured internationally. The amount of time available for parents to take holidays varies from country to country depending on vacation extensions and number of public holidays. Travelling fewer times in any country is, therefore, not necessarily an indication of low family wealth. It may be a consequence of restricted leisure opportunities.

The problem of item bias is not confined to contextual differences. It may also derive from communication problems. When multiple-language versions of survey instruments are involved, communication problems are likely to occur. Poor translation of items may also affect word or phrase complexity. Ercikan (1998), who explored item bias in her work for the International Association for the Evaluation of Educational Achievement (IEA), discovered differences of interpretation among Canadian English- and French-speaking students. According to Ercikan, 18 IEA test items were biased; that is, they did not function equally for anglophone and francophone examinees, as is illustrated by the following example:

> In one item, a particular word was determined to be in more common usage in French than in English. In English "the animal preyed on other animals" was translated into French as "the animal fed on other animals." The fact that "prey" is a less common word than "fed" could have made the item more ambiguous and difficult for the English-speaking group. (Ercikan, 1998, p. 548)

Because communication problems and contextual differences that affect item functioning can be anticipated, procedures to identify and minimize item bias can be adopted. One way to detect item bias is to pre-test survey questions (Greenfield, 1997; Wirth & Kolb, 2004). Once data have been collected, item bias can be explored by judgemental procedures (methods that rely solely on expert judges to select potentially biased items) and by using statistical methods such as differential item functioning (DIF), which has been employed in this study.

DIF examines item bias in latent variable analysis. The construct under study must, therefore, be measured by multiple items. From a statistical-methodological

perspective, DIF occurs when respondents from different groups show differences in probability of item endorsement after having been matched according to the trait being measured (Zumbo, 1999). Individuals are usually expected to display similar response patterns on items that measure the same construct. When the DIF technique is used, item bias is signalled when response patterns change systematically among different groups with similar trait scores.

Although often used interchangeably, the terms "DIF" and "item bias" are not synonymous, as Clauser and Mazor (1998) rightly affirm. DIF is merely a statistical tool to investigate item bias—a technique, as Zumbo (1999) points out, that only flags potentially biased items. Statistical flags such as DIF should not be used to automatically discard items (Shepard, Camilli, & Williams, 1985). Before concluding that an item is biased, researchers must provide a more complete account of the nature of bias by employing other analytical approaches, including, for example, judgemental evaluation by experts.

Methods for investigating item bias are typically used when new measures are created, when existing measures are adapted to new contexts or for different populations not initially intended when the measures were developed, or when existing measures are translated (Zumbo, 2007). As has been said, pre-tests may be also useful to detect group differences that affect measurement invariance.

This empirical investigation did not involve judgemental evaluation of EYE-DA test items. Analysis was confined to statistical flagging of item bias using the DIF approach. A more detailed account of the data analysis follows.

Data Analysis

Potential item bias in the EYE-DA evaluation was explored using data from 2008–9 and 2009–10, the first two cycles of New Brunswick's provincial assessment. Of the 2008–9 sample, 4,857 were anglophone; 1,747 were francophone; and 97 were Aboriginal (6,701 children in total). Of children assessed in 2009–10, 4,205 were anglophone; 1,774 were francophone; and 86 were Aboriginal (6,065 in total). The EYE-DA samples comprise approximately 97% of preschool children in the province. As Jodoin and Gierl (2001) observed, the power to detect DIF may decrease in unbalanced conditions. Because the

number of anglophone students is higher than that of francophone and Aboriginal students, the procedures used in this study to flag DIF might underestimate item bias. To assure that results are reliable, we examined item bias in the EYE-DA using statistical testing and graphical representation of DIF, and conducted analysis on the 2008–9 data and used the 2009–10 data to cross-validate the results.

The EYE-DA, administered prior to kindergarten entry, assesses the learning needs of children and helps educators identify those who may need assistance to develop the skills appropriate to kindergarten and Grade 1. The test evaluates four developmental domains associated with school preparedness and emerging literacy skills (Willms, 2009):

- Domain A – *Awareness of Self and Environment*, for example, recognition of colours, animals, feelings, and positional concepts.
- Domain B – *Cognitive Skills*, for example, knowledge of the alphabet, counting, reasoning, and grouping similar items.
- Domain C – *Language and Communication Skills*, for example, knowledge of receptive and expressive vocabulary, sentence completion, and oral communication.
- Domain D – *Gross and Fine Motor Development*, for example, walk backwards, skip forward, cutting, and drawing.

Each domain contains 12 items, which are identical in the 2008–9 and 2009–10 EYE-DA cycles.

In this study, DIF analysis was carried out with the logistic regression method, which allows estimation of DIF effect size and tests uniform and non-uniform DIF. Uniform DIF occurs when the difference in probability of correct answer between the focal and the reference groups of respondents of comparable ability is constant along the trait continuum. Non-uniform DIF takes place when the difference in probability of correct answers between these groups is not the same at all ability levels (Clauser & Mazor, 1998). Uniform and non-uniform DIF were examined herein. Equation 1 shows the logistic model used to detect DIF.

Equation 1

General logistic model for DIF detection

$$ln\left[\frac{p_i}{1-p_i}\right] = b_0 + b_1 Ability + b_2 FocalGroup + b_3 Ability * FocalGroup$$

where "pi" is the proportion of examinees who completed item "i" consistently and received the full points.[3] "Ability" refers to the children's total score on the domain measured by the item "i."[4] EYE-DA domains were analyzed separately. "FocalGroup" is a dummy variable that identifies the examinee's group membership: anglophone children, the reference group, were coded "zero"; francophone and Aboriginal children, the focal groups, were coded "one." The product of the dummy variable "FocalGroup" and the predictor "Ability" is an interaction term, which tests for non-uniform DIF.

Significance of DIF and its size effect were assessed by means of statistical testing.

Statistical Testing of DIF

DIF significance was tested by means of a chi-square test, as suggested by Swaminathan and Rogers (1990). The chi-square statistic was calculated for three logistic models:

Model #1 (no DIF): $Y = b_0 + b_1 Ability$

Model #2 (uniform DIF): $Y = b_0 + b_1 Ability + b_2 FocalGroup$

Model #3 (non-uniform DIF): $Y = b_0 + b_1 Ability + b_2 FocalGroup + b_3 Ability{*}FocalGroup$

Uniform DIF was deemed significant when the chi-square difference between models #2 and #1 was bigger than the critical value of chi-square with one degree of freedom, at a significance level of 0.05. DIF was considered non-uniform when the chi-square difference between models #2 and #3 was statistically significant.

The major shortcoming of the chi-square test is that trivial DIF effects may be statistically significant when the test is based on a large sample size (Zumbo, 1999, p. 27). Zumbo and Thomas (1997) suggest, therefore, that the magnitude of DIF be also taken into account. Effect size was examined by comparing the Nagelkerke R-square of each model above using the same strategy applied to the chi-square test. We used two available guidelines to classify DIF in logistic regression models, one defined by Zumbo and

3 The EYE-DA items are scored on a scale that ranges from 0 to 3. To facilitate interpretation of analyses, items were dichotomized—responses zero to two were scored as "0" and three was scored as "1"—and DIF detection accomplished by binary logistic regression.

4 The children's scores were estimated by means of IRT graded-response models (Samejima, 1970).

Thomas (1997) and a less conservative approach offered by Jodoin and Gierl (2001). The effect size can be considered "negligible," "moderate," or "large":

Table 1. Classification Criteria.

DIF Effect Size	Zumbo-Thomas	Jodoin-Gierl
Negligible	$R^2\Delta < 0.13$	$R^2\Delta < 0.035$
Moderate	$0.13 \leq R^2\Delta \leq 0.26$	$0.035 \leq R^2\Delta \leq 0.070$
Large	$R^2\Delta > 0.26$	$R^2\Delta > 0.070$

Note: R2Δ in the test for uniform DIF represents the incremental R2 between model #1 and model #2; in the test for non-uniform DIF, R2Δ represents the difference between the R2 of model #3 and model #2.

If the chi-square and at least one R-square test pointed to existence of DIF, a cross-validation sample was used to confirm the results. DIF was considered to be consistent if statistical tests flagged an item as biased in the 2008–9 and 2009–10 EYE-DA data.

Purification of the matching criterion was used in the process of conducting DIF analysis (Holland & Thayer, 1988). That is, whenever the chi-square and at least one R-square test pointed to existence of DIF in both EYE-DA cycles, the ability trait was recalculated without the biased items, and DIF was tested again for all of the other items.

An item was declared to be potentially biased if DIF was considered to be significant by the chi-square test, relevant by at least one R-square test, and consistent if the results were cross-validated by DIF analysis on the 2009–10 sample. A more conservative approach to data analysis was necessary insofar as DIF methods may yield unstable results when flagging biased items (Hambleton & Jones, 1995).

Graphical Representation of DIF

In addition to statistical testing, graphical representation of DIF was also conducted. Developed by Willms, Tramonte, and Chin (2007), the graphical representation of DIF suggested in this study has been found to be helpful for visualizing the location of DIF (the location in the trait continuum at which DIF takes place), the direction of DIF (the direction indicating which subgroup is potentially favoured by item bias), its impact (the proportion of respondents being affected by DIF), and its relative size. The DIF size can be estimated by subtracting the probability of a correct answer of the focal group from that of the reference group for all values of ability along the trait continuum.

Results

The chi squared statistic was highly influenced by large sample sizes and trivial DIF effects were often significant, as shown in Table 2. No item was flagged as potentially biased based on the Zumbo-Thomas measure of effect size for R squared. According to the Jodoin-Gierl R-squared test, four items displayed uniform DIF. No item displayed significant non-uniform DIF.

Table 2. Number of Items Flagged as Biased by Chi- and R-squared Tests.

Cultural Groups	DIF Type	Test*	Domain				
			A	B	C	D-Fine	D-Gross
Francophone and Anglophone	Uniform	Chi-squared	11	8	10	4	4
		ZT R-squared	0	0	0	0	0
		JG R-squared	1	1	2	0	0
	Non-uniform	Chi-squared	3	2	2	0	3
		ZT R-squared	0	0	0	0	0
		JG R-squared	0	0	0	0	0
Aboriginal and Anglophone	Uniform	Chi-squared	2	3	3	0	0
		ZT R-squared	0	0	0	0	0
		JG R-squared	0	0	0	0	0
	Non-uniform	Chi-squared	2	1	0	2	0
		ZT R-squared	0	0	0	0	0
		JG R-squared	0	0	0	0	0

*ZT (Zumbo and Thomas) and JG (Jodoin and Gierl) tests.

**The total number of items in domains A, B, and C is 12; the total number of items in domain D, fine and gross motors, are 7 and 5, respectively.

Table 3 shows the uniform DIF results for the four items flagged as potentially biased. As mentioned above, DIF is only found when the Jodoin-Gierl R-squared test is considered. The results are consistent across the 2008–9 and 2009–10 EYE-DA cycles.

Table 3. Chi-squared and R-squared Tests of Items Flagged as Potentially Biased.

Domain-Item	Year	Chi-squared Test		R-squared test		
		Difference	Significance at 0.05	Difference	Significance*	
					ZT	JG
a5	2008–9	235.027	Yes	0.040	Neg	Mod
	2009–10	201.735	Yes	0.040	Neg	Mod
b7	2008–9	349.819	Yes	0.041	Neg	Mod
	2009–10	438.880	Yes	0.059	Neg	Mod
c6	2008–9	427.460	Yes	0.057	Neg	Mod
	2009–10	591.131	Yes	0.093	Neg	Larg
c9	2008–9	241.488	Yes	0.038	Neg	Mod
	2009–10	272.454	Yes	0.050	Neg	Mod

*ZT (Zumbo and Thomas) test. JG (Jodoin and Gierl) test. Negligible (Neg); moderate (Mod); and large (Larg) DIF.

On the whole, DIF seems to be more relevant to comparisons between francophone and anglophone children than between Aboriginal and non-Aboriginal anglophone examinees. Apparently, DIF size of items in domain D—which assesses children's physical development—is smaller than DIF size of the items in the other domains—which include awareness of self and environment, language and communication, and cognitive skills.

Below, DIF is graphically represented for all the items that displayed uniform DIF by test domain. The y-axis indicates the DIF size, which was estimated by subtracting the probability of the correct answer of the focal group from that of the reference group for all values of ability along the trait continuum. The x-axis shows the location in the ability continuum at which DIF takes place. The graph also shows the direction of DIF, showing which subgroup is potentially favoured by item bias. The histogram represents the distribution of the students' total score—which combines their answer to all items in the domain under consideration using IRT graded-response models (Samejima, 1969) and indicates the proportion of examinees being affected by DIF. Finally, the graph shows the overall DIF, which was estimated by calculating the average of DIF size for all of the items. The overall DIF is useful to indicate whether the test as a whole benefits a specific group of examinees. The graphs use data from the EYE-DA 2008–9.

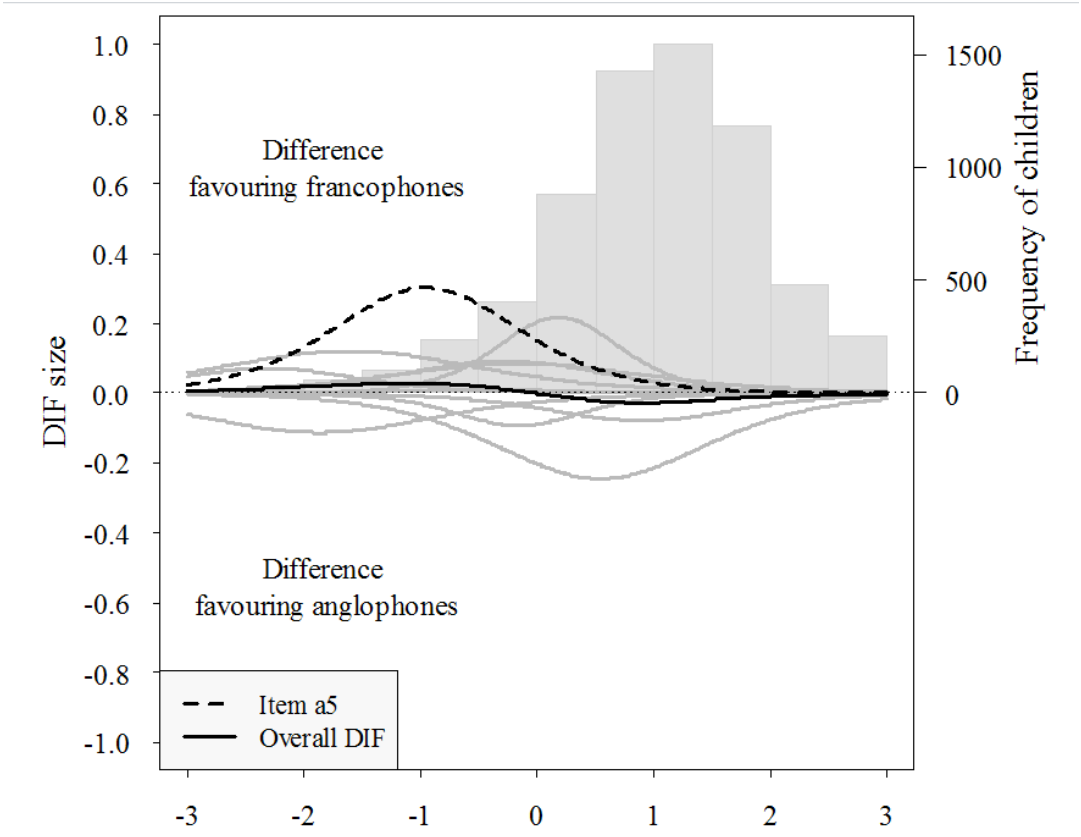

Figure 3. Graphical representation of DIF – Domain A.

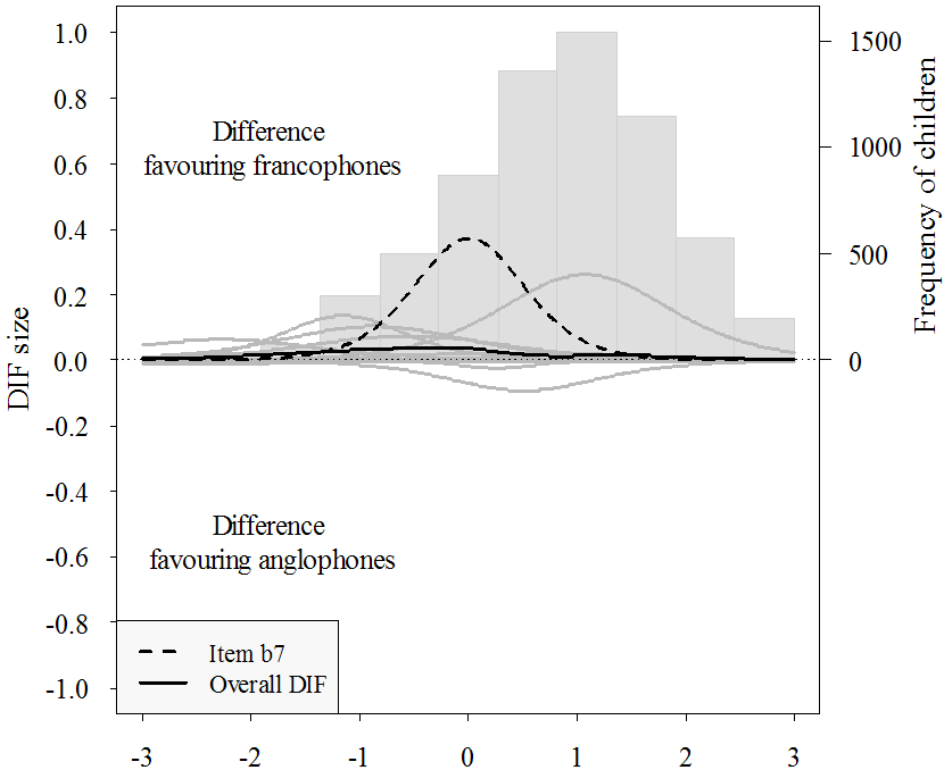

Figure 4. Graphical representation of DIF – Domain B.

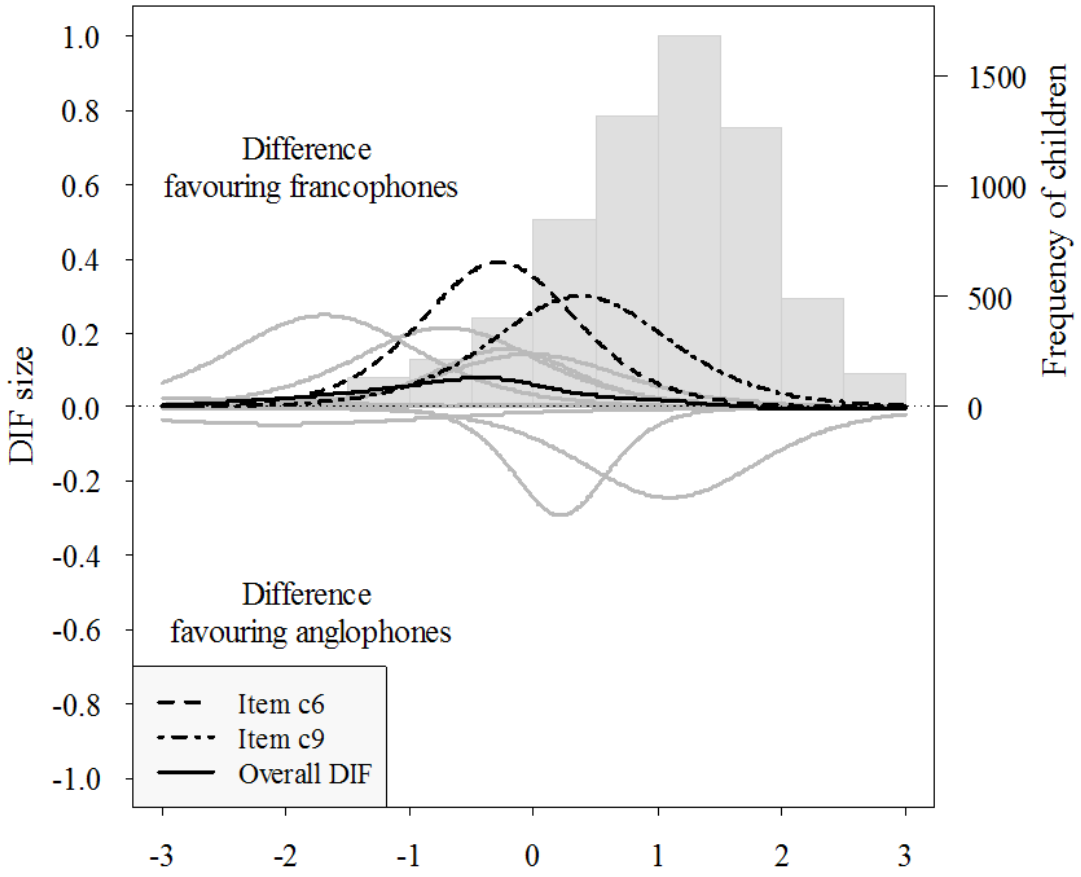

Figure 5. Graphical representation of DIF – Domain C.

In all items flagged as potentially biased, francophone students had higher probability of a correct answer than their anglophone peers with comparable ability level. Moreover, there is little overlap between the curves of DIF and the distribution of students' ability, showing that the majority of children were not affected by item bias. Generally, DIF was higher in the left side or in the middle of the ability continuum; whereas the score distribution is concentrated in the right-hand of the scale. DIF tended to affect, however, children with lower scores, having an impact on the measure quality of those who are the most vulnerable and should therefore be considered for early interventions. The overall DIF in each EYE-DA domain is low along the trait continuum.

Conclusion

This article explored the problem of measurement equivalence in the EYE-DA, an assessment tool that measures children's early educational development. The EYE-DA has been used in several different contexts and in several different languages, including English, French, Chinese, and Spanish. In each case, tests of item bias are conducted to discern whether the assessment exhibits DIF for differing populations and subpopulations. As has been said, quantitative assessments of minorities are controversial; for example, several authors admonish inappropriate assessments that disregard social, cultural, and language differences of Aboriginal students (Philpott, Nesbit, Cahill, & Jeffery, 2004). This article attempted to shed light on this question by investigating item bias in a cross-cultural test of Aboriginal, anglophone, and francophone students.

Results revealed that linguistic and cultural differences between francophone and anglophone children appear more problematic with respect to instrument transferability than cultural differences between Aboriginal and non-Aboriginal anglophone children. Educators may argue that educational assessments deal with content unfamiliar to Aboriginal children and that they are therefore unfairly disadvantaged. These results indicate, however, that educational assessments can be suitable to Aboriginal children. In cases where score disparities are evident, they may be better explained by differences in socio-economic factors, rather than by problems with the instrument.

Future research is necessary to better understand the problem of measurement equivalence in educational assessments in Canada. The sample analyzed in this article, which comprises preschoolers of New Brunswick, is diverse because it involves children from three cultural groups: anglophone, francophone, and Aboriginal children in New Brunswick. However, there are over 600 First Nations bands in Canada with markedly different cultures, and with over 60 different languages (Statistics Canada, 2013). Therefore, we cannot claim that the results from this study can be generalized to all Aboriginal children across Canada.

Similarly, different results may be found if other francophone communities in Canada are surveyed. For example, francophone children in Quebec, where 79% of the population only speaks French, may differ considerably from francophone children in New Brunswick, where 32% of the population only speaks French (Statistics Canada, 2006). Moreover, Canadian-French vocabulary varies across provinces, as Emenogu and

Childs (2003) point out. As a result, test translations may be more or less appropriate depending on the francophone community where the instrument is administered. Analysis of item bias may therefore yield different results.

This study also provides an important contribution to the discussion about instrument equivalence among young children, which has not been extensively explored (He & Wolfe, 2010). Results suggest that item bias due to translation issues may be crucial in early childhood assessments that involve children with different linguistic backgrounds. Items from domain C, which assessed language and communication skills, seemed more subjected to bias than other domains with more straightforward directions—domains A (which required children to identify colours and animals, for example), B (with questions on counting, reasoning, and grouping similar items), and D (which assessed children's physical development). Perhaps the same items that evaluate language and communication skills would not display DIF if the instrument were administered to older children, for whom poor translation might not increase word difficulty.

The Early Years Evaluation tools are designed to provide leading indicator data that educators can use to maximize the accuracy of the planning and design decisions they make. It was not designed to diagnose specific learning problems or to make comparisons among provinces, school districts, or schools, or between cultural groups. The Learning Bar Inc., which administers the Early Years Evaluation, stresses the importance of using the data to assess changes in results year over year for the same community, and to use the results to gauge the kinds of resources each child needs to succeed at school, both socially and academically. Nevertheless, school administrators at the district and provincial levels use EYE data to make decisions about the allocation of resources and the effects of various interventions aimed at reducing inequalities. For these purposes, empirical analysis of item bias is crucial for ensuring that test instruments employed are equally applicable among populations with differing cultural and linguistic backgrounds. With increased use of standardized tests in school reforms, analysis of measure equivalence must guarantee that administrators are supplied with accurate information about children's performance, especially when minorities are involved.

References

Batista-Foguet, J. M., Fortiana, J., Currie, C., & Villalbí, J. R. (2004). Socio-economic indexes in surveys for comparisons between countries. S*ocial Indicators Research*, *67*(3), 315–332.

Bolt, D. M., & Gierl, M. J. (2006). Testing features of graphical DIF: Application of a regression correction to three nonparametric statistical tests. *Journal of Educational Measurement*, *43*(4), 313–333.

Bussiere, P., Cartwright, F., Crocker, R., Ma, X., Oderkirk, J., & Zhang, Y. (2001). *Measuring up: The performance of Canada's youth in reading, mathematics and science—OECD PISA Study—First results for Canadians aged 15*. Ottawa, ON: Human Resources Development Canada, Council of Ministers of Education, Canada and Statistics Canada.

Carr-Stewart, S. (2006). First Nations education: Financial accountability and educational attainment. *Canadian Journal of Education*, *29*(4), 998–1018.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44.

Davidov, E., Schmidt, P., & Schwartz, S. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly*, *72*(3), 420–445.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*(4), 355–368.

Douglas, J. A., Stout, W., & DiBello, L. V. (1996). A Kernel-smoothed version of SIBTEST with applications to local DIF inference and function estimation. *Journal of Educational and Behavioral Statistics*, *21*(4), 333–363.

Emenogu, B., & Childs, R. A. (2003). *Curriculum and translation differential item functioning: A comparison of two DIF detection techniques*. Retrieved from http://eric.ed.gov/PDFS/ED476424.pdf

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543–553.

Gopaul-McNichol, S., & Armour-Thomas, E. (2002). *Assessment and culture: Psychological tests with minority populations*. San Diego, CA: Academic Press.

Greenfield, P. M. (1997). Culture as process: Empirical methods for cultural psychology. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology: Vol. 1. Theory and method*. Boston, MA: Allyn & Bacon, 301–346.

Hambleton, R. K., & Jones, R. W. (1995). Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly*, *18*(1), 21–36.

He, W., & Wolfe, E. W. (2010). Item equivalence in English and Chinese translation of a cognitive development test for preschoolers. *International Journal of Testing*, *10*(1), 80–94.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer (Ed.), *Test validity* (pp. 129–145). Hillsdale, NJ: L. Erlbaum Associates.

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

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology*, *16*(2), 131–152.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349.

Kankaras, M., & Moors, G. (2009). Measurement equivalence in solidarity attitudes in Europe. *International Sociology*, *24*(4), 557–579.

Labercane, G., & McEachern, W. (1995). Striving for success: First Nations education in Canada. *Education*, *115*(3), 323–323.

Pashley, P. J. (1992). *Graphical IRT-based DIF analyses*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED385576

Philpott, D., Nesbit, W., Cahill, M., & Jeffery, G. (2004). Educational assessments of First Nations students: A review of the literature. In W. Nesbit (Ed.), *Cultural diversity and education: Interface issues* (pp. 77–102). St. John's, NL: Memorial University.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, *34*(4, Pt. 2), 100.

Schaeffer, N. C. (1988). An application of item response theory to the measurement of depression. *Sociological Methodology*, 271–307.

Scrams, D. J., & McLeod, L. D. (2000). An expected response function approach to graphical differential item functioning. *Journal of Educational Measurement*, *37*(3), 263–277.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement*, *22*(2), 77–105.

Statistics Canada. (2006). *2006 Census*. Retrieved from Statistics Canada website: http://www12.statcan.ca/census-recensement/2006/dp-pd/hlt/97-555/T401-eng.cfm

Statistics Canada. (2013). *Aboriginal peoples in Canada: First Nations people, Métis, and Inuit*. Ottawa, ON: Minister of Industry.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370.

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.

Van de Vijver, F. J. R. (2000). The nature of bias. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 87–106). Mahwah, NJ: Lawrence Erlbaum.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment*, *13*(1), 29–37.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *47*(4), 261–329.

Willms, J. D. (2009). Pre-schoolers benefit from new skills assessments. *Education Canada*, *49*(5), 36–39.

Willms, J. D., Tramonte, L., & Chin, N. (2007). *The use of item response theory for scaling behavioural outcomes in the NLSCY for the successful transitions project*. Unpublished report. Ottawa, ON: Human Resources and Skills Development Canada.

Wirth, W., & Kolb, S. (2004). Designs and methods of comparative political communication research. In F. Esser, & B. Pfetsch (Eds.), *Comparing political communication: Theories, cases, and challenges* (pp. 87–112). New York, NY: Cambridge University Press.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores*. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65–82). Charlotte, NC: Information Age Publishing.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. (Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science). Prince George, BC: University of Northern British Columbia.