Quality Criteria for Measurement in Social Research:
A Study of Item Bias in the Early Years Evaluation Direct Assessment (EYE-DA)¹

by
Luana Marotta

Abstract
Item bias occurs when survey questions do not have the same meaning and significance for all respondents. This is of particular concern when diverse social groups are involved. Item bias threatens measurement equivalence and causes intergroup comparisons to be invalid. Often encountered by educational researchers, item bias is also relevant to other types of quantitative research in the social sciences equally committed to improve measurement quality.

The analytical approach used to study item bias is called Differential Item Functioning (DIF). DIF occurs when respondents with similar values in the measured constructs (for example, socioeconomic status and cognitive ability) provide different patterns of responses to the same question. Using data from the Early Years Evaluation-Direct Assessment (EYE-DA), an assessment tool that measures children’s early educational development, an empirical study of the problem is offered. I examined whether EYE-DA test items—devised by Anglophone researchers—may be biased for Francophone and Aboriginal New Brunswickers. DIF analysis yielded consistent results, flagging few EYE-DA items as biased. Linguistic and cultural differences between Francophone and Anglophone children were more challenging to test transferability than cultural differences between Aboriginal and Anglophone examinees.

Introduction
“Measurement,” often used as a synonym for quantification (Lazarsfeld, 1969, p.94), refers to a process in which numbers are assigned to observations and reflect characteristics under study (Upshaw, 1968, p.60). Measurement enables social scientists to link abstract concepts to empirical indicants (Zeller & Carmines, 1980, p.2). It provides an “x-ray” of social structure in quantitative research (Lazarsfeld, 1962, p.760), serves to clarify theoretical frameworks, and suggests new variables for consideration (Blalock, 1970, p.88).

Sociologists have long been concerned about assessing measurement quality because of its importance to social inquiry.² Measurement errors lead researchers to incorrect inferences about relationships among variables, to incorrect theories about social realities—to what Hubert Blalock has called “the gap between theory and research” (1982, p.12).

Bias, one of the most frequent measurement errors in the social sciences (Hammersley & Gomm, 2000, p.151), occurs whenever complications in research processes distort outcomes in a

¹ This paper is an abbreviated version of my graduate thesis with the same title.
² Measurement became a popular tool early in the 20th century, at which time empirical research developed as an independent branch of Sociology in the United States (Lizon, 2006, p.227). The institutionalization of empirical research was followed by development of sample surveys that have enabled sociologists to access data highly relevant to their studies (Goldthorpe, 1997, p.408). The practice of quantitative analysis among sociologists gradually increased as survey research advanced.
systematic way (Kidder et al., 1986; Babbie 2007, 2008). It threatens measurement validity—that is, measurement’s capacity to reflect concepts researchers expect to measure. Bias can result from problems associated with data collection, instrument inadequacy, or errors in data analysis (Deming, 1944; Suchman, 1962; Fowler, 2009).

This study, which is concerned with questions about measurement errors, focuses on a specific form of bias called item bias. Survey or test items are biased when they systematically fail to measure what they are intended to measure, of particular concern when diverse social groups are involved. Offered herein is a systematic approach to assessing item bias in which the Early Years Evaluation - Direct Assessment (EYE-DA) is used as a model. The EYE-DA is an individually administered direct measure of the developmental status and the school preparedness of pre-kindergarten children 3 to 5 years old. The EYE-DA, employed by school districts across Canada, has been used in a number of research studies. Data collected for analysis, which comes from New Brunswick, Canada, encompasses over 13,000 children from Anglophone, Francophone, and Aboriginal communities.

Analysis is provided as to whether EYE-DA items, devised by Anglophone researchers, may be biased for Francophone and Aboriginal children—in other words, whether certain test items may be understood differently by the examinees and, thus, affect test outcomes. The analytical approach used to examine item bias is called Differential Item Functioning (DIF; Holland & Wainer, 1993). DIF is a statistical procedure used to flag item bias in Latent Analysis—analysis by which researchers estimate indices of non-observable constructs (e.g. power, degree of solidarity, and socioeconomic status) using observable variables. DIF investigates whether members of different social groups who hold like values with respect to the construct being measured provide different response patterns to the same question.

In the natural sciences—as opposed to the social sciences—measurement is more stable regardless of setting. A Celsius thermometer, for example, measures temperature the same way around the world. Measurement in the social sciences, however, tends to be subject to change as context varies: measures often refer to objects and life events, or situations wherein meaning and significance depends on understating certain phenomena within particular cultural contexts (Pawson, 1989, p.20). Social scientists, however, are interested in equivalent measures that allow them to make reliable comparisons of diverse social groups. The claim that social measures are not comparable across different actors or situations is “a devastating one” (Blalock, 1982, p. 57). Researchers must thus identify aspects of the research process that may threaten exchangeability of survey instruments, including item bias.

A review of pertinent literature follows, after which I describe methodological issues involved in the empirical investigation of item bias. Research findings are offered in the third section; the conclusions and implications of the study, in the fourth.

**Literature Review**

Item bias, which refers to distortions in item level, occurs when items function differently for certain groups of respondents. It violates the assumption of measurement invariance, which holds that measurement properties should not be affected by the context (Zumbo, 2009, p.76). Item

---

3 Aboriginal student respondents were enrolled in locally controlled, band-operated, First Nations schools (Labercane & McEachern, 1995; Carr-Stewart, 2006).
bias, often encountered by educational and psychological researchers, is equally relevant to sociologists interested in designing surveys that yield comparable results.

When researchers design surveys, they expect items to measure abstract constructs such as occupational status (Ganzeboom et al., 1992) or power (Banzhaf, 1965). Items are biased when they fail to measure the construct for every individual accurately. Measurement errors are systematic, rather than random, insofar as items tend to be misunderstood by specific groups of respondents, as is illustrated by Figure 1:

**Figure 1**

*Illustration of a Biased Item*

In Figure 1, Item A is entirely biased against Group 2. Item A does not measure what it should be measuring—the construct marked in blue—for all individuals of the group. The red lines indicate that Item A measures an inappropriate construct for the individuals in Group 2.

Item bias can result from either contextual or communication issues. The former refers to circumstances wherein attributes of social structure affect the significance of item content. Contextual differences influence the item’s capacity to equally reflect the underlying construct being measured among different populations. Communication problems, which most often stem from ambiguities of meaning, occur when respondent interpretations are affected by group affiliation.
When bias results from contextual issues, item content becomes inappropriate with respect to the construct being measured after the item is adapted to a different context, as shown in Figure 2. 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 3.

Figure 2
Item Bias Stemming from Contextual Issues
Bias stemming from contextual differences has been illustrated by Batista-Foguet et al. (2004), who have demonstrated that certain items included in the Family Affluence Scale (FAS) are not weighted equally among countries. The item “Number of holidays spent by the family,” for example, has a different impact on the construct “family wealth” when FAS is measured internationally. Vacation privileges and the number of public holidays differ and, thus, vary from country to country (2004, p.328). Traveling fewer times in any country is, therefore, not necessarily an indication of low family wealth. It may be a consequence of restricted leisure opportunities.

Item bias may also derive from communication problems, which can occur when researchers and respondents do not share the same understanding of survey items. Communication problems are likely to occur when diverse versions of survey instruments are involved. Poor translation of items may affect word or phrase complexity, for example. Ercikan (1998), who explored item bias while working for the International Association for the Evaluation of Educational Achievement (IEA), discovered differences of interpretation among native English- and French-
speaking Canadian students. According to Ercikan, eighteen of the IEA test items that he measured were biased—that is, they did not function equally for Anglophone and Francophone examinees. Vocabulary used in one item, he notes, may have had greater obscurity in English than in French:

“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 affecting item functioning can be anticipated, procedures can be adapted so as to identify areas of concern and minimize item bias. One way to detect item bias is to pre-test survey questions (Greenfield, 1996; Wirth & Kolb, 2004). Once data have been collected, bias can be explored by expert judges tasked with selecting potentially biased items, and/or by means of Differential Item Functioning (DIF), the statistical methodology used in this study.

DIF examines potential item bias in Latent Analysis, which aims to reduce data complexity by explaining associations between observed items in terms of fewer latent factors (Bartholomew et al., 2002, p.236). By examining potential item bias in latent analysis, DIF requires that multiple items measure the construct under study. From a statistical-methodological perspective, DIF occurs when respondents from different groups show differences of probability of item endorsement after having been matched according to the trait being measured (Zumbo, 1999, p.12). Individuals are expected to display similar response patterns with respect to items that measure the same construct. When the DIF technique is employed, item bias is signaled when response patterns systematically change among different groups with similar trait scores.

Item bias in latent analysis threatens “scalar equivalence” (Hui & Triandis, 1985; Van de Vijver & Tanzer, 2004). Scalar equivalence presumes that “a numerical value on the scale [or index] refers to the same degree, intensity, or magnitude of the construct regardless of the population of which the respondent is a member” (Hui & Triandis, 1985, p.135). Item scaling presupposes that item characteristics are similar for all respondents (Schaeffer, 1988, p.272). If indices contain a biased item, it may therefore be concluded that constructs under analysis have not been measured on the same metric.

Though often used interchangeably, “DIF” and “item bias” are not synonyms, as Clauser and Mazor (1998) rightly affirm. DIF is merely a statistical tool to investigate item bias—a technique, as Zumbo points out, that flags potentially biased items (1999, p.13). Statistical flags such as DIF should not be used to discard items automatically (Shepard et al., 1985, p.102). Before concluding that an item is biased, researchers must provide a more complete account of the nature of bias by employing other analytical approaches—evaluation by experts, for example.

Methods for investigating item bias are typically used when new measures are created, when measures are being adapted to new contexts or to different populations, or when existing measures are being translated (Zumbo, 2007a, p.223). As noted above, pre-tests may be also helpful in detecting group differences that affect measurement invariance.

DIF analysis herein determined whether EYE-DA items, developed by Anglophone researchers, are potentially biased when they are used without modification for students of other cultures. A detailed analysis follows.
Data Analysis

Potential item bias in the EYE-DA test was explored using data from 2008-09 and 2009-10, the first two cycles of New Brunswick’s provincial assessment. Of the 2008-09 sample (6950 children in total), 4872 were Anglophone; 1981 Francophone; 97 Aboriginal. Of children assessed in 2009-10 (6342 in total), 4229 were Anglophone; 2025 Francophone; 88 Aboriginal. Item bias was empirically analyzed by means of DIF analysis on the 2008-09 data; 2009-10 data were used to cross-validate the results. Analysis took into account the four EYE-DA domains: Awareness of Self and the Environment, Cognitive Skills, Language and Communication, and Motor Development (Fine and Gross Motor). Each domain contains twelve items, identical in the 2008-09 and 2009-10 EYE-DA cycles.

Nearly all DIF methods investigate the interaction between group membership and response patterns to items once respondents are matched with respect to the construct being measured (Angoff, 1993, p.13). The total score on all items is often used as the matching criterion. Scores for each assessed EYE-DA domain were estimated by means of Item Response Theory (IRT) graded-response models (Samejima, 1970).

Some DIF methods available for flagging item bias are the IRT (Thissen, Steinberg & Wainer, 1993), the Mantel-Haenzel (Holland & Thayer, 1988), the Standardization Procedure (Dorans & Kulick, 1986), and the Logistic Regression (Swaminathan & Rogers, 1990). All approaches use statistical tests that detect existence of DIF and measure the magnitude of its effect—what Scrams and McLeod have called “global” approaches to DIF (Scrams & McLeod, 2000). In this study, DIF analysis was carried out by means of the Logistic Regression Method, which consists of modeling the probability of item endorsement according to group membership and the total scores on constructs being measured by items. The method allows for the testing of uniform and non-uniform DIF (Clauser and Mazor, 1998). Uniform DIF occurs when DIF size—the difference in probability of correct answers between focal and reference groups of comparable ability—is constant along the trait continuum. Non-uniform DIF occurs when DIF size is greater for respondents with specific ability levels. Equation 1 shows the logistic model used in this study to detect DIF.

---

4 The relatively large EYE-DA sample comprises approximately 97% of pre-kindergarten children in the province.
5 Cross-validation did not apply when DIF analysis was carried out for Aboriginal examinees. A careful analysis of the 2008-09 and 2009-10 samples revealed that the group of Aboriginal children assessed in the first EYE-DA cycle may not be equivalent to the one evaluated in the second cycle. The EYE-DA cycles one and two were administered in eight Band-operated schools, out of which only two had a similar number of children doing the test in both cycles. The other six schools had a change in the number of students from cycle one to two of 104 per cent on average. In other words, the number of students in six Aboriginal schools increased or decreased approximately 104 per cent from 2008-09 to 2009-10—the Francophone and Anglophone schools had, on average, a loss or gain of 38 and 33 per cent students, respectively. Given such a great change in the number of students in Aboriginal schools and small sample sizes—97 and 88 Aboriginal children tested in 2008-09 and 2009-10, respectively—the groups of Aboriginals in the two EYE-DA cycles may be not comparable.
6 Unidimensionality, supported when one score is enough to synthesize correlated items (Embretson & Reise, 2000, p.37), was examined herein by means of Exploratory Factor Analysis (Stevens, 1996, p.386). Exploratory Factor Analysis has shown that Domains A, B, and C can be considered one-dimensional. A single score was estimated for each of these domains. Analysis revealed, however, that Domain D is conspicuously divided into two sub domains, Fine and Gross Motor Development. Scores were therefore estimated separately for each sub domain.
7 For instance, item bias stemming from imprecise translation may affect word complexity, causing items to display greater DIF with respect to students with low scores.
Equation 1

General Logistic Model for DIF Detection

\[
\ln \left( \frac{p_i}{1 - p_i} \right) = b_0 + b_1\text{Ability} + b_2\text{FocalGroup} + b_3\text{Ability} \times \text{FocalGroup}
\]

“\(p_i\)” is the proportion of examinees that completed Item “i” consistently and received full points.\(^8\) “Ability” refers to the children’s total score on the domain that was measured by the item “i”. “FocalGroup” is a dummy variable used to identify the examinee’s group membership. Referent-group Anglophones were coded “zero” (0); focal-group Francophone and Aboriginal children, “one” (1). Each focal group was analyzed separately. The product of the dummy variable “FocalGroup” and the predictor “Ability” are interaction terms that test for non-uniform DIF.

Statistical Testing of DIF

DIF significance was assessed by means of a Chi-square test, as suggested by Swaminathan and Rogers (1990). Chi-square statistics were calculated for three models:

- Model 1: \(Y = b_0 + b_1\text{Ability}\)
- Model 2: \(Y = b_0 + b_1\text{Ability} + b_2\text{FocalGroup}\)
- Model 3: \(Y = b_0 + b_1\text{Ability} + b_2\text{FocalGroup} + b_3\text{Ability} \times \text{FocalGroup}\)

Uniform DIF was deemed significant when the Chi-square difference between models 1 and 2 was bigger than the critical value of Chi-square with one degree of freedom at a significance level of 0.01. DIF was considered non-uniform when the Chi-square difference between models 2 and 3 was statistically significant.

The magnitude of DIF was examined by comparing the R-square of each of the three models using a strategy similar to that of the Chi-square test. DIF was considered to be relevant if the R-square difference between one model and another was at least of 0.035 (Jodoin & Gierl, 2001).

An item was declared potentially biased if DIF was considered significant by the Chi-square test, relevant by the R-square test, and consistent if the results had been 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 biased items are flagged (Hambleton & Jones, 1995, p.14).

\(^8\) Although EYE-DA items have four categories, the four were dichotomized and DIF detection accomplished by binary logistic regression in order to facilitate interpretation of analyses.
Graphical Representation of DIF

In addition to statistical tests to occurrence of DIF and its size effect, graphical DIF was also considered. Recently enhanced by Willms, Tramonte, and Chin (2007), graphical representation is 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. DIF size can be estimated by subtracting the probability of corrected answer of the focal group from that of the reference group. The relative size of DIF among diverse items can be compared and the overall DIF estimated by calculating average DIF size for items as a whole.

Graphical DIF approaches have been used elsewhere (Pashley, 1992; Douglas et al., 1996; Scrams & McLeod, 2000; Bolt & Gierl, 2006). Unlike global DIF approaches, which indicate only the existence of DIF and its size effect, graphical approaches provide complementary information to DIF analysis, as noted above. The approach used in this study implements a feature new to graphical DIF. In addition to direction, location, and relative size (typical of most graphical DIF approaches), the representation suggested by Willms, Tramonte, and Chin (2007) displays DIF impact. Score distribution is represented together with DIF size so as to determine the proportion of individuals potentially affected by item bias.

Graphical analysis combined statistical testing with complementary information provided by the aforementioned approach. Items were highlighted when DIF was shown significant by statistical tests used during logistic regression analysis. The results are discussed in the next section.

Findings

The DIF analysis produced remarkable results. When DIF was tested for Francophone children, seven out of forty eight items were consistently flagged as potentially biased in both EYE-DA cycles DIF. When DIF was tested for Aboriginal children, only one item displayed DIF. As has been said, cross-validation could not be done when DIF was tested for Aboriginal examinees because the 2008-09 and 2009-10 samples of Aboriginal children may not be equivalent.

DIF detection was primarily based on the R-square test, particularly when Francophone and Anglophone samples were analyzed. Because Chi-square statistics are highly influenced by sample size, trivial DIF effects became significant when large numbers were considered. R-square statistics, which depend on the response patterns of the independent variables, are not influenced by the sample size.9

DIF size is graphically represented below, by test domain, for all items of the first EYE-DA cycle. Score distribution is also reproduced in order to illustrate how many students were affected by DIF. The graphs show the overall DIF of each domain, which corresponds to the average DIF size of the items and depict DIF location (that is, indicate which groups of respondents were negatively or positively affected). Items were highlighted when DIF was shown to be significant by statistical tests.

9 Current software packages typically provide the Wald coefficient, designed to test the significance of individual coefficients for independent variables of logistic models. Results obtained from the Wald test are similar to those of the Chi-square test. It, too, is affected by sample size.
Figure 4
Size of DIF between 2008-09 Anglophone and Francophone Examinees

DIF size Domain A

DIF size Domain B

DIF size Domain C

DIF size Domain D Gross Motor

- 10 -
Figure 5
Size of DIF between 2008-09 Anglophone and Aboriginal Examinees
On the whole, DIF seems more relevant when comparing Francophone with Anglophone student examinees than Aboriginal with Anglophone. Indeed, results suggest that the test instrument may be equally applicable for Aboriginal and Anglophone children. Language and cultural differences between Francophone and Anglophone children appears to have a greater impact on exchangeability of EYE-DA items.

The overall DIF in each EYE-DA domain is relatively low along the trait continuum, as is shown by the bold line in the graphs. The negative impact of DIF on items that disfavored focal Aboriginal and Francophone groups was compensated for by the negative impact of DIF on items disfavoring the Anglophone reference group, resulting in an average DIF size close to zero.

DIF size of items in Domain D gross and fine motor appears smaller than DIF size of items in other domains. This phenomenon is readily explained. The items of Domain D are more straightforward. Item 1, for example, asks children to stand on one foot for a certain period of time; Item 9 asks them to draw a person; Item 10 asks them to print their first names. The items
in the other domains, more subjective and ambiguous in nature, invite communication challenges. In Item 8 of Domain C, for instance, test administrators must evaluate how well children pronounce “Dogs like to run and play” (“Les chiens aiment courir et jouer”); “Most rabbits have soft fur” (“La plupart des lapins ont le poil doux”); and “A turtle moves slowly” (“La tortue avance très lentement”). Item 9 of the same domain presents three pictures and asks respondents tell a story about the pictures. Test administrators are required to evaluate students based on story coherence and use of appropriate grammar.

The majority of children were not affected by DIF. DIF was generally higher in the left side or in the middle of the trait continuum, whereas score distribution tended to be concentrated in the right-hand side of the scale. DIF tended to affect, however, children with lower scores, having an impact on the measure quality of those in need.

Conclusion

This study dealt with the measurement error item bias, which results from challenges that affect exchangeability of survey items. Item bias is of concern to social scientists doing quantitative research because their studies usually involve diverse social groups for which survey items may not be equally applicable. Ideally, survey items cover the same construct for every respondent and produce comparable outcomes. Measurement invariance cannot be taken for granted in social research, however, for measurement tools are not context free (Zumbo, 2009, p.76).

Survey research if often criticized—notably by phenomenologists—for failure to accommodate variations of meaning across different cultures (Pawson, 1989, p.20). Survey studies, critics argue, result in standardized measures not necessarily equivalent for groups under study. Herbert Blumer (1956), a preeminent scholar of interpretivism, is one such critic. Blumer pointed out that most variables in the social sciences are tied to specific cultural contexts, and that objects and events of human experience (meaningless in and of themselves) have meaning conferred on them (p.687). Blumer evidently believed that different people assign different meanings and significance to the same measure. Aaron Cicourel also censured those who assume language and common-sense meaning to be self-evident (1964, p.17). Survey researchers, Cicourel argued, arbitrarily assume that all respondents interpret questions unambiguously.

Even if measurement invariance may be unachievable, measurement in the social sciences should not be deemed impossible. Social research is based on agreements concerning everyday terms and the concepts those terms represent, as Earl Babbie (2007) suggests. Because there are several ways to ensure that measurements fully integrate agreed-upon meanings (Babbie, 2007, p.146), researchers must endeavor to control errors and to increase measurement quality.

The number of respondents for whom an item is invalid is variable and therefore difficult to estimate. The exact amount of measurement error will, to a certain degree, remain unknown. With item bias analysis, however, errors are more easily detected because they are systematic and affect response patterns for specific groups of people. Researchers can often predict outcomes based on respondent characteristics. In the EYE-DA, for example, children who do well on the overall test are expected to have higher outcomes on individual items. Children may not have mastered every skill assessed by the EYE-DA test: they may have difficulty with certain items.

---

10 Interestingly, this item displayed DIF against Anglophone, not Francophone children.
When the majority of students in groups with high scores systematically fail to answer specific items correctly, however, item bias must be considered.

Measurement errors—seldom exclusively associated with group affiliation—may be randomly distributed over sample populations. Random errors, though more difficult to detect, are in fact less problematic for data analysis. Random errors do not affect comparisons of group outcomes as much as systematic errors do. The magnitude of random errors is expected to be the same for all groups (Camilli & Shepard, 1994, p.7); furthermore, measurement deviations in one group usually compensate for those in another group. Systematic errors, however, cause measures to be less valid for particular groups.

A crucial step of empirical analysis consists of selecting groups for which measures may not function as intended. It is not realistic to examine item bias for every possible group combination in a sample. Researchers must choose group differences most relevant to the research questions. Some criteria surface more often in the literature concerning item bias, especially gender and ethnicity. However, preconceived notions of which criteria have the potential to create item bias can blind researchers to other significant group differences. Careful pre-tests of survey items are vital for detecting unanticipated challenges that may affect item function.

This study was concerned with empirical analysis of item bias after collection of data. I have discussed a particular statistical procedure to detect item bias called “Differential Item Functioning” (DIF). DIF examines whether those of different social groups holding similar values with respect to the construct being measured provide different response patterns to the same question. DIF is a method for flagging item bias by detecting group differences in item performance caused by aspects irrelevant to survey purposes (contextual and communication issues discussed in the literature review). In a cross-national study of socioeconomic status, for example, respondents resident in developing countries are likely to score lower than those of developed countries. Group differences of performance with respect to measures of socioeconomic status are relevant for study purposes and refer to item impact (Clauser & Mazor, 1998)—that is, “true” differences among the groups of target constructs. Item bias occurs when group differences for individual items persist after respondents are matched on constructs of interest.

As explained, DIF methods may produce unstable results (Hambleton & Jones, 1995, p.14). I therefore adopted a conservative approach to data analysis. Items were flagged as potentially biased when both Chi- and R-square tests pointed to significant DIF, and when DIF was found in cross-validation samples. Empirical analysis revealed the Chi-square statistic to be highly influenced by sample size. When the relatively large Anglophone and Francophone samples were examined, most items displayed significant DIF. With such large samples, the R-square test yielded more reliable results.

DIF size between Anglophone and Francophone children was in general greater than that between Anglophone and Aboriginal. Although some educators have claimed that educational assessments deal with content unfamiliar to Aboriginal children (and that they are therefore unfairly disadvantaged), DIF results indicate that the EYE-DA test is equally suitable for Anglophone and Aboriginal children. If score disparities exist between Anglophone and Aboriginal examinees, they may be better explained by structural factors such as socioeconomic differences, not by problems with the EYE-DA instrument. Linguistic and cultural differences appear more problematic with respect to test transferability between Anglophone and
Francophone children. The number of items flagged as biased when DIF was tested for Francophones was small, however (7 out of 48). If bias is confirmed by expert judgment, flagged items can be excluded, changed, or substituted without substantially changing the test results.

Francophone and Aboriginal New Brunswickers are not necessarily representative of Francophone and Aboriginal communities elsewhere. Diverse sample are needed for a more complete account of item bias in Canada’s educational surveys. Aboriginal New Brunswickers may be culturally closer to the Anglophone community than those more geographically isolated. Cultural differences appear sharper among Aboriginal communities of Northern Canada. For example, in most Canadian provinces, less than one per cent of Aboriginals speak Aboriginal languages at home; in Nunavut, on the other hand, the average increases to 54.2 per cent. If remote Aboriginal communities are sampled, cultural differences may have greater impact on the transferability of educational instruments.

Different results may also be found if Canada’s other Francophone communities are surveyed. Given that cultural boundaries are often influenced by geographic location, Francophones from Quebec, where 67% of the population only speaks French, may be more culturally distant from Canadian Anglophones than Francophone New Brunswickers, where 10% of the population only speaks French and 33% is bilingual. Cultural differences between Francophone and Anglophones may be more challenging to test exchangeability if other Francophone communities are studied; accordingly, item bias may be more likely to occur if a broader sample is surveyed.

Using Celsius thermometers as an example, I stressed in my Introduction that measurement in the natural sciences is more stable than measurement in the social sciences, wherein meaning and significance often vary with context. I also stated that social scientists, like natural science researchers, are concerned about measurement equivalence. They, too, want to ensure that comparisons among people under study are valid. Nonetheless, because subjects of study in the social sciences—human beings and institutions—are somewhat more diverse and complex than in the natural sciences, researchers have more difficulty predicting and preventing ambiguities that may arise during the measurement process. Scalar equivalence comparable to thermometers is impossible to achieve in the social sciences.

Even though social science measurement is far from perfect, it is a useful tool for inquiry in that it permits researchers to compare diverse groups of individuals with relative ease, especially when large populations are being studied. Measurement errors cannot be completely avoided but, as I have demonstrated, their effect on validity can be controlled to a certain extent. Item bias, one kind of measurement error, must be carefully considered along with other challenges that jeopardize measurement quality.

---

13 Norbert Elias (1978), who provides a comprehensive discussion of the sciences and the complexity of their subject matter, argues that, physicists, biologists, and sociologists deal with processes of different levels of integration, which are more or less regular and predictable.
Item bias has been extensively discussed within the field of educational testing, but item bias is as relevant to quantitative research of all social science practitioners committed to enhancing measurement equivalence and meaningful group comparisons. DIF, the analytical approach used in this study to investigate item bias in an educational assessment tool, can be effectively applied to assessment of job satisfaction (Hulin & Mayer, 1986; Collins et al., 2000), domestic environments (Bingenheimer et al., 2005), personality (Panter et al., 1997), and socioeconomic status (Batista-Foguet et al., 2004).

Given increased globalization of commerce and cultural encounters, social scientists have noticed increased attention to the study of groups with conspicuous cultural and language differences. Empirical analysis of item bias is vital to ensuring that survey or test instruments employed are equally applicable among populations with salient groupings.

References


Bartholomew, D. J., Steele, F., Moustaki, I., & Galbraith, J. (2002). The analysis and interpretation of multivariate data for social scientists (2nd ed.) Chapman and Hall/CRC.


Lizón, Á. (2006). Estadística y causalidad en la sociología empírica del XX. *Papers, 80*.


