Journal of Educational Research & Policy Studies

Volume 5 Number 1 2005

An Electronic Reprint of:

Zumbo, B. D., & Gelin, M.N. (2005). A Matter of Test Bias in Educational Policy Research: Bringing the Context into Picture by Investigating Sociological / Community Moderated (or Mediated) Test and Item Bias. *Journal of Educational Research and Policy Studies, 5,* 1-23.

Journal of Educational Research & Policy Studies

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

Bruno D. Zumbo, Michaela N. Gelin
University of British Columbia

Abstract
The present study introduces and demonstrates a new methodology for item and test bias studies: "moderated differential item functioning (DIF)." This technique expands the DIF methodology to incorporate contextual and sociological variables as moderating or mediating effects of the DIF. Specifically, this paper explores differential domain functioning (DDF) – the focus of interpretation for this test is on the "domain" rather than the item. This moderated DDF effect is demonstrated on a multiple-choice and constructed-response provincial assessment test that was designed to match a specified mathematics curriculum. Participants were 45,728 grade four students, 45,022 grade seven students, and 43,525 grade 10 students in British Columbia, Canada. The data from these participants was narrowed down to create four contrast groups of communities that reflect differences in contextual variables: rural low-income, rural affluent, urban low-income, and urban affluent. Gender DDF was explored using a general linear statistical model. After statistically matching males and females on their mathematical ability, gender DDF was moderated by the contextual variables. Thus, this "moderation" approach allows one to investigate the effect of sociological, community-based contextual variables that may help one understand the complex functioning of DIF in large scale testing. In other words, what the authors are advocating is to take a more "sociological" and "ecological" approach to help educators understand differences in item and test performance.

Introduction

Determining whether an item on a test displays bias or impact has a number of significant implications for researchers, selection personnel, test takers, and policy makers. The primary issue is one of consequential matters of test fairness and equity. That is, there should be a level playing field where, for example, male and female students have equal opportunities to do well in a large-scale assessment, and hence being treated equitably in terms of test score performance.

Differential item functioning (DIF) is a statistical technique that is used to identify differential item response patterns between groups of test-takers (e.g., male vs. female, Caucasian vs. African American) and thus aids in identifying potentially biased test items. In assessing response patterns, the comparison groups, for example males and females, are first statistically matched on the underlying construct of interest (e.g., verbal abilities or mathematics achievement) and then the DIF methods evaluate the response patterns to individual test items. Thus, as Zumbo (1999) posits, DIF occurs when examinees with the same underlying ability on the construct measured by the test, but who are from different groups, have a different probability of correctly answering (or endorsing) the item. He continues with a conceptualization of the basic principle of DIF: "If different groups of test-takers (e.g., males and females) have roughly the same level of something (e.g., knowledge), then they should perform similarly on individual test items regardless of group membership" (p. 5).

DIF is different than previous classical test theory techniques used to assess bias because DIF *matches* the groups on the *latent* variable of interest; previous bias studies compared mean scores either without any matching technique or simply compared the factor structure for the groups of interest. See Zumbo (2003), and Zumbo and Koh (2005) for a demonstration of how comparing factor structures may miss item bias. Previous studies that found group differences on observed scores, such as group comparisons of scale or item means, may be misleading because respondents are not first matched on the construct of interest. Thus, matching groups on the variable measured by the test is important for determining whether item responses are equally valid for different groups. However, it should be noted that DIF is a statistical method designed to flag potentially problematic items. Therefore, it is the

first step in determining whether there is item bias or item impact. Further study would be needed by content experts to determine whether one has bias or impact.

Item bias is a value judgment with social, political, and ethical implications, and thus, takes into account the purpose of the test. Specifically, item bias requires that the source of the differential functioning of the item is *irrelevant* to the purpose of the test and/or interpretation of the measure. In essence, item bias is an artifact of the testing procedure. That is, item bias would occur if one group of test-takers (e.g., males) were less likely to get an item correct (or endorse the item) than the comparison group of test-takers (e.g., females) because the item is tapping a factor over-and-above the factor of interest. For example, if females were less likely to endorse an item from an achievement test of mathematical ability than males because the question required prior knowledge of sports terms that the females are not familiar with, then the item is biased. In addition, having knowledge of sporting terms is irrelevant for the purpose of the test. Thus, for item bias to occur DIF must be apparent; however, as Zumbo (1999) cautions, "DIF is a necessary, but not sufficient, condition for item bias" (p. 12).

Item impact is evident when one group of examinees is found to endorse the item more than the other group of examinees because the two groups truly differ on the underlying ability or factor being measured by the test. That is, item impact occurs when the item measures a *relevant* characteristic of the test, and 'real' differences between the two groups of interest are found. For example, if females were less likely to endorse an item from an achievement test of mathematical ability than men matched on mathematical ability, and men and women truly differed on mathematical aptitude, item impact is present.

The distinction between whether the group differences are based on irrelevant or relevant characteristics of the measure is really a question of the purpose of the measure. Therefore, one needs to be clear about the purpose of the test before conducting the analysis. As well it is important to note that if an item is flagged as displaying DIF it does not mean that the item should be automatically omitted from the scale. Rather, experts in the appropriate area should carefully analyze items that are flagged as displaying DIF. For example, if a mathematics achievement item were flagged as displaying

DIF then mathematics educators should carefully analyze why the item was flagged.

Uses of DIF

There are three general uses for DIF:

1. *Fairness and equity in testing.* This purpose of DIF is often because of policy and legislation in which the groups (e.g., visible minorities or language groups) are defined ahead of time.
2. *Dealing with a possible "threat to internal validity."* In this case, DIF is often investigated so that one can make group comparisons and rule-out measurement artifact as an explanation for the group difference. The groups are identified ahead of time and are often driven by an investigators research questions (e.g., gender differences in depression).
3. *Trying to understand the (cognitive and/or psycho-social) processes of item responding and test performance, and investigating whether these processes are the same for different groups of individuals*. In this context the groups are not identified ahead of time and instead latent class or other such methods are used to "identify" or "create" groups and then these new "groups" are studied to see if one can learn about the process of responding to the items.

Purpose and Structure of this Paper

A variety of statistical methods have been developed over the years to aid the researcher in identifying DIF items, for the purposes described above. This paper introduces a new methodology to study the role of contextual variables in differential item functioning (DIF) analyses. DIF is a statistical methodology that is often used in the process of developing new assessment measures, evaluating differential item response patterns of existing measures, and validating test score inferences for policy research (Zumbo & Hubley, 2003). Typically, DIF is explored by using gender or ethnic grouping variables. In addition, a few DIF studies have explored cognitive factors to help understand when and why DIF occurs. Furthermore, no DIF method, to our knowledge, has explicitly incorporated moderator or mediator effects of contextual sociological / community variables to help understand when, why, and to what degree DIF may occur. Our focus in the present paper is moderated DIF; however, our methodology is easily extended to mediated

DIF and hence will not be discussed further in this paper. In a DIF framework, moderator effects would occur if a moderator variable (e.g., socioeconomic status) influences the *direction* or *magnitude* of DIF. In other words, the DIF depends on some other contextual factor(s).

Moderated DIF

Given that moderated DIF will be explored using a statistical modeling approach, we will describe it within the context of an example using regression DIF methodology. In a typical regression DIF analysis one models each item as:

$$Y = b_0 + b_1\text{TOT} + b_2\text{GRP} + b_3\text{TOT*GRP}, \quad (1)$$

where TOT denotes the conditioning variable, GRP the grouping variable(s), and TOT*GRP the interaction term of the grouping effects variable and the matching variable.

The moderated DIF would be an extension of the above stated model wherein the gender DIF effect (i.e., the direction and/or magnitude of DIF) depends on the level of a third variable (i.e., the moderating variable), such as the income group of the examinees. In this sense, the DIF effect is moderated by level of income. Finally, given that family income can be considered a contextual (or systemic) variable, the DIF effect would be moderated by a contextual variable. That is, DIF not only depends on the grouping variable (e.g., gender), but the presence or absence of DIF is moderated by another variable, such as family income.

One can expand equation (1) to incorporate the moderating variables:

$$\begin{aligned} Y = b_0 + b_1\text{TOT} + b_2\text{GRP} + b_3\text{TOT*GRP} \\ + b_4\text{INCOME} \\ + b_5\text{TOT*INCOME} \\ + b_6\text{GRP*INCOME} \\ + b_7\text{TOT*GRP*INCOME}, \end{aligned} \quad (2)$$

where the notation is the same except for INCOME which denotes the moderating variable, in this case, income level.

Just as typical DIF modeling has a natural hierarchy of entering variables into the model, wherein the sequence is first the conditioning or matching variable, second the main effect, and third the interaction terms, moderated DIF modeling also has a natural hierarchy. That is, the expanded statistical model includes the following variables in sequence: (a) the conditioning or

matching variable (i.e., total scale score), (b) the main effect DIF grouping variable (e.g., gender), (c) the interaction between the conditioning variable and DIF grouping variable, (d) the main effect moderating variable (i.e., contextual variable), (e) the interaction between the conditioning variable and moderating variable, (f) the interaction between the DIF grouping variable and moderating variable, and finally, (g) the three-way interaction between the conditioning variable, DIF grouping variable and moderating variable. In short, the number of explanatory variables in the DIF regression model is increased with an eye toward a better understanding of the sociological process of gender differences. Given the contextual nature of this variable it may be at the individual or community level. For the described model above, the regression variable (b) is the uniform DIF, (c) is the non-uniform DIF, (f) is the moderated uniform DIF effect, and (g) is the moderated non-uniform DIF effect. If a moderated DIF effect is found, post-hoc analyses could be used to investigate at which levels of the moderating variable the DIF effect is present.

Numerous theories in the social sciences postulate the existence of moderated relationships. It can be argued that moderator variables are also relevant in the area of assessment. For example, standardized assessments must be fair so that examinees with equal ability levels have an equal probability of correctly answering each task. In general, equality among examinees is commonly investigated in relation to gender and ethnicity; however, equality among examinees also implies comparisons among other contextual factors such as those at the community level (e.g., rural versus urban locations), as well as those at the individual level (e.g., parental education). Thus, moderated DIF methodology allows one to take into account theoretical relationships among contextual variables.

This paper will address the notion of using socio-cultural variables as explanatory variables in DIF and particularly as moderating variables. The moderated DIF methodology will be introduced in the context of a case study. Note that the case study is particularly unique because it exploits the availability of linked data to address the matter of contextual variables.

Case Study

Moderated DIF will be demonstrated in the case of numeracy from a standardized assessment. The numeracy component was chosen because of

the heightened awareness from educators and policymakers that mathematics education has an important role in our technological world, and is thus often called the "critical filter" for success (Frempong & Willms, 1999). Although gender differences in mathematics performance are decreasing over time and the average gender difference in mathematics performance is small (Friedman, 1994; Frost, Hyde, & Fennema, 1994; Hyde, Fennema, & Lamon, 1990), gender differences on tests involving numeracy continue to be of major concern to educational researchers (e.g., Friedman, 1994; Leder, 1992; Ryan & Fan, 1996; Tate, 1997).

To accurately interpret gender differences in performance assessments DIF is essential because it aids us in ruling out item and test bias in explanations for the observed gender differences in test performance. Likewise, fairness among examinees is important because standardized achievement tests for elementary and secondary school students are commonly used to (a) provide information on student learning in selected areas of the curriculum in relation to national standards, (b) assist in curriculum and program development, and (c) to aid policy–makers and school administrators in decision-making. Standardized achievement tests are also useful for identifying subpopulations of students who require additional help and support. More recently, these tests are also being used as part of an educational accountability system that assesses teacher, school, and district performance (British Columbia Ministry of Finance and Corporate Relations, 1996; Raham, 1998). As previously mentioned, fairness among examinees implies fairness among *all* possible subgroups of examinees including those from different ethnic groups, SES groups, and those who live in different neighborhoods (e.g., rural and urban).

A number of recent studies focusing on gender-related DIF in mathematics assessments have identified item characteristics such as item format, and item content which may influence students' performance on mathematics tests (Garner & Englehard, 1999; Harris & Carlton, 1993; Lane, Wang, & Magone, 1996; O'Neil & McPeek, 1993; Ryan & Chiu, 2001; Ryan & Fan, 1996; Scheuneman & Grima, 1997). The interaction of gender DIF and sociological/ community moderating variables, however, has not been investigated. Although different contextual variables such as classroom size (Ehrenberg, Brewer, Gamoran, & Willms, 2001), socio-economic status (SES) (Chao & Willms, 2000; Frempong & Willms, 1999), teaching practices (McCaffrey et al., 2001), and parental styles (Chao & Willms, 2000) have been explored in

relation to school achievement, such variables have not been investigated as moderating variables in DIF analyses or even item or test bias.

Participants

The participants in this study consisted of 45,728 grade four students, 45,022 grade seven students, and 43,525 grade 10 students in British Columbia who took the Foundation Skills Assessment (FSA) examination in the spring of 2000 and whose school-level information matched the 1996 Stats Canada census survey. The matching of school-level census attributes to FSA data was important as it allowed the school code to be subsequently matched to student level FSA data. The major assumption of the methodology is that for a given set of students in a school with the same postal code, the aggregate value of the census characteristic is equal to that of the community in the area with the postal code. Due to time differences between reporting periods (i.e., 1996 census data and 2000 FSA data) demographic characteristics for some students who wrote the FSA were unavailable. In those cases, the missing census data was the result of schools that did not exist in 1996 and were opened in a following year. Efforts were made to update the census data to more current school codes, but there were still some school codes that did not return census attributes. In those cases, the missing census data was the result of school postal codes being incorrect or not available at the time the census data was compiled. In total, 867 students in grade four who took the FSA were eliminated from the data extraction because their student records could not be matched to census data. As a result, the total number of grade four students included in the data extraction was 47,014, of which 1,286 had missing values for some of the contextual variables (i.e., family income < 30K; rural versus urban) investigated in this study, and a total of 45,728 cases were therefore used for the analyses of grade four students. Similarly, 1,419 grade seven students' records and 2,324 grade 10 students' records could not be matched to the census data. The sample sizes for each grade level, by gender and age, are provided in Table 1 (see following page).

The Foundation Skills Assessment

Moderated gender DIF will be demonstrated using students' responses on the numeracy component of a provincial examination called the Foundation Skills Assessment (FSA). The numeracy component was designed to measure critical thinking skills in mathematics that are embedded in the British

Columbia education curricula for students in grades four, seven, and ten. The main purpose of the FSA was to help the province, individual schools, and districts evaluate how well important foundation skill areas are being addressed in order to make plans for improvement. It is because of this main purpose that the matter of moderated gender DIF by contextual (sociological) variables is so important.

Table 1: *Number of Students for Each Grade, by Gender and Age*

Grade	*N*	**Male** *(n)*	**Female** *(n)*	**Age** Mean (Std. Deviation)	Min.	Max.
4	45,728	23,019	22,709	9.72 (0.47)	9	11
7	45,022	22,911	22,111	12.72 (0.49)	12	14
10	43,525	22,302	21,223	15.76 (0.55)	15	18

The numeracy component consists of 32 multiple choice items and four constructed response items that were designed to measure four major content domains: number, patterns and relations, shape and space, and statistics and probability (see Appendix for the table of specifications for the FSA). Inferences and comparisons are typically made based on these four domains. In addition to examinees item-by-item test information from the 2000 assessment year, their individual and school-level information is statistically linked with the 1996 Statistics Canada census data so that contextual variables at both school and individual levels can be investigated.

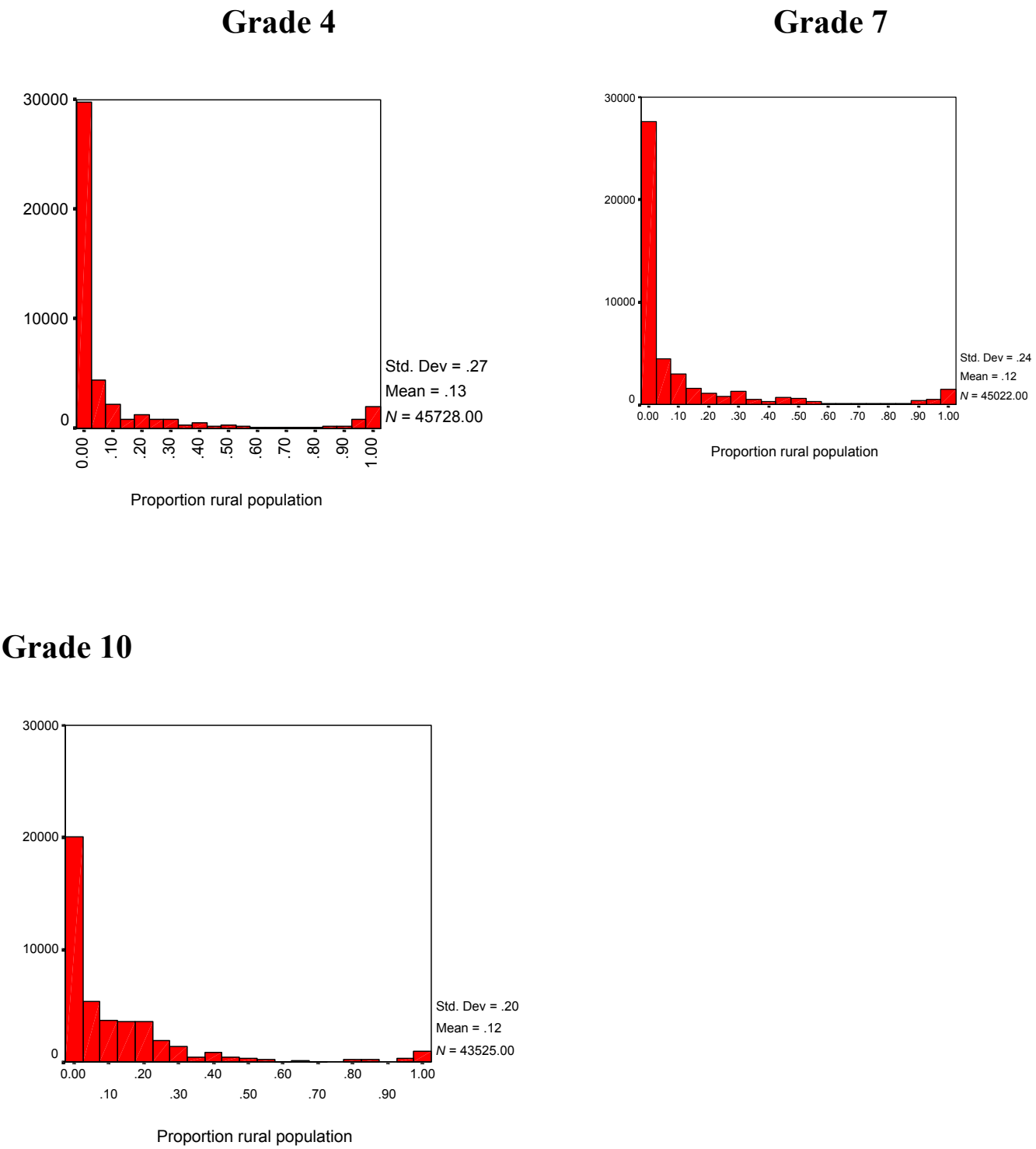
Moderator Variables for the Case Study

Moderated DIF explores external explanations (rather than solely internal cognitive explanations) for the potential item or test bias. Because of the wide use of rural vs. urban community and average family income as explanatory variables in the sociology of education literature, these variables are our moderator variables to introduce moderated DIF/DDF.

Rural versus urban contextual variable. The rural versus urban community is defined as the proportion of families who live in a rural location. The definition of rural is based on the Statistics Canada 1996 Census definition of rural as those areas with a population concentration of less than 1,000 and a population density of up to 400 per square kilometer (Statistics Canada,

2001a). Moreover, "community" in this study is the school community, and therefore, those students who attend the same school are designated as living in the same community. To create rural versus urban contrast groups, those students who attended schools in which the school community was designated as having 0% of families living in a rural area, were assigned as "urban." In contrast, those students who attended schools in which the school community was designated as having 50% or more of families living in a rural area were

Figure 1: *Proportion of Rural Population by Grade*



assigned as "rural." Because of these divisions, many cases were omitted from analyses because they were not clearly in an urban or rural community (see Figure 1).

Family income contextual variable. The second contextual variable is the family income level. There were two family income variables used in this study: the proportion of families whose income is (a) less than $20,000 per year, and (b) less than $30,000 per year. Canada's low-income thresholds are based on the size of households and size of community. Thus, large urban communities have higher income thresholds because of the higher cost of living, particularly housing costs (First Call, 2002). As cited in the latter reference and based on Statistics Canada data (Statistics Canada, 2001b) a three-person family living in an urban community with gross yearly income of $30,000 would be counted as living with a low income. That same family living in a rural community would not be counted as living with a low income. Rather, a three-person family living in a rural community with a gross yearly income of $20,000 would be counted as living with a low income. Therefore, in order to create appropriate income contrast groups, the rural or urban community location as described above was taken into account. For those students' who lived in a rural community, we used the family income variable less than $20,000, and for those students who lived in an urban community we used the family income variable less than $30,000. For each situation, affluent versus low-income families was designated to be in either the 10th percentile or less and the 90th percentile and above, respectively. Cases where the income level was between the 10th and 90th percentile were omitted from analyses because they did not fit within the contrast group definition. See Table 2 (on following page) for the community location by income level breakdown for each grade.

It is important to note that both these contextual variables are at the school-level and thus common to all students in sub-groups of the testing environment and not at the individual student level. As a result, students who attend the same school are grouped together. For example, for each student in the same school, their family income is based on the average family income of the school location. Likewise, all students who attend the same school are assumed to live in the same rural or urban location. Although these two variables are continuous, for methodological reasons we have made these two variables binary contrasting grouping variable(s). It should be further noted

that because most testing programs do not collect individual-level sociological/ contextual variables, it is anticipated that most of the studies investigating the moderating effect of community variables will have to rely on linked census data and hence community-level (rather than individual) moderating variables. By using these moderating variables, a more "sociological" and "ecological" approach is helpful in order to understand differences in item and test performance.

Table 2: *Cross Tabulation of Contrast Groups for Each Grade by Gender*

Grade	Contrast Group		Female	Male	Total
4 (*N*=4559)	**Rural**	**Low-income**	193	233	426
		Affluent	221	215	436
	Urban	**Low-income**	930	890	1820
		Affluent	917	960	1877
7 (*N*=4142)	**Rural**	**Low-income**	163	166	329
		Affluent	163	163	326
	Urban	**Low-income**	848	857	1705
		Affluent	874	908	1782
10 (*N*= 2577)	**Rural**	**Low-income**	150	143	293
		Affluent	108	108	216
	Urban	**Low-income**	466	539	1005
		Affluent	514	549	1063

These external reasons could be factors related to the particular testing context, such as opportunity-to-learn, facilities and resources available,

socioeconomic variables, and other characteristics of the environment in which the testing, learning, and day-to-day living are taking place. To investigate the relationship between such "macro-level" and external variables, linked data was used – linking school and community-level variables with the item response strings. The complex hierarchical structure of the data where students are in classrooms, classrooms in school, and schools in districts, etc. is already incorporated into our model, and thus, hierarchical linear modeling approaches are unnecessary. Contextual variables linked to item responses are already "grouped" at the school-level. Specifically, examinees' test information was statistically linked with census data based on the school in which the examinee was enrolled.

Analyses

Scale-level analyses. As a preliminary to the DIF modeling, the dimensionality of the numeracy items will be investigated via multi-group confirmatory factor analyses. As described, because the domain-level scores are the primary focus of interpretation, the analyses will be conducted at the test domain level for each grade. Hence, the DIF analyses will be referred to as "differential domain functioning" (DDF) for the remainder of the paper. General linear statistical modeling will be used to investigate the moderated DDF effects of family income and of the community location (rural versus urban).

Scale-level analyses: factor analysis models. The FSA numeracy component was hypothesized to be unidimensional because the item scores are summated to form a single score to measure numeracy ability. A simultaneous multi-group (by gender) maximum likelihood confirmatory factor analysis of a Pearson covariance matrix was conducted using LISREL 8.53 (Jöreskog & Sörbom, 2002). The Pearson matrix is appropriate because the observed variables are continuous. Table 3 (see following page) shows the results of the Chi-squared difference tests for full invariance models compared to the baseline models. The full invariance hypothesis tests equality of loadings and the equality of uniquenesses between genders. The full invariance model test was repeated for community group within each grade, and thus the full invariance model tests were conducted at an $\alpha = .01$, a Bonferroni-corrected α for the five model tests within each grade. The full invariance models between genders are rejected by the data for grade four overall and for grade

ten rural affluent. Therefore, strong invariance models (i.e., only equality of loadings across genders) were assessed for the two rejected full invariance models. In both cases, the strong invariance model was not rejected by the data: Grade four overall $\triangle\chi^2$ (4) = 7.73, p = .102, Grade ten rural affluent $\triangle\chi^2$ (4) = 8.59, p = .072, and therefore strong invariance holds in those cases where full invariance does not.

Table 3*: Simultaneous Tests for Full Invariance Model of Numeracy Between Genders Overall and Among Contrast Groups*

	Group		Full Invariance		
			$\triangle\chi^2$	$\triangle df$	*P*
Grade 4	Overall		20.90	8	**0.007**
	Rural	Affluent	4.53	8	0.806
		Low Income	8.14	8	0.420
	Urban	Affluent	14.11	8	0.079
		Low Income	12.46	8	0.132
Grade 7	Overall		14.74	8	0.064
	Rural	Affluent	12.89	8	0.116
		Low Income	8.84	8	0.356
	Urban	Affluent	15.33	8	0.053
		Low Income	8.65	8	0.373
Grade 10	Overall		16.49	8	0.036
	Rural	Affluent	20.90	8	**0.007**
		Low Income	7.72	8	0.461
	Urban	Affluent	13.22	8	0.105
		Low Income	13.95	8	0.083

Note: *P*-values in bold are statistically significant.

Item-level Analyses: Differential Domain Functioning Results

Table 4 (see following page) lists the results of the DDF analyses for each domain and grade. Table 5 (see following page) lists the results of the more complex model, moderated DDF, taking into account the community location and income level of families within the community. Upon comparing Tables 4 and 5 one sees the following:

1. From Table 4, small (measured by the effect size) DDF effects are found for some domains and in some grades. For example, there is both small uniform and non-uniform DDF for the number domain in grade four. However, this DDF is not apparent in grade ten.
2. Likewise, from Table 5, small DDF effects are apparent in some community groups across domains within grades. For example, there is a small uniform DDF for the number domain within rural low-income communities.
3. Comparing Tables 4 and 5, it is evident that the DDF found in Table 4 is not the same when one takes into account the community characteristics – i.e., the moderated DDF. For example, the DDF found in the number domain for grade four students is only apparent for rural low-income communities. Likewise, the number DDF found in Table 4 for grade seven students is not at all present in the moderated case in Table 5.

Discussion

The purpose of this paper is to introduce and demonstrate a new methodology for item and test bias studies: Mediated and moderated DIF (or in our case, moderated DDF), with a focus for our case study on moderation. This "moderation" or "mediation" approach allows one to investigate the effect of sociological, community-based contextual, variables that may help one understand the complex functioning of DIF in large-scale testing. Conventional DIF methodology either (a) ignores all other factors than the DIF variable, (b) focuses on cognitive variables, variables that characterize the person, or (c) focuses variables that characterize the item or task such as item format, item content, or item context within the test. This results of this study suggest that measurement specialists should broaden their view on what effects test performance to include characteristics of the situation in which the person is learning and/or taking the test.

From this case study, it is evident that if one ignores the rural or urban community location or income level one does not get the whole picture of

Table 4: *P-Values and, Where Appropriate, Effects Sizes for Gender DDF Effects*

		Domain			
		Number	**Patterns and Relations**	**Shape and Space**	**Statistics and Probability**
Grade 4	**Uniform DDF**	**.002** η^2**=.002**	.111	.256	.902
	Non-uniform DDF	**.004** η^2**=.002**	.149	.308	.881
Grade 7	**Uniform DDF**	**.003** η^2**=.002**	.314	.939	**.0001** η^2**=.010**
	Non-uniform DDF	.046	.468	.259	.009
Grade 10	**Uniform DDF**	.401	**.0001** η^2**=.010**	.697	**.003** η^2**=.003**
	Non-uniform DDF	.756	.011	.749	.627

Note: The per DDF tests were conducted at an α =.0063; a Bonferroni-corrected α for the 8 DDF tests per grade.

gender DDF. This study went beyond the conventional explanatory variables for DIF and considered that in British Columbia, where the FSA is conducted, there are large socio-geographic differences in the province. A majority of the population lives in large urban (or geographically close to urban) communities; however, a substantial number of individuals live in rural settings. Likewise, the urban/rural split brings with it differences in wealth, both personal wealth and community economic well-being. In short, the socio-geographic features are related to variables that impact education and opportunities to learn. Furthermore, the geography of British Columbia is such that the

Table 5: *P-Values and, Where Appropriate, Effect Sizes for Moderated Gender DDF Effects by Contrast Group*

	Domain	Rural				Urban			
		Affluent		Low-Income		Affluent		Low-Income	
		Uniform DDF	Non-Uniform DDF	Uniform DDF	Non-Uniform DDF	Uniform DDF	Non-uniform DDF	Uniform DDF	Non-uniform DDF
Grade 4	Number	.302	.474	**.003 η^2=.021**	.008	.473	.219	.192	.604
	Patterns and Relationships	.943	.966	.615	.961	.533	.495	.166	.248
	Shape and Space	.178	.173	.374	.447	.499	.530	.149	.076
	Statistics and Probability	.547	.777	.644	.591	.015	.049	.331	.712
Grade 7	Number	.257	.365	.024	.065	.011	.094	.210	.344
	Patterns and Relationships	.482	.609	.482	.459	.536	.389	.480	.971
	Shape and Space	.013	.039	.586	.816	.185	.188	.122	.097
	Statistics and Probability	.064	.087	.178	.220	.087	.163	.040	.661
Grade 10	Number	.367	.554	.977	.999	.076	.364	.898	.396
	Patterns and Relationships	**.002 η^2=.046**	.023	.531	.467	**.000 η^2=.019**	**.005 η^2=.007**	.070	.565
	Shape and Space	.384	.705	.228	.568	.857	.807	.868	.710
	Statistics and Probability	.024	.062	.832	.116	.01	.630	.266	.294

Note: The per DDF tests were conducted at an $\propto$ =.0063; a Bonferroni-corrected $\propto$ for the 8 DDF tests per community group, per grade.

individuals who live in rural communities are also less educated with lower participation rates in post-secondary education. In the end, moderated or mediated DDF will only be of value if one works in collaboration with educational sociologists and educational economists and policy makers (who are familiar with contextual community variables) to help understand why the community contextual variables are moderating the DDF.

Acknowledgements

A British Columbia Ministry of Education Research Grant funded this project. We would like to thank Pat McCrea for helping us secure the Ministry data. An earlier version of this paper was presented at the 2003 meeting of

Division D (Measurement and Research Methodology) of the American Educational Research Association in Chicago Illinois, U.S.A.

Appendix: *Table of Specifications for the FSA 2000 Numeracy Component*

	Content Areas	**Table of Specification Percentage**	**# of Marks**	**% of Test**
Grade 4	**Number** • Students apply their number sense to solve problems using whole numbers from 0 to 10,000 and proper fractions. • They use the basic arithmetic operations in whole number contexts.	35 - 45%	19	39
	Patterns and Relationships • Students investigate, establish and present rules for numerical and non-numerical patterns.	15 - 25%	8	17
	Shape and Space • Students estimate, measure and compare quantities using decimal numbers and standard units of measure. • They describe, classify and relate three-dimensional objects and two-dimensional shapes. • They use numbers and directional words to describe the relative positions of objects.	20 - 30%	12	25
	Statistics and Probability • Students collect, assess, validate and graph data. • They conduct simple probability experiments to explain outcomes.	10 - 20%	9	19

Grade 7	**Number** • Students solve problems involving numbers including decimal fractions and integers. • They use ratios, rates, percentages and decimal numbers in various contexts.	35 - 45%	20	41
	Patterns and Relationships • Students use expressions containing variables to make predictions. • They use variables and equations to express and summarize relationships.	15 - 25%	9	19
	Shape and Space • Students solve problems involving the properties of circles and their relationships to angles and time zones. • They link angle measurements to the properties of parallel lines. • They analyze patterns and designs using congruence, symmetry, translation, rotation and reflection.	20 - 30%	11	23
	Statistics and Probability • Students analyze data using measures of variability and central tendency. • They solve problems using probability.	10 - 20%	8	17

Grade 10	**Number** • Students solve problems involving numbers, including rational and irrational numbers. • They perform basic operations on the real number system and apply these skills in various practical, real-life or technical contexts.	25 - 35%	16	33
	Patterns and Relationships • Students use patterns to solve problems. • They simplify and manipulate algebraic expressions and make connections between algebraic and graphical representations.	20 - 30%	13	27
	Shape and Space • Students use trigonometry to analyze real-life situations. • They use geometry to analyze interrelationships among shapes.	25 - 35%	12	25
	Statistics and Probability • Students interpret, draw inferences and communicate statistical information. • They use probability terminology and determine permutations and combinations of possible events.	10 - 20%	7	15

Source: The British Columbia Ministry of Education (2000).

References

British Columbia Ministry of Education (2000, January/February). FSA 2000 Numeracy Specifications. *BC Education News*, p. 9.

British Columbia Ministry of Finance and Corporate Relations. (1996, June). *Executive summary, Report on accountability in the K-12 education system.* Internal Audit Branch, Office of the Comptroller General. Victoria, British Columbia, Canada.

Chao, R. K., & Willms, J. D. (2000). *Family income, parenting practices, and childhood vulnerability: A challenge to the "culture of poverty" thesis.* (Policy Brief Rep. No. 9). New Brunswick: Canadian Research Institute for Social Policy.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2001, November). Does class size matter? [Electronic version]. *Scientific American*, *6*, 1-6.

First Call – BC Child and Youth Advocacy Coalition (2002, November). BC Campaign 2000 Fact Sheet #1: What is poverty? Retrieved March 17th, 2003, from http://www.firstcallbc.org/whatsnew/02Fact sheet 1-What is Poverty.pdf

Frempong, G., & Willms, J. D. (1999). *Mathematics: The critical filter.* (Policy Brief Rep. No. 5). New Brunswick: Canadian Research Institute for Social Policy.Hyde, J.S., Fennema, E., & Lamon, S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155.

Friedman, L. (1994). Meta-analytic contributions to the study of gender differences in mathematics: The relationship of mathematical and spatial skills. *International Journal of Educational Research, 21*, 361-371.

Frost, L.A., Hyde, J.S., & Fennema, E. (1994). Gender, mathematics performance, and mathematics-related attitudes and affect: A meta-analytic synthesis. *International Journal of Educational Research, 21*, 373-385.

Garner, M., & Engelhard, G.J. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*, 29-51.

Harris, A.M., & Carlton, S.T. (1993). Patterns of gender differences on mathematics items on the scholastic aptitude test. *Applied Measurement in Education, 6*, 137-151.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155.

Jöreskog, K.G., & Sörbom, D. (2002). LISREL (Version 8.53) [Computer software]. Chicago, IL: Scientific Software International, Inc.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice, 15*, 21-27, 31.

Leder, G. (1992). Mathematics and gender: Changing perspectives. In D. Grows (Ed.), *Handbook of research on mathematics teaching and learning.* Reston, VA: National Council of Teachers of Mathematics.

McCaffrey, D.F., Hamilton, L.S., Stecher, B.M., Klein, S.P., Bugliari, D. & Robyn, A. (2001). Interactions among instructional practices, curriculum, and student achievement: The case of standards-based high school mathematics. *Journal for Research in Mathematics Education, 32*, 493-517.

O'Neill, K.A., & McPeek, W.M. (1993). Item and test characteristics that are associated with differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning* (pp. 255-279). Hillsdale, NJ: Erlbaum.

Raham, H. (1998, July-August). Building school success through accountability. *Policy Options,* 13-17.

Ryan, K.E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*, 73-90.

Ryan, K.E., & Fan, M. (1996). Examining gender DIF on a multiple-choice test of mathematics: A confirmatory approach. *Educational Measurement: Issues and Practice, 15*, 15-20, 38.

Scheuneman, J.D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance of female and black examinees. *Applied Measurement in Education, 10*, 299-319.

Statistics Canada. (2001a, November). Definitions of rural (Catalogue no. 21-0006-XIE). *Rural and Small Town Canada Analysis Bulletin, 3*(3), 1-17.

Statistics Canada. (2001b, November). *Low income cutoffs from 1991 to 2000 and low income measures from 1990 to 1999*. Income Statistics Division. (Catalogue no. 75F0002M-01007).

Tate, W.F. (1997). Race-ethnicity, SES, gender, and language proficiency trends in mathematics achievement: An update. *Journal for Research in Mathematics Education, 28*, 652-679.

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. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing, 20*, 136-147.

Zumbo, B.D., & Hubley, A.M. (2003). Differential Item Functioning and Item Bias. In R. Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment,* pp. 505-509. Thousand Oaks, CA.: Sage Press.

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of Differences in Item-Level Characteristics in Scale-Level Measurement Invariance Tests of Multi-Group Confirmatory Factor Analyses. *Journal of Modern Applied Statistical Methods, 4*, 275-282.