LANGUAGE ASSESSMENT QUARTERLY, *4*(2), 223–233

Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going

Bruno D. Zumbo
University of British Columbia, Canada

The purpose of this article is to reflect on the state of the theorizing and praxis of DIF in general: where it has been; where it is now; and where I think it is, and should, be going. Along the way the major trends in the differential item functioning (DIF) literature are summarized and integrated providing some organizing principles that allow one to catalog and then contrast the various DIF detection methods and to shine a light on the future of DIF analyses. The three generations of DIF are introduced and described with an eye toward issues on the horizon for DIF.

Methods for detecting differential item functioning (DIF) and item bias typically are used in the process of item analysis when developing new measures, adapting existing measures for use in new settings or with populations not initially intended when the measure was developed, adapting existing measures to new languages and/or cultures, or more generally validating test score inferences. DIF methods allow one to judge whether items (and ultimately the test they constitute) are functioning in the same manner in various groups of examinees. In broad terms, this is a matter of measurement invariance; that is, is the test performing in the same manner for each group of examinees?

In what follows I reflect on the state of the praxis and theorizing of DIF in general: where it has been; where it is now; and where I think it is, and should, be going. Along the way I intend to integrate and summarize major trends in the DIF

Correspondence should be addressed to Bruno D. Zumbo, Department of ECPS, 2125 Main Mall, Scarfe Building, The University of British Columbia, Vancouver, British Columbia, Canada V6T 2G8. E-mail: bruno.zumbo@ubc.ca

literature, provide some organizing principles that allow one to catalog and then contrast the various DIF detection methods, and shine a light on what I believe is the future of DIF analyses. Those involved in this work have come to address a number of critical, and recurring, issues that face the future of DIF. These critical issues are threaded throughout.

I propose that we consider three generations of DIF praxis and theorizing. In so doing, I am not suggesting distinct historical periods and a natural linear stepwise progression toward our current thinking. In fact, in using the expression "generations of DIF" I want to suggest quite the contrary. Note also that given the general purpose of this article, throughout I use the terms *test* and *measure* interchangeably.

THE FIRST GENERATION: MOTIVATIONS FOR THE PROBLEM AND CONCEPT FORMATION

In the first generation of DIF, the more commonly used term was *item bias*. Concerns about item bias emerged within the context of test bias and high-stakes decision making involving achievement, aptitude, certification, and licensure tests in which matters of fairness and equity were paramount. Historically, concerns about test bias have centered around differential performance by groups based on gender or race. If the average test scores for such groups (e.g., men vs. women, Blacks vs. Whites) were found to be different, then the question arose as to whether the difference reflected bias in the test. Given that a test comprises items, questions soon emerged about which specific items might be the source of such bias.

Given this context, many of the early item bias methods focused on (a) comparisons of only two groups of examinees; (b) terminology such as *focal* and *reference* groups to denote minority and majority groups, respectively; and (c) binary (rather than polytomous) scored items. Due to the highly politicized environment in which item bias was being examined, two interrelated changes occurred. First, the expression *item bias* was replaced by the more palatable term *differential item functioning,*' or DIF in many descriptions. DIF was the statistical term that was used to simply describe the situation in which persons from one group answered an item correctly more often than equally knowledgeable persons from another group. Second, the introduction of the term *differential item functioning* allowed one to distinguish item impact from item bias. Item impact described the situation in which DIF exists, because there were true differences between the groups in the underlying ability of interest being measured by the item. Item bias described the situations in which there is DIF because of some characteristic of the test item that is not relevant to the underlying ability of interest (and hence the test purpose).

SECOND GENERATION: EMBODYING THE NEW TERMS AND BUILDING FRAMEWORKS FOR EMPIRICALLY INVESTIGATING DIF

The transition to the second generation of DIF was signaled by the widespread acceptance of the term *DIF* rather than *item bias*, and the concomitant separation of *impact* and *bias*, but also in the introduction of new statistical DIF methods that embodied these ideas and distinctions. Carrying on from the first generation, consumers of DIF methodology and technology were mostly educational and psychological measurement specialists. As a result, research primarily focused on developing sophisticated statistical methods for detecting or "flagging" DIF items rather than on refining methods to distinguish item bias from item impact and providing explanations for why DIF was occurring. Along the way in this second generation, although a relatively small number of nonmeasurement specialists became interested in exploring DIF and item bias in tests, it was apparent that much of the statistical terminology and software being used was not very accessible to many researchers.

With an eye toward encapsulating the work of this second generation of DIF, I overview how the field views DIF in this generation. At least three frameworks for thinking about DIF have evolved in the literature: (a) modeling item responses via contingency tables and/or regression models, (b) item response theory (IRT), and (c) multidimensional models. Although these frameworks may be seen as interrelated, they are freestanding. Each framework provides useful organizing principles for describing DIF and developing methods for detecting DIF in items.

Modeling Item Responses via Contingency Tables and/or Regression Models

A statistical implication of the definition of DIF that arose in the first generation (i.e., persons from one group answering an item correctly more often than equally knowledgeable persons from another group) is that one needs to match the groups on the ability of interest *prior* to examining whether there is a group effect. That is, the definition of DIF implies that after conditioning on (i.e., statistically controlling for) the differences in item responses that are due to the ability being measured, the groups still differ. Thus, within this framework, one is interested in stating a probability model that allows one to study the main effects of group differences (termed *uniform DIF*) and the interaction of group by ability (termed *nonuniform DIF*) after statistically matching on the test score.

This class of DIF methods, in essence, consists of conditional methods in that they study the effect of the grouping variable(s) and the interaction term(s) over-and-above (i.e., while conditioning on) the total score. In this sense, they share a lot in common with analysis of covariance (ANCOVA) or Attribute × Treatment interaction (ATI) methods. Building on this similarity, it is important to recognize that nearly all DIF methods are applied in what would be called an observational or quasi-experimental study design, and so one must keep in mind all of the commonly known caveats around making causal claims of grouping variable effects in observational studies involving intact groups.

This framework for DIF has resulted in two broad classes of DIF detection methods: Mantel-Haenszel (MH) and logistic regression (LogR) approaches. The MH class of methods (Holland & Thayer, 1988) treats the DIF detection problem as one involving, in essence, three-way contingency tables. The three dimensions of the contingency table involve (a) whether one gets an item correct or incorrect and (b) group membership, while conditioning on (c) the total score discretized into a number of category score bins. The LogR class of methods (Swaminathan & Rogers, 1990) entails conducting a regression analysis (in the most common case, a logistic regression analysis as the scores are binary) for each item wherein one tests the statistical effect of the grouping variable(s) and the interaction of the grouping variable and the total score after conditioning on the total score. One clear contrast between the MH and LogR methods is that one needs to discretize the conditioning variable in the MH methods whereas one does not have to do so with the LogR methods. The MH assumes no interaction (like ANCOVA) whereas the LogR allows for an interaction (like ATI methods).

IRT Models

Referring back to the definition of DIF formulated in the first generation of DIF, one can approach DIF from an IRT framework. In this case, one considers two item characteristic curves (ICCs) of the same item but computed from two groups. In the IRT context, if the items exhibit DIF, then the ICCs will be identifiably different for the groups. The ICCs can be identifiably different in two common ways. First, the curves can differ only in terms of their threshold (i.e., difficulty) parameter, and hence the curves are displaced by a shift in their location on the theta continuum of variation. Second, the ICCs may differ not only on difficulty but also on discrimination (and/or guessing), and hence the curves may be seen to intersect. Within this context, the former represents uniform DIF (i.e., a main effect of group), whereas the latter represents nonuniform DIF (i.e., an interaction of group by ability).

In its essence, the IRT approach is focused on determining the area between the curves (or, equivalently, comparing the IRT parameters) of the two groups. It is noteworthy that, unlike the contingency table or regression modeling methods, the IRT approach does not match the groups by conditioning on the total score. That is, the question of “matching” only comes up if one computes the difference function between the groups conditionally (as in MH or LogR). Comparing the IRT parameter estimates or ICCs is an unconditional analysis because it implicitly assumes that the ability distribution has been “integrated out.” The mathematical expression “integrated out” is commonly used in some DIF literature and is used in the sense that one computes the area between the ICCs *across the distribution* of the continuum of variation, theta.

A problem occurs in the IRT context because it is a latent variable modeling approach. Because the scale for theta in any IRT model is arbitrary, one must set it during calibration. How is this resolved? Computing algorithms like BILOG (and other such 2PL/3PL varieties of calibration software) set the mean of the ability distribution at zero. Some Rasch calibration software typically set the mean of the item difficulties at zero, whereas others fix a single item parameter estimate, much like one does in confirmatory factor analysis to fix the scale of the latent variable.

Another issue that arises in IRT DIF is that if the two groups have different ability distributions, then the scales for the groups will be arbitrarily different. This is a problem because, in the case of DIF, one wants the two groups on the same scale or metric. If the two groups are not on the same metric, any DIF results will be impossible to interpret. This matter of a common metric is important to highlight because, in several recent studies, some Rasch analysts have ignored this matter and computed the difference between the item difficulty parameter for the two groups with a *t* statistic, falsely relying on Rasch invariance claims to justify the computation and incorrectly ignoring the need for a common metric.

The most common IRT methods for DIF include signed area tests (which only focus on uniform DIF), unsigned area tests (which allow for nonuniform DIF), and nested model testing via a likelihood ratio test, which is most easily conducted for uniform DIF. In addition, one can approach this via nonparametric IRT using the software TestGraf (Ramsay, 2001). An advantage of nonparametric IRT is that it provides a graphical method and needs far fewer items and examinees than other IRT approaches.

Multidimensional Models

Based on the principle that DIF occurs because of some characteristic of the test item that is not relevant to the underlying ability of interest (and hence the test purpose), a long-standing framework has evolved for DIF based on the dimensionality of items. This framework begins with the assumption that all tests are, to some extent, multidimensional. The informal rationale has been that there is typically one primary dimension of interest in a test, but there may also be other dimensions within that test that produce construct-irrelevant variance. For example, in a problem-based test of mathematics, the test will consist of some primary dimension that reflects mathematics ability as well as some other dimensions that may reflect other secondary abilities such as reading comprehension or verbal abilities. These other dimensions are often correlated with the primary dimension. As part of this informal rationale, it was not uncommon to think of DIF as arising from dimensions other than those of primary interest in the test. Ackerman (1992) provided a thorough discussion of the basis for the multidimensional framework.

Stout and his colleagues (e.g., Shealy & Stout, 1993) formalized some of this thinking and introduced a new DIF test statistic, simultaneous item bias test (SIBTEST) based on their framework. The multidimensional approach to DIF, as implemented in SIBTEST, allows for a variety of scenarios that comprise differential dimensionality as the source for DIF. Because this method involves a type of factor analysis, it requires the analyst to study sets (or bundles) of items, rather than individual items for DIF.

That is, as has been noted by the proponents of the multidimensional approaches to DIF detection (see Gierl, 2005, for a nice overview of Roussos and Stout's work), the conventional manner in which one investigates DIF, outside of the multidimensional approach, is to individually examine all items on a test for DIF and then, if the results suggest DIF, those items are further studied by content specialists and others to ascertain possible reasons for the observed DIF and determine whether item impact or bias is present. Given that such DIF studies usually occur in the context of observational (rather than experimental) studies, the sources or causes of DIF may be difficult to establish. Thus, the conventional approach is an inductive or exploratory approach to investigating DIF.

Alternatively, as suggested by Roussos and Stout (1996), one could approach the DIF detection issue from a more theory-based and hypothetico-deductive strategy. That is, one would consult (with the aid of a content specialist) the relevant literature and determine whether any predictions (i.e., scientific hypotheses) can be made for where and why and for who DIF may be present. Once this has been accomplished, one then goes about testing the predictions using Stout's SIBTEST DIF detection methods. The attractiveness of this strategy for many is the hope that a theory-based approach will provide an explanation for why DIF would be present (i.e., from a multidimensional framework, the literature would identify the secondary dimension(s)) and whether the DIF reflects item impact or bias. Of course, the confirmatory (i.e., theory-based) strategy is most fruitful when the content literature is well developed. Unfortunately, what is not mentioned in this literature is that one could use any of the DIF methods to detect DIF in this hypothetico-deductive strategy, and not only SIBTEST.

TRANSITIONING TO THE THIRD GENERATION

The matter of wanting to know why DIF occurs is an early sign of the third generation of DIF theorizing. The multidimensional approach, in its praxis, places the source of DIF in the test structure and provides a well-integrated hypothesis-testing strategy. Of course, the concern for the sources (causes) of DIF predates the Rousos and Stout multidimensional approach. For example, the "why" concerns can be clearly seen in Angoff (1993) when he wrote about long-standing Educational Testing Service DIF work: "It has been reported by test developers that they are often confronted by DIF results that they cannot understand; and no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF values" (p. 19).

It is not widely acknowledged in the DIF research literature that a very useful strategy and corresponding statistical methodology was introduced early in the history of the second-generation DIF by Muthen and his colleagues to address the question of the sources, "the why question," of DIF (Muthen, 1985, 1988, 1989; Muthen, Kao, & Burstein, 1991; Muthen & Lehman, 1985). This class of approaches exploits among other things the multiple-indicators, multiple causes structural equation model, and how this model relates to item response theory. One way of conceptualizing Muthen's work is that it is a merging of the "Modeling Item Responses via Contingency Tables and/or Regression Models" and "Item Response Theory Models" frameworks just described while allowing for possible multidimensionality. An essential difference between the Muthen approach and the Roussos–Stout approach is that Muthen's approach more explicitly (and easily) allows the researcher to focus on sociological, structural community, and contextual variables as explanatory sources of DIF (Zumbo & Gelin, 2005). Muthen's work, from my view, most clearly signals the third generation of DIF methodology.

Third Generation: Current and Future Directions

In essence, the transition to the third generation is best characterized by a subtle, but extremely important, change in how we think of DIF—in essence, revisiting the first generation. That is, the third generation of DIF is most clearly characterized as conceiving of DIF as occurring because of some characteristic of the test item *and/or testing situation* that is not relevant to the underlying ability of interest (and hence the test purpose). By adding "testing situation" to the possible reasons for DIF that have dominated the first two generations of DIF (including the multidimensional model), one greatly expands DIF praxis and theorizing to matters beyond the test structure (and hence multidimensionality) itself, hence moving beyond the multidimensional model of DIF. For example, a number of studies focusing on gender-related DIF have investigated item characteristics such as item format and item content, which may influence students' performance on tests; however, contextual variables such as classroom size, socioeconomic status, teaching practices, and parental styles have been largely ignored in relation to explanations for (and causes of) DIF (Zumbo & Gelin, 2005).

The third generation of DIF is best represented by its uses, the praxis of DIF. There are five general uses that embody the third-generation praxis of DIF analyses and motivate both the conceptual and methodological developments in third-generation DIF.

Purpose 1: Fairness and equity in testing. This purpose of DIF is often because of policy and legislation. In this purpose, the groups (e.g., visible minorities or language groups) are defined ahead of time before the analyses and often set by the legislation or policy. Although this use of DIF is still important today, this is where the first two generations of DIF were clearly situated and DIF was conceived of and created with this purpose in mind.

Purpose 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). This purpose evolved as DIF moved away from its exclusive use in large-scale testing organizations and began to be used in day-to-day research settings. In essence, DIF is investigated so that one can make group comparisons and rule out measurement artifact as an explanation for the group differences.

Purpose 3: Investigate the comparability of translated and/or adapted measures. This use of DIF is of particular importance in international, comparative, and cross-cultural research. This matter is often referred to as construct comparability. Please see Kristjansson, Desrochers, and Zumbo (2003) and Hambleton, Merenda, and Spielberger (2006) for a discussion of developments in translation and adaptation.

Purpose 4: Trying to understand item response processes. In this use DIF becomes a method to help understand the cognitive and/or psychosocial processes of item responding and test performance and investigating whether these processes are the same for different groups of individuals. In this use, DIF becomes a framework for considering the bounds and limitations of the measurement inferences. In Zumbo’s (2007) view of validity, DIF becomes intimately tied to test validation, but not only in the sense of test fairness. The central feature of this view is that validity depends on the interpretations and uses of the test results and should be focused on establishing the inferential limits (or bounds) of the assessment, test, or measure (Zumbo & Rupp, 2004). In short, invalidity is something that distorts the meaning of test results for some groups of examinees in some contexts for some purposes. Interestingly, this aspect of validity is a slight, but significant, twist on the ideas of test and item bias of the first-generation DIF.

That is, as Zumbo (2007) and Zumbo and Rupp (2004) noted, test and item bias aim analyses at establishing the inferential limits of the test—that is, establishing for whom (and for whom not) the test or item score inferences are valid.

In this context the groups may be identified ahead of time by the researcher. However, in this use of DIF it is most fruitful if the groups are not identified ahead of time and instead latent class or mixture modeling 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. One can approach this from developments in mixture latent variable modeling developed by Muthen in his Mplus software, as well as by other forms of mixture and latent class models. Two excellent exemplars of this sort of work are Li, Cohen, and Ibarra (2004) and DeAyala, Kim, Stapleton, and Dayton (2003).

Purpose 5: Investigating lack of invariance. In this purpose DIF becomes an empirical method for investigating the interconnected ideas of (a) lack of invariance, (b) model-data fit, and (c) model appropriateness in model-based statistical measurement frameworks like IRT and other latent variable approaches—for example, invariance is an assumption for some forms of computer-based testing, computer-adaptive testing, linking, and many other IRT uses. A particularly promising approach that builds beyond DIF is Rupp's (2005) new techniques for examining examinee profiles.

The direction and focus of third-generation DIF praxis and theorizing has been shaped by its origins in test bias and high-stakes decision making involving achievement, aptitude, certification, and licensure tests. Current directions in DIF research find their inspiration from considering many testing situations outside of test bias, per se. Today, in addition to matters of bias, DIF technology is used to help answer a variety of basic research and applied measurement questions wherein one wants to compare item performance between or among groups when taking into account the ability distribution. At this point, applications of DIF have more in common with the uses of ANCOVA or ATI than test bias per se.

This broader application has been the impetus for a variety of current and future directions in DIF development, such as test translation and cross-cultural adaptation. Many novel applications of DIF occur because previous studies of group differences compared differences in mean performance without taking into account the underlying ability continuum. An example of such an application in language testing would be a study of the effect of background variables such as discipline of study, culture, and hobbies on item performance.

Moving beyond the traditional bias context has demanded developments for DIF detection in polytomous, graded-response, and rating scale (e.g., Likert) items. Furthermore, because nonmeasurement specialists are using DIF methods increasingly, it has been necessary to develop more user-friendly software and more accessible descriptions of the statistical techniques as well as more accessible and useful measures of DIF effect size for both the binary and polytomous cases (Kristjansson, Aylesworth, McDowell, & Zumbo, 2005; Zumbo, 1999).

Clearly the pronouncements I hear from some quarters that psychometric and statistical research on DIF is dead or near dying are obviously overstated. The academic journals are chock full of psychometric studies on the technical issues of DIF, because as we move into the third generation of DIF many methodological problems appear on the horizon. Several examples from my work suffice to demonstrate the liveliness of DIF research. For example, in a forthcoming paper I propose the use of statistical methods for probing interactions (e.g., the Johnson–Neyman technique and other contemporary variations on this method) as a way of understanding nonuniform DIF—a problem that has plagued DIF research for decades. I have ongoing research focusing on complex data situations wherein one has students nested within classrooms, classrooms nested within larger school organizations, and a myriad of contextual variables at each level that are potentially related to DIF. New methods are being developed to study the contextual variables while remaining true to the complex data structure with random coefficient models and generalized estimating equations.

ACKNOWLEDGMENTS

I thank Professor Antony J. Kunnan for inviting me to write this article and for patiently waiting for it to materialize. I also thank Dr. Anita M. Hubley, Professor Steve Sireci, Dr. Brian French, Professor Susan Maller, as well as my current and former doctoral students Anne M. Gadermann, Martin Guhn, Amery Wu, Yan Liu, Zhen Li, Dr. Andre A. Rupp, Dr. Michaela Gelin, Dr. Kim Koh, Dr. Charles Ochieng Owour, Dr. Krista Breithaupt, and the many participants from around the world in my day-long workshops on DIF and Measurement Invariance for many spirited conversations and debates that allowed me to formulate, try out, and develop the ideas presented in this article.

REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Erlbaum Associates.

DeAyala, R. J., Kim, S-H., Stapleton, L. M., & Dayton, C. M. (2003). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing, 2,* 243–276.

Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice, 24*(1), 3–14.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2006). *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Kristjansson, E., Aylesworth, R., McDowell, I., & Zumbo, B. D. (2005). A comparison of four methods for detecting DIF in ordered response items. *Educational and Psychological Measurement, 65,* 935–953.

Kristjansson, E. A., Desrochers, A., & Zumbo, B. D. (2003). Translating and adapting measurement instruments for cross-cultural research: A guide for practitioners. *Canadian Journal of Nursing Research, 35*, 127–142.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4,* 115–136.

Muthen, B. O. (1985). A method for studying the homogeneity of test items with respect to other relevant variables. *Journal of Educational Statistics, 10,* 121–132.

Muthen, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Lawrence Erlbaum Associates.

Muthen, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 551–585.

Muthen, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: An application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28,* 1–22.

Muthen, B. O., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10,* 133–142.

Ramsay, J. O. (2001). *TestGraf: A program for the graphical analysis of multiple-choice test and questionnaire data* [Computer software and manual]. Montreal, Canada: McGill University.

Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20,* 355–371.

Rupp, A. A. (2005). Quantifying subpopulation differences for a lack of invariance using complex examinee profiles: An exploratory multi-group approach using functional data analysis. *Educational Research and Evaluation, 11,* 71–95.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159–194.

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

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

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

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.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 73–92). Thousand Oaks, CA: Sage.