EARLY EDUCATION AND DEVELOPMENT, *18*(3), 453–472


# Does the EDI Measure School Readiness in the Same Way Across Different Groups of Children?

Martin Guhn
*Human Learning, Development, and Culture*
*University of British Columbia*

Anne Gadermann and Bruno D. Zumbo
*Measurement, Evaluation, and Research Methodology*
*University of British Columbia*

The present study investigates whether the Early Development Instrument (Offord & Janus, 1999) measures school readiness similarly across different groups of children. We employ ordinal logistic regression to investigate differential item functioning, a method of examining measurement bias. For 40,000 children, our analysis compares groups according to gender, English-as-a-second-language (ESL) status, and Aboriginal status. Our results indicate no systematic measurement differences regarding Aboriginal status and gender, except for 1 item on which boys are more likely than girls to be rated as physically aggressive by Kindergarten teachers. In contrast, ESL children systematically receive lower ratings on items of the language and communication domains—as expected by definition of ESL status—but not within the physical, social, and emotional domains. We discuss how our results fit with child development research and the purpose of the Early Development Instrument, thus supporting its validity.

## INTRODUCTION

The Early Development Instrument (EDI; Offord & Janus, 1999) is

Correspondence should be addressed to Martin Guhn, University of British Columbia, 400-2725 Melfa Road, Vancouver, BC V6T 1N4, Canada. E-mail: guhn@interchange.ubc.ca

> a teacher-completed measure of children's school readiness at entry to grade 1 [that] was designed to provide communities[1] with a feasible, acceptable, and psychometrically reliable instrument that [can] be used for whole populations of children to monitor community efforts to improve early years' outcomes over time. (Janus & Offord, 2007, p. 12)

In other words, the EDI is a community tool to assess "early years' outcomes" or "school readiness" at an aggregated community or population level.

In the fields of public health and epidemiology, this concept of measuring and reporting certain characteristics (e.g., health outcomes, incidence of illnesses) of people at a population level is very common. In education, this approach is less common but has also been used (e.g., when reporting achievement scores, dropout rates). However, in regard to characteristics of children, and particularly in regard to a holistic view of school readiness as assessed by the EDI, this approach has not, to our knowledge, been used before. Traditionally, children's school readiness has been assessed at an individual level, for the purpose of assigning individualized prevention and intervention strategies to children with perceived needs.

Thus, the construct of school readiness as defined by the EDI is to be understood quite differently than the traditional notion of school readiness; that is, school readiness is seen as a characteristic of interest at an aggregated group level (e.g., community or school), not at an individual level. Accordingly, it is advised that interpretations of EDI data be conducted solely at such group (i.e., community, school, etc.) levels, and the EDI is explicitly not a tool to diagnose (and to thus label) individual children. Despite the fact that EDI data are aggregated and then interpreted at the aggregated level, the data are nonetheless obtained at an individual level, namely via teaching ratings. Thus, the fact that EDI data are interpreted at the aggregate level does not mitigate the need to examine psychometric properties of the EDI in regard to the individual-level data, because a systematic bias at the individual level could result in systematic bias at the aggregate level.

In this article we investigate item bias, a pivotal aspect of test validity, for the EDI. Investigating item bias is important, because item bias presents a threat to the validity and fairness of a test or scale (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Zumbo, 2007). Bias at the item level, if large enough, may translate to bias at the domain or scale score level. This is of particular relevance for tests that are (primarily) interpreted at the domain or scale score level, as is the case for the EDI. Accordingly, we also investigate whether cumulative effects of (potential) bias at the item level lead to bias at the domain score level.

With regard to validity, these issues are of particular importance, because inferences that are made on the basis of domain or scale scores that are biased are not

---

[1]Here, we use the term *community* to refer solely to the concept that is delineated in the *British Columbia Atlas of Child Development* (Kershaw, Irwin, Trafford, & Hertzman, 2005), which uses it synonymously with *neighborhood*.

equally appropriate, meaningful, and useful for different subgroups of the target population. This highlights the fact that the investigation of test and item bias is targeted at establishing the inferential limits of a test, that is, for which group of respondents the inferences made on the basis of the test scores are valid and for which they are not (Zumbo, 2007, in press). In broad terms, this is a matter of measurement invariance; that is, is the EDI performing in the same manner for each group of examinees (e.g., boys and girls)?

It is important at this point to highlight how the EDI is administered. Kindergarten teachers are asked to rate each of their students on 103 items, which are then separated into five developmental domains: (a) Physical Health and Well-Being, (b) Social Competence, (c) Emotional Maturity, (d) Language and Cognitive Development, and (e) Communication Skills and General Knowledge (cf. Janus & Duku, this issue). Inferences from the EDI are then based, as noted above, at a group level (e.g., a community) on the five domains.

Because the EDI items involve binary and rating (Likert) response formats, we employed ordinal logistic regression in order to examine item bias (Zumbo, 1999). Ordinal logistic regression is a method to statistically identify the presence of differential item functioning (DIF). The presence of DIF at the item level is a necessary, but not sufficient, condition for item bias (Camilli & Shepard, 1994). That is, if an item is flagged as displaying DIF, it does not necessarily mean that item bias is present. Rather, one has to ascertain whether the statistical presence of DIF is due to item bias or item impact. The following definitions (cf. Zumbo, 1999) illustrate the distinction between these different terms:

- *DIF*. DIF occurs when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item after matching on the underlying ability that the item is intended to measure. The existence of DIF—a statistical property—indicates the presence of either item impact or item bias, and the distinction between the two cannot be inferred by statistical analysis alone.
- *Item impact*. The presence of DIF indicates item impact when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item due to *true* differences between the groups in the underlying ability being measured by the item.
- *Item bias*. The presence of DIF indicates item bias when examinees from different groups have differing probabilities of responding correctly to (or endorsing) an item due to differences between the groups in regard to some characteristic of the test item or testing situation that is *not relevant* to the test purpose.

These definitions illustrate that DIF is, as mentioned previously, a necessary but not sufficient (statistical) condition for item bias. Therefore, in the case of the statistical presence of DIF, subject matter experts should be consulted to differentiate theoretically and conceptually between item bias (implying that the item is mea-

suring construct-irrelevant differences) and item impact (implying that the item is measuring construct-related differences; Camilli & Shepard, 1994).

It thus needs to be emphasized that a statistical examination of items can only indicate the presence of DIF; the statistical analysis itself cannot make a distinction between item bias and item impact.

Procedures to identify DIF, and thus potential item bias, are frequently used in the process of developing and adapting educational and psychological measures, as well as for the validation of test score inferences. In particular, the analysis of DIF is performed to examine five issues that are foundational for establishing test validity (Zumbo, in press): (a) fairness and equity in testing for test participants from different groups; (b) ruling out measurement artifact as potential threat to internal validity; (c) identifying group differences in item responding that— pending further investigation—arise from group differences that are either criterion-relevant or -irrelevant, such as differences in ability, differences in cognitive processing, and/or differences in contextual or psychosocial factors; (d) translation and/or adaptation of measures to different languages or cultures; and (e) as part of item response theory and other such latent variable modeling. In this article, we primarily examine the first two issues, with a passing nod to the third. In this context, it is important to recall that the EDI is filled out by the Kindergarten teacher, and not by the children themselves. Accordingly, any DIF on the EDI is to be understood as a difference between groups *with respect to the perception and rating of the Kindergarten teacher.* Therefore, for the EDI, the issues of (a) fairness and equity in testing, (b) ruling out measurement artifact as potential threat to internal validity, and (c) identifying group differences in item responding are all to be interpreted in light of the fact that the ratings reflect the perceptions of the Kindergarten teacher.

## Research Objective

The EDI is, as was mentioned previously, a *community tool* to measure school readiness of groups of children. The implications for research investigating the validity of the EDI is that one needs to examine the decisions and inferences that are made based on EDI data at a group (e.g., community or population) level. According to this purpose, the reporting of EDI results has occurred by grouping children at the community level, as well as at the school district or health district level (Kershaw, Irwin, Trafford, & Hertzman, 2005).

The EDI has been used across diverse communities and school or health districts within Canada (and also in Europe, Australia, and South America; Janus et al., 2007). Given this diversity of communities, it is pivotal to examine DIF in order to allow for meaningful comparisons across these communities and districts. After all, it is important to find out whether the EDI is measuring school readiness similarly across different groups of children and, likewise, across communities with diverse compositions of groups of children (Zumbo & Gelin, 2005). Our analysis addresses this issue within the context of the Canadian province of British Co-

lumbia (BC). The province of British Columbia (equal to the size of France, Germany, and the Netherlands combined, with a population of about 4.5 million) is made up of about 500 communities.

The definition and boundaries of these communities are based on research with, and reports of, the people living in these communities. These communities differ largely with respect to their demographic, cultural, geographic, and socioeconomic characteristics (Kershaw et al., 2005). Accordingly, the question for us was to decide which criteria for the grouping of Kindergarten children have significance for the BC context and should therefore be used for our DIF analysis.

We decided to focus on three criteria, namely (a) student gender, (b) student English-as-a-second-language (ESL) status (i.e., ESL vs. non-ESL/native speaker), and (c) student Aboriginal status (i.e., Aboriginal vs. non-Aboriginal background).

These groupings have commonly been used in developmental research. Gender differences in regard to school readiness are of general interest to developmental researchers as well as educators and parents (e.g., Angenent & de Man, 1989; Dauber, Alexander, & Entwisle, 1993; Duncan et al., 2006; Gullo & Burton, 1992; McCoy & Reynolds, 1999). Examining DIF—and thus the presence of item bias or item impact—with regard to gender contributes important information as far as the interpretation of gender differences is concerned.

Likewise, differences in school readiness with regard to ESL status are also of importance to educators and others, particularly in regard to language, reading, and writing acquisition (e.g., Chiappe & Siegel, 1999; Lesaux & Siegel, 2003). In this area, an examination of DIF with respect to ESL status can contribute important interpretative information toward, for example, policies regarding language instruction and educational support for ESL children.

Finally, examining DIF with respect to Aboriginal status is of particular sociopolitical relevance because it is associated with issues regarding the cultural relationship between Aboriginal and European immigrant education (see Miller, 1996, for a historical account of schooling and education from an Aboriginal perspective).

## METHOD

### Sample

Our sample consisted of 43,900 Kindergarten children from the entire province of BC, Canada. Data collection occurred during the spring terms of five consecutive school years, 1999/2000 through 2003/2004. Of the children, 48.6% were female, 51.4% were male. According to the information provided by the Kindergarten teachers on the EDI, 17.0% of the children were non-native speakers (i.e., ESL), and 6.7% were Aboriginal. A comparison of our EDI data set with a data set from the British Columbia Ministry of Education in regard to the designations ESL/non-ESL and Aboriginal/non-Aboriginal showed that both the ESL and Aboriginal children were slightly underrepresented; in other words, on the EDI, Kinder-

garten teachers did not assign ESL status or Aboriginal status to as many children as were indicated as such by the Ministry data.

Therefore, it was examined whether this underrepresentation occurred in a systematic way. However, a comparison between the groups (Group 1: children for which the ESL or Aboriginal designation in the EDI database coincided with the Ministry's designation; Group 2: children that were designated as ESL or Aboriginal only by the Ministry data[2]) showed that there were no statistically significant differences in regard to the groups' respective EDI scores. Given this finding, we could assume for our further analyses that our EDI results were not systematically influenced (i.e., biased) by an under- or misrepresentation of ESL Kindergarten children due to differences in teachers' and the Ministry's classification criteria for ESL status.

Participation in the EDI survey was voluntary, even though it was facilitated and supported by the Ministry of Education. Overall, participation was extremely high, with representation from all 59 school districts in BC. Of the schools that opted out, a relatively high proportion of schools were among those that are located on Aboriginal reserves (for an illustration of potential reasons, the interested reader is referred to Miller, 1996).

## Measure

All children were rated on the EDI by their Kindergarten teachers. The EDI contains demographic information (e.g., gender, age, first language, Aboriginal background) and 103 binary and Likert-scale items on five developmental domains: Physical Health and Well-Being (13 items), Social Competence (26 items), Emotional Maturity (30 items), Language and Cognitive Development (26 items), and Communication Skills and General Knowledge (8 items).

The following is a sample question[3] from the EDI for the Communication Skills and General Knowledge Domain: "How would you rate this child's ability to tell a story?" Response options are *very good/good*, *average*, *poor/very poor*, and *I don't know*. For data analysis purposes, all responses on binary items were coded 0 or 10; 3-point Likert-scale items were coded 0, 5, and 10; and 5-point Likert-scale items were coded 0, 2.5, 5, 7.5, and 10. All items contain an additional response option, *I don't know* (coded 99), which was not included in the statistical analyses. For every item, 10 designates the highest (i.e., most positive, most developmentally desirable) score.

---

[2]In BC, both the EDI and the Ministry of Education data included individual child information. Thus, children that were identified as ESL or Aboriginal in the EDI database but not in the Ministry database could be individually identified.

[3]The EDI is available (in English and French) at www.offordcentre.com/readiness/EDL_viewonly.html

For every domain, the average score was calculated, ranging from 0 to 10. In addition, the five domain scores were combined into a total EDI score ranging from 0 to 50. It needs to be noted that only the domain scores are reported in practice (Janus et al., 2007; Kershaw et al., 2005[4]), in concert with the recommendations of the authors of the EDI; here, we use the total score purely for methodological research purposes.

## DIF Analyses

In this section, we provide a brief nontechnical introduction to DIF analysis using ordinal logistic regression. For a comprehensive, in-depth coverage of the method, the interested reader is referred to Zumbo (1999, 2007), and Shimizu and Zumbo (2005).

There are several ways to examine DIF, and thus measurement and test bias, statistically. For tests that consist of items with binary and ordinal (e.g., Likert-scale) response formats, Zumbo (1999) developed a method that integrates binary and ordinal logistic regression. In this method, as the first step, groups of participants (e.g., boys and girls) are matched on the variable of interest (e.g., the total EDI score as an indicator of overall school readiness). Then, the probability of obtaining a certain score on the item under investigation is calculated for both groups, for each total EDI score level, respectively. Accordingly, the logistic regression model includes variables to represent (a) the groups, (b) the score for the variable of interest, and (c) the interaction between the group status and the score for the variable of interest (Shimizu & Zumbo, 2005). For the analysis of DIF, the predictor variables are not entered simultaneously but successively for the following three models (Gelin, Carleton, Smith, & Zumbo, 2004; Zumbo, 1999):

Model 1: The conditioning variable (i.e., the total EDI score) is the sole predictor.
Model 2: The conditioning variable (i.e., the total EDI score) *and the grouping variable* are in the equation.
Model 3: The conditioning variable (i.e., the total EDI score), the grouping variable, *and the interaction term representing the interaction of the total EDI score and the grouping variable* are in the equation.

These three models correspond to the following three equations, in which $y \times_{\text{item score}}$ represents the predicted item score; $b_0$ and $b_1$, respectively, stand for the regression intercept and regression coefficient; $\text{TOTAL}_{\text{EDI score}}$ denotes the conditioning variable, the total score of the EDI; GROUP refers to the grouping variables gender, ESL status, or Aboriginal status; $\text{TOTAL}_{\text{EDI score}} \times \text{GROUP}$ represents the interaction term between the total EDI score and either gender, ESL status, or Aboriginal status; and *e* designates the error term.

[4]*The British Columbia Atlas of Child Development* is available at www.earlylearning.ubc.ca

Model 1: $y \times_{\text{item score}} = b_0 + b_1\text{TOTAL}_{\text{EDI score}} + e$

Model 2: $y \times_{\text{item score}} = b_0 + b_1\text{TOTAL}_{\text{EDI score}} + b_2\text{GROUP} + e$

Model 3: $y \times_{\text{item score}} = b_0 + b_1\text{TOTAL}_{\text{EDI score}} + b_2\text{GROUP} + b_3(\text{TOTAL}_{\text{EDI score}} \times \text{GROUP}) + e$

This sequence allows one to calculate how much variance the grouping variable (in Model 2) explains over and above the conditioning (i.e., matching) variable (in Model 1). The difference between Model 1 and Model 2 can then be tested for significance via a chi-square test, and an effect size can be calculated via a comparison of the $R^2$ values. Similarly, a comparison of Model 2 and 3 allows one to calculate how much variance the interaction term (in Model 3) explains over and above the effects of the conditioning and grouping variables (in Model 2; Zumbo, 1999). In other words, this analysis allows us to address the following questions: (a) Is there a significant group difference? If so, what is its effect size?; and (b) Is there a significant interaction? If so, what is its effects size?

Accordingly, in DIF terminology, *uniform* DIF refers to the group differences (i.e., the main effect, comparing Models 1 and 2), and *nonuniform* DIF refers to the Group × Total Score interaction (i.e., the interaction effect, comparing Models 2 and 3).

For the interpretation of effect sizes from ordinal logistic regression DIF analyses, Jodoin and Gierl (2001) have suggested guidelines. According to their criteria, effect sizes of $R^2 < .035$ are considered negligible, those between .035 and .070 moderate, and those =.070 are large.

## Statistical Analysis

We used Zumbo's ordinal logistic regression DIF methodology (1999). The existence of DIF was examined for each of the 103 items of the EDI for each of the following group comparisons, respectively: (a) Gender (girls vs. boys), (b) ESL designation (ESL vs. non-ESL), and (c) Aboriginal background (Aboriginal vs. non-Aboriginal).

Models 1 through 3, as described previously, were fit for each of the 103 items separately. For every analysis, Kindergarten children were matched based on their total EDI score. Although the total EDI score is not reported in practice (see the "Measure" section), we conditioned (i.e., matched) on the total score for the following two reasons: (a) A factor analysis of the items indicated that there was one dominant factor (suggesting that the total score was a proxy for a child's overall school readiness), and (b) when matching on domain scores (with different scale lengths), each item had a different and, potentially, relatively large contribution to the matching criterion (e.g., for the Communication Skills and General Knowledge domain, consisting of 8 items, each item contributed an eighth to the matching score).

In the second step, for those items flagged with DIF—using the criteria for statistical significance as well as the effect size criteria proposed by Jodoin and Gierl (2001) described in the "DIF Analyses" section—we investigated whether DIF items had an effect at the domain level. In other words, we examined whether DIF of a single or multiple items on one domain resulted in differential functioning at the domain level. This is of particular relevance, because EDI scores are reported solely at the domain scale level, and hence one would want to see the cumulative effect of the item-level properties on the domain score. The examination was done graphically, as statistical significance tests and effect size estimations for the analysis of differential domain-level functioning have yet to be developed.

In the third step, after the matching on the total EDI score, the probabilities for obtaining a certain domain score for each of the groups being compared and for the entire range of the matching score were calculated by adding up the predicted item scores of the domain score under investigation. These domain score probabilities could then be presented graphically in a curve that is the domain-level equivalent of an item response function from item response theory. The total was then divided by the number of items on the domain scale for ease of interpretation (the reported EDI domain scores are also average scores, likewise ranging from 0 to 10). The predicted average domain scores were then plotted for the respective group comparisons to visually represent the differential domain functioning. In essence, in the language of psychometrics, we translated the item characteristic curves to domain-level characteristic curves, which were then plotted and compared on the same graph.

The last step in our analysis was to conceptually examine whether those items flagged with DIF were indicative of item bias or of item impact. Accordingly, subject matter experts were consulted to scrutinize whether our findings coincided with the research literature (suggesting item impact) or whether our findings were more likely to be consequences of the measurement process (suggesting item bias).

## RESULTS

### DIF Grouping Variables

In addition to the theoretical, educational, and sociocultural reasons for selecting the grouping criteria discussed previously (gender, ESL status, and Aboriginal status), it is noteworthy that the EDI results for each of these three comparisons showed statistically significant differences of substantial effect size. In Table 1, these differences are presented for each of the three group comparisons and for each of the five EDI domains individually. The differences are expressed in the raw score metric—the actual differences between the groups' respective average do-

TABLE 1
Mean Group Differences and Effect Sizes for Group Comparisons

| *EDI Domain* | *Gender Comparison* | | *ESL Comparison* | | *Aboriginal Comparison* | |
|---|---|---|---|---|---|---|
| Physical Health and Well-Being | 0.32 | (*d* = .30)[a] | 0.14 | (*d* = .13) | 0.58 | (*d* = .50) |
| Social Competence | 0.75 | (*d* = .43) | 0.46 | (*d* = .25) | 0.66 | (*d* = .36) |
| Emotional Maturity | 0.77 | (*d* = .52) | 0.32 | (*d* = .21) | 0.51 | (*d* = .31) |
| Language and Cognitive Development | 0.56 | (*d* = .27) | 0.80 | (*d* = .40) | 0.86 | (*d* = .41) |
| Communication Skills and General Knowledge | 0.56 | (*d* = .29) | 2.15 | (*d* = 1.10) | 0.68 | (*d* = .33) |
| Total EDI score[b] | 2.96 | (*d* = .43) | 3.86 | (*d* = .56) | 3.27 | (*d* = .45) |

*Note.* Mean group differences were on a 10-point scale. Means were higher for girls, native speakers, and non-Aboriginal children. EDI = Early Development Instrument; ESL = English as a second language.

[a]Effect sizes (Cohen's *d*) of .2, .5, and .8 are considered as small, medium, and large, respectively (Cohen, 1992). [b]Sum of domain scores; 50-point scale.

main scores, on a 10-point scale—and as standardized effect sizes (Cohen's *d*). In the table note, we include Cohen's (1992) general guidelines for interpreting effect sizes.

## DIF Analyses

*ESL.* For the ESL versus non-ESL comparison, seven items displayed (uniform) DIF. (No items displayed nonuniform DIF.) In all, 5 of these items with DIF belonged to the Communication Skills and General Knowledge domain (with a total of 8 items), the other 2 to the Language and Cognitive Development domain (with a total of 26 items; see Table 2). In Table 2, the three items of the Communication Skills and General Knowledge domain that displayed DIF of negligible effect size are also included (in italics).

In Figure 1, one item displaying DIF is represented graphically as an illustrative example. This item is the first item on the Communication Skills and General Knowledge domain and the one with the largest DIF effect size ($\Delta R^2 = .091$, $p < .001$). An examination of the group differences at the domain score level (for the Communication Skills and General Knowledge domain) showed that ESL children, on average, received lower scores (2-point difference on a 10-point scale) than their non-ESL counterparts when matched at the same level of school readiness (i.e., on the total EDI score). Corresponding to the term *differential item functioning* we refer to this difference as *differential domain functioning*. Figure 2 graphically represents these group differences at the domain level (i.e., the differential domain functioning). For the Language and Cognitive Development domain, the two items with DIF did not have an effect at the domain score level.

TABLE 2
Effect Sizes for EDI Items With Uniform Differential Item Functioning in the ESL Versus Non-ESL Comparison

| *Item* | *Effect Size*[a] *($\Delta R^2$)* |
|---|---|
| Communication Skills and General Knowledge domain (8 items) | |
| How would you rate this child's …[b] | |
| … ability to use language effectively in English? | .091 |
| … ability to listen in English? | .041 |
| … ability to tell a story? | .067 |
| *… ability to take part in imaginative play?* | *.020* |
| … ability to communicate own needs in a way understandable to adults and peers? | .035 |
| … ability to understand on first try what is being said to him/her? | .035 |
| *… ability to articulate clearly, without sound substitutions?* | *.028* |
| Would you say that this child …[c] | |
| *… answers questions showing knowledge about the world?* | *.033* |
| Language and Cognitive Development domain (26 items) | |
| Would you say that this child …[c] | |
| … is showing awareness of rhyming words? | .048 |
| … understands simple time concepts (e.g., today, summer, bedtime)? | .036 |

*Note.* EDI = Early Development Instrument; ESL = English as a second language.
[a]Effect sizes of $R^2 < .035$ are considered negligible, those between .035 and .070 moderate, and ones =.070 large (Jodoin & Gierl, 2001). Negligible effect sizes are in italics. [b]Response options for these items are on a 3-point Likert scale: *very good/good* (10), *average* (5), *poor/very poor* (0), and *don't know.* [c]Response options for these items are binary: *yes* (10), *no* (0), and *don't know.*

*Gender.* In the analysis for gender, one item showed (uniform) DIF, namely an item from the Emotional Maturity domain: "Would you say this child gets into physical fights?" Response format was a 5-point Likert scale: *often* or *very true* (0), *sometimes* or *somewhat true* (5), never or *not true* (10), and *don't know.* (No item displayed nonuniform DIF.)

The effect size for the DIF of this item was $\Delta R^2 = .053$ ($p < .0001$). In Figure 3, the item score probabilities for boys and girls depending on their total EDI score are plotted. The graph illustrates that boys, on average, and at every total EDI score level (the matching criterion), had a higher probability of obtaining a teacher rating designating them as more physically aggressive than girls. (Figure 3 is thus equivalent to Figure 1 in that both depict DIF at the item level.)

The graph in Figure 4 presents the probabilities for obtaining a certain score on the Emotional Maturity domain for boys and girls at every level of the matching score (i.e., total EDI score). As can be seen, the two plots (Figure 4; for Girls and Boys) are nearly identical, showing that the DIF on that one item had no effect at the domain score level. In other words, our graphic examination suggests that there was no substantial differential domain functioning, despite the fact that one item

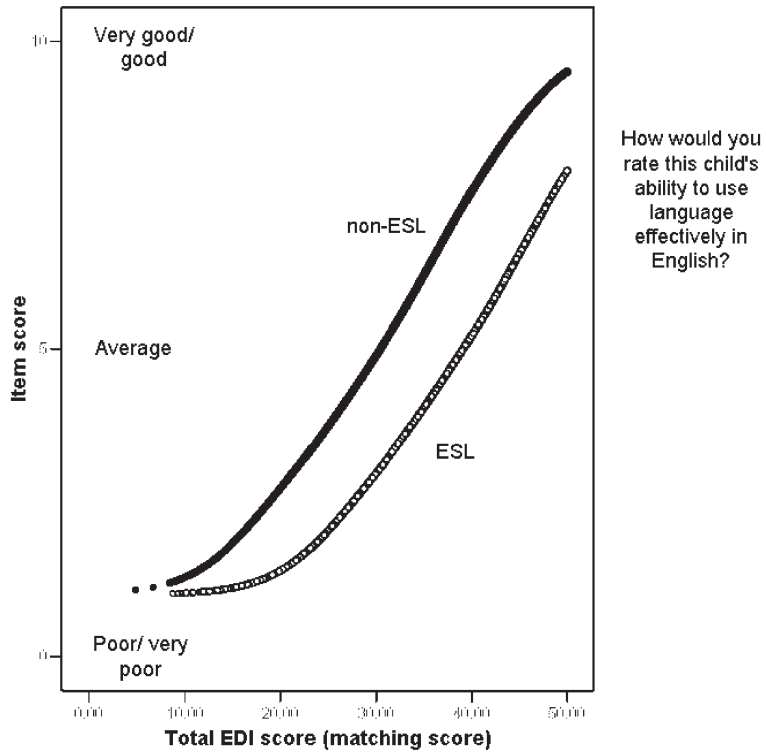


FIGURE 1 Item "Ability to use English effectively" on the Communication Skills and General Knowledge domain, which displayed differential item functioning in the ESL comparison. ESL = English as a second language; EDI = Early Development Instrument.

on this scale displayed DIF. (Figure 4 is equivalent to Figure 2 in that both depict the effect—or lack of effect—of the item DIF at the domain level.)

*Aboriginal background.* In the analysis comparing children designated as Aboriginal with those designated as non-Aboriginal, no item showed DIF.

## DISCUSSION

Our DIF analysis identified several items that displayed uniform DIF of substantial effect size. In one case, DIF at the item level resulted in DIF at the domain level—or, to be more exact, in differential domain functioning. In regard to fairness and measurement bias, our results support the validity of the EDI and suggest that the EDI is measuring school readiness similarly across groups of Kindergarten chil-

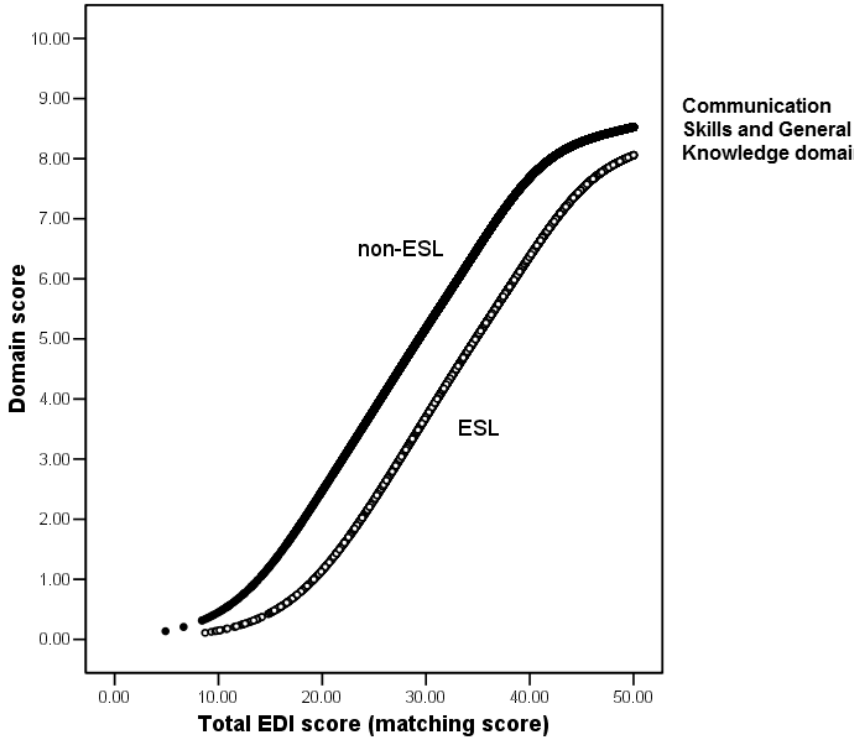


FIGURE 2 Differential domain functioning (Communication Skills and General Knowledge domain) due to (cumulative) item-level differential item functioning. ESL = English as a second language; EDI = Early Development Instrument.

dren, grouped according to gender, ESL status, or Aboriginal background. In other words, the Kindergarten teachers' ratings of the children on the EDI are not biased by their perceptions of children's gender, ESL, or Aboriginal status. We make this general conclusion based on arguments that suggest that all cases of DIF on the EDI are cases of item impact, meaning that group differences on these items reflect actual group differences in the underlying ability or skill that is being measured rather than construct-irrelevant variance. In the following sections we discuss the results and our pertaining arguments in detail and also address a number of questions raised by the results.

## ESL Status Group Comparison

Most of the items that displayed DIF were identified in the comparison between ESL and non-ESL children. In this comparison, seven items displayed DIF. Five of these items were on the Communication Skills and General Knowledge domain,

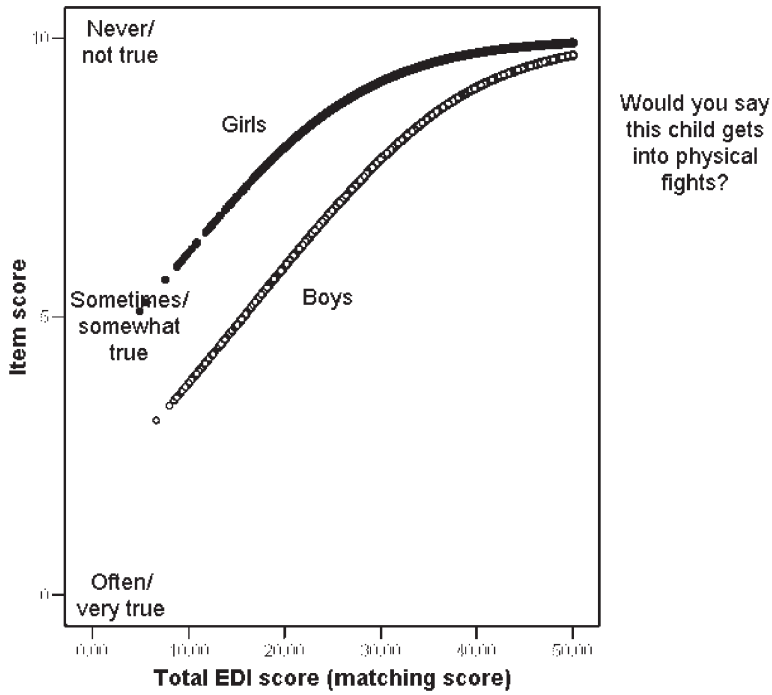


FIGURE 3 Item "Gets into physical fights" on the Emotional Maturity domain, which displayed differential item functioning in the gender comparison. EDI = Early Development Instrument.

and the other two were on the Language and Cognitive Development domain. On the Language and Cognitive Development domain, the presence of 2 items (out of 26) with DIF did not have an effect on the average score for this domain. However, for the Communication Skills and General Knowledge domain, the (cumulative) presence of DIF on five out of eight items did add up to the point that it clearly affected the domain-level score.[5]

The size of this effect at the domain score level was quite substantial, as the following points illustrate: EDI results are, as mentioned, reported at the domain score level for each community or district. For the five EDI domains, the ranges of the average scores for the 59 school districts, on a 10-point scale, were as follows: 8.08 to 9.08 (Physical Health and Well-Being), 7.38 to 8.96 (Social Competence), 7.35 to 8.71 (Emotional Maturity), 7.44 to 9.02 (Language and Cognitive Develop-

[5]Due to the absence of a statistical test, we cannot refer to this difference as *statistically* significant, even though the size implies *practical* significance.

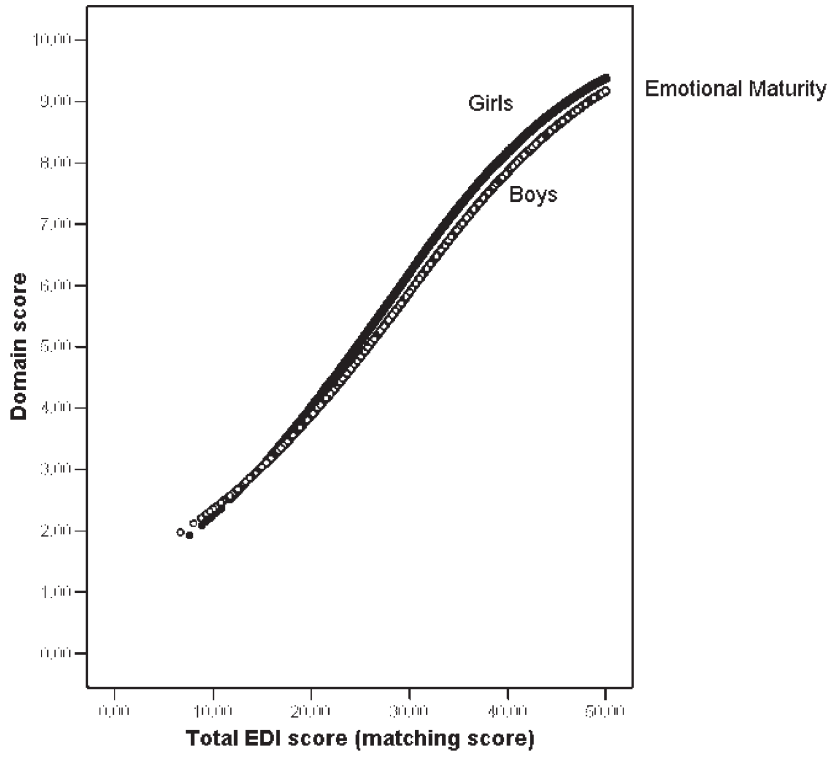


FIGURE 4 No differential domain functioning (on the Emotional Maturity domain), despite item-level differential item functioning. EDI = Early Development Instrument.

ment), and 6.29 to 8.36 (Communication Skills and General Knowledge). As our results indicate, the group difference between ESL and non-ESL children on this scale is about 2 points. Numerous communities in BC have more than 50% of children with ESL status, and for such communities, the average domain score for Communication Skills and General Knowledge is going to be, on average, about 1 point below that for communities without ESL children.

What does this mean in practical terms? For a district to drop by 1 point on the average score of the Communication Skills and General Knowledge domain is equivalent to dropping from the top quintile to the lowest quintile,[6] and the same is true at the community level. Commonly, the relative ranking of districts or communities derived from the EDI average scores, as well as an associated percentage of

[6]In the *British Columbia Atlas of Child Development* (Kershaw et al., 2005), quintiles are used for the reporting of results at the community and district levels.

vulnerable children within a district or community, has been used as an argument to either back up funding requests (in the case of perceived need; i.e., relatively low average scores) or to praise community initiatives or political action (in the case of relatively high scores).

Considering such usage of EDI scores, the question is what the implications of our finding are. Does the same Communication Skills and General Knowledge score have the same practical implications for a community with a high proportion of ESL children as opposed to one with a low proportion? Do separate group norms for ESL and non-ESL children provide an answer? We advise against doing so, because in our opinion this could invite false inferences. It might, for example, convey the misleading perception that a level of Communication Skills and General Knowledge that is considered insufficient for native English-speaking children is perceived as "normal" for ESL children—with the implication that normal might translate into "ok" or "acceptable." From a societal and educational point of view, however, the goal ought to be that (almost) all children reach a sufficient level of school readiness, and thus communication skills, so that they can thrive in school. The challenge therefore is how communities and schools can jointly provide support for families and their children with relatively low English communication skills. By definition, a large proportion of these children comes from an ESL background, because the ESL designation is not assigned to children who are non-native speakers, but only to those children who (a) are non-native speakers and (b) are deemed in need of targeted ESL language support in school.

An additional argument for advising against norms for ESL groups is that there is a wide variation among different subgroups of the ESL population. Rather than masking this variation by providing overall ESL norms, we suggest that further investigations of specific ESL subgroups identify language-specific needs that can then be addressed via educational support. Finally, it needs to be noted that studies examining ESL in relation to school success have identified that the socioeconomic status of the children's families and communities has a strong relation to the children's academic achievement (for the BC context, see Toohey & Derwing, 2006). Analyses linking EDI scores to socioeconomic status at the community level support this claim (Kershaw et al., 2005). In regard to the ESL comparison, we would therefore like to conclude by saying that the DIF analysis identified group differences that, in fact, are to be expected on those items that refer to English communication skills. After all, that distinction is the main criterion for designating children as ESL. Accordingly, the displayed DIF most probably is item impact, and not item bias, as it refers to actually occurring differences between the groups on the characteristic that is being measured.

## Gender Group Comparison

For our gender comparison, we identified one item with (uniform) DIF. This item belonged to the EDI Emotional Maturity domain and was related to physical ag-

gression (i.e., "Would you say this child gets into physical fights?"). On this item, boys had a higher probability of obtaining a higher (i.e., more physically aggressive, because the item is reverse coded) score than girls, after matching boys and girls on their total EDI school readiness score. In other words, boys with the same overall school readiness as girls were, on average, perceived and rated as more physically aggressive than girls by their Kindergarten teachers. This finding coincides with numerous child development studies that suggest that boys, on average, tend to be more physically aggressive (e.g., Alink et al., 2006; Hyde, 1984). Therefore, it can be assumed that this finding also represents a case of item impact and not item bias, as the statistical gender DIF can be assumed to reflect actual group differences with respect to the characteristic that is being measured (emotional maturity, as a domain of school readiness). It is important to add that the DIF identified for this item had no effect at the domain score level. This can be attributed to the fact that all other items did not display DIF of substantial effect size, and because the influence of 1 item on a scale with 30 items is relatively small. It can thus be concluded that, in regard to gender, the reporting of EDI scores, which is done solely at the domain score level, is unaffected by the presence of DIF in one item.

### Aboriginal Background Group Comparison

In regard to the comparison between Aboriginal and non-Aboriginal children, our analyses did not identify any DIF, implying that the EDI is not affected by measurement bias in regard to children's Aboriginal status. In other words, children's Aboriginal status did not seem to systematically bias Kindergarten teacher's ratings. This being said, it must, however, be emphasized that the EDI results may not—despite the large sample size—be representative of the diversity of Aboriginal children and their communities, because numerous on-reserve schools (with a high number/proportion of Aboriginal children) opted out from participating in the EDI assessment.[7]

## Conclusion

Current and future EDI-related research has been and will have to continue to be conducted in order to validate the EDI in an ongoing and context-dependent manner. The studies in this special issue provide an illustrative overview of how research projects in combination address different aspects of validity. One question raised by our findings pertains to the group differences that are, for each of the

[7]Currently, the Human Early Learning Partnership at the University of British Columbia, the organization that coordinates the EDI project in BC, is collaborating with numerous stakeholders toward developing an early childhood education tool that more clearly integrates Aboriginal values and their cultural diversity.

three groupings, consistent across all five developmental domains of the EDI, and, in some cases, of a large effect size.

We refer the interested reader to a number of studies that have examined related issues in the BC context (e.g., Ministry of Education, British Columbia, 2006; Toohey & Derwing, 2006), the Canadian context (e.g., Bonneau & Lauzon, 2006; Bowlby, 2006), or in similar U.S. contexts (e.g., Duncan et al., 2006). Further EDI-related studies examining these group differences specifically can hopefully provide further information to meaningfully interpret these differences.

The DIF analysis of the EDI data from Kindergarten children in BC presents results that are foundational for the establishment of the EDI's validity. Due to the representativeness of our sample (nearing census dimensions) in regard to ethnic diversity, demographics, all socioeconomic status strata, community contexts, and so on, our results may be assumed to also be generalizable to other North American jurisdictions that share some of the same characteristics (e.g., high degree of ethnic diversity). In regard to Kindergarten teachers' perceptions and ratings of children, the results suggest that the EDI is fair and unbiased in regard to gender, ESL status, and Aboriginal status. The identified cases of DIF can compellingly be related to research findings and be referred to as item impact, accordingly. These findings provide useful information for the interpretation of other EDI-related research, as measurement bias can, to a certain degree, be ruled out as a confounding issue.

## ACKNOWLEDGMENTS

Portions of this paper were included in a paper presented at the Annual Meeting of the American Educational Research Association, as part of a symposium entitled, "Translating school readiness assessment into community actions and policy planning: The Early Development Instrument Project," San Francisco, CA, April, 2006. The authors wish to acknowledge the support from the Human Early Learning Partnership (HELP) at the University of British Columbia, Vancouver, BC, Canada.

## REFERENCES

Alink, L. R. A., Mesman, J., van Zeijl, J., Stolk, M. N., Juffer, F., Koot, H. M., et al. (2006). The early childhood aggression curve: Development of physical aggression in 10- to 50-month-old children. *Child Development, 77,* 954–966.

Angenent, H., & de Man, A. F. (1989). Intelligence, gender, social maturity, and school readiness in Dutch first-graders. *Social Behavior and Personality, 17,* 205–209.

Bonneau, C. E., & Lauzor, J. (2006). First nations learners and extracurricular activities: Barriers and bridges to participation. *Educational Insights, 10*(1). Retrieved November 10, 2006, from http://www.ccfi.educ.ubc.ca/publication/insights/v10n01/pdfs/bonneau.pdf

Bowlby, G. (2006). *Provincial drop-out rates—Trends and consequences.* Retrieved November 10, 2006, from www.statcan.ca/english/freepub/81-004-XIE/2005004/drop.htm

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Chiappe, P., & Siegel, L. S. (1999). Phonological awareness and reading acquisition in English- and Punjabi-speaking Canadian children. *Journal of Educational Psychology, 91,* 20–28.

Cohen, J. (1992). . *Psychological Bulletin, 112,* 155–159.

Dauber, S. L., Alexander, K. L., & Entwisle, D. R. (1993). Characteristics of retainees and early precursors of retention in grade: Who is held back? *Merrill-Palmer Quarterly, 39,* 326–343.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2006, November). *School readiness and later achievement* (Working paper). Evanston, IL: Northwestern University, Institute for Policy Research.

Gelin, M. N., Carleton, B. C., Smith, M. A., & Zumbo, B. D. (2004). The dimensionality and gender differential item functioning of the Mini Asthma Quality of Life Questionnaire (MINIAQLQ). *Social Indicators Research, 68,* 91–105.

Gullo, D. F., & Burton, C. B. (1992). Age of entry, preschool experience, and sex as antecedents of academic readiness in kindergarten. *Early Childhood Research Quarterly, 7,* 175–186.

Hyde, J. S. (1984). How large are gender differences in aggression? A developmental meta-analysis. *Developmental Psychology, 20,* 722–736.

Janus, M., Brinkman, S., Duku, E., Hertzman, C., Santos, R., Sayers, M., et al. (2007). *The Early Development Instrument: A population-based measure for communities. A handbook on development, properties, and use.* Hamilton, Ontario, Canada: Offord Centre for Child Studies.

Janus, M., & Offord, D. (2007). Development and psychometric properties of the Early Development Instrument (EDI): A measure of children's school readiness. *Canadian Journal of Behavioural Science, 39,* 1–22.

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

Kershaw, P., Irwin, L., Trafford, K., & Hertzman, C. (2005). *The British Columbia atlas of child development* (1st ed.). Victoria, British Columbia, Canada: Human Early Learning Partnership and Western Geographical Press.

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,* 933–953.

Lesaux, N., & Siegel, L. S. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39,* 1005–1019.

McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology, 37,* 273–298.

Miller, J. R. (1996). *Shingwaukís vision: A history of native residential schools*. Toronto, Ontario, Canada: University of Toronto Press.

Ministry of Education, British Columbia. (2006, September). *Foundation skills assessment. 2001/02–2005/06.* Retrieved November 10, 2006, from

Offord, D., & Janus, M. (1999). *Early Development Instrument. A population-based measure for communities (2004/05 version).* Retrieved November 20, 2006, from www.offordcentre.com/readiness/EDI_viewonly.html

Shimizu, Y., & Zumbo, B. D. (2005). A logistic regression for differential item functioning primer. *Japan Language Testing Association Journal, 7,* 110–124.

Toohey, K., & Derwing, T. M. (2006, July). *Hidden losses: How demographics can encourage incorrect assumptions about ESL high school students' success* (Working Paper). Vancouver, British Columbia, Canada: Simon Fraser University.

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, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defence.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics,* (pp. 45–79). Amsterdam: Elsevier Science.

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

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