Metodologia Delle Scienze Sociali

Statistical Methods for Investigating Item Bias in Self-Report Measures

The University of Florence Lectures on Differential Item Functioning

Bruno D. Zumbo, Ph.D.
Professor
University of British Columbia
Vancouver, Canada

Presented at the Universita degli Studi di Firenze
July 3, 2008, Florence (Firenze), Italy

1

1

Addendum since first publication

- With the growing interest in this lecture, this material has come to be referred to in the research literature, and among psychometricians and data analysts as:
  - "*The University of Florence Lectures on Differential Item Functioning*",
  - or the "*Florence Lectures on DIF*".

Citation: Zumbo, B.D. (2008). *Statistical Methods for Investigating Item Bias in Self-Report Measures, [The University of Florence Lectures on Differential Item Functioning]*. Universita degli Studi di Firenze, Florence, Italy.

2

2

- I want to begin by thanking Prof.ssa Filomena Maggino for arranging for this visit and for being so generous with her time in the planning and organization.
- It is a tremendous honour for me to be here.

3

3

Overview

- Methods for detecting differential item functioning (DIF) and scale (or construct) equivalence typically are used in developing new measures, adapting existing measures, or validating test score inferences.
- DIF methods allow the judgment of whether items (and ultimately the test they constitute) function in the same manner for various groups of examinees, essentially flagging problematic items or tasks. In broad terms, this is a matter of measurement invariance; that is, does the test perform in the same manner for each group of examinees?
- You will be introduced to a variety of DIF methods, some developed by the presenter, for investigating item-level and scale-level (i.e., test-level) measurement invariance. The objective is to impart psychometric knowledge that will help enhance the fairness and equity of the inferences made from tests.

4

4

Agenda

To Do:

1. **What is measurement invariance, DIF, and scale-level invariance?**
2. **Construct versus item or scale equivalence and the three generations of DIF**
3. **Description of DIF methods**
4. **DIF for Ordinal / Likert / Rating Scale Items**
5. **Scale Level Effect of DIF: Graphical Method**
6. **Recommendations**

Note: Nearly all of this presentation will deal with "measures" and other "self-report" instrument but, if time permits, I will discuss some new ideas at the end about how to extend this work to "indicators" using a structural equation modeling framework.

Reference list at the end of the notes.

5

5

Why are we looking at item or scale-level bias?

- Technological and theoretical changes over the past few decades have altered the way we think about test validity and the inferences we make from scores that arise from tests and measures.
- If we want to use the measure in decision-making (or, in fact, simply use it in research) we need to conduct research to make sure that we do not have bias in our measures. Where our value statements come in here is that we need to have organizationally and socially relevant comparison groups (e.g., gender or minority status).

6

6

Why are we looking at item or scale-level bias?

- In recent years there has been a resurgence of thinking about validity in the field of testing and assessment. This resurgence has been partly motivated by the desire to expand the traditional views of validity to incorporate developments in qualitative methodology and in concerns about the consequences of decisions made as a result of the assessment process.
- For a review of current issues in validity we recommend the recent papers by Zumbo (2007b), Hubley and Zumbo (1996) and the edited volume by Zumbo (1998).
- I will focus on item bias first and then turn to the scale-level approaches toward the end.

7

7

Agenda

Done:

1. What is measurement invariance, DIF, and scale-level invariance?

To Do:

2. Construct versus item or scale equivalence and the three generations of DIF
3. Description of DIF methods
4. DIF for Ordinal / Likert / Rating Scale Items
5. Scale Level Effect of DIF: Graphical Method
6. Recommendations

8

8

Item and Test Bias

- Methods for detecting differential item functioning (DIF) and item bias typically are used in the process of developing new measures, adapting existing measures, or 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?

9

9

Item and Test 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 is comprised of items, questions soon emerged about which specific items might be the source of such bias.

10

10

Item and Test 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 inter-related changes occurred.
  - First, the expression 'item bias' was replaced by the more palatable term 'differential item functioning' or DIF in many descriptions.

11

11

Item and Test Bias

  - 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 or testing situation that is not relevant to the underlying ability of interest (and hence the test purpose).

12

12

Item and Test Bias

- Traditionally, consumers of DIF methodology and technology have been educational and psychological measurement specialists.
- As a result, research has 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.

13

13

Item and Test Bias

- Although this is changing as increasing numbers of non-measurement specialists become interested in exploring DIF and item bias in tests, it has become apparent that much of the statistical terminology and software being used is not very accessible to many researchers.
- In addition, I believe that we are doing these days is what I like to call the "***Third Generation DIF***" (Zumbo, 2007a). My presentation today will try and highlight this 3rd Generation.

14

14

Third Generation DIF (Zumbo, 2007a)

- **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).**

15

15

Third Generation DIF

- **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 multi-dimensional 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, socio-economic 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.**

16

16

Third Generation DIF

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.
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.

17

17

Third Generation DIF

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.
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 (2007b) view of validity, DIF becomes intimately tied to test validation; but not only in the sense of test fairness.

18

18

Third Generation DIF

4. Trying to understand item response processes. ... continued…

- 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) note, test and item bias aim analyses at establishing the inferential limits of the test – i.e., establishing for whom (and for whom not) the test or item score inferences are valid.

19

19

Third Generation DIF

4. Trying to understand item response processes. … continued…

- 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, as well as by other forms of mixture and latent class models.

20

20

Third Generation DIF

5. **Investigating lack of invariance.**

- In this purpose DIF becomes an empirical method for investigating the interconnected ideas of:
  - lack of invariance ,
  - model-data fit, and
  - 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.

Third Generation DIF

- 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.

Third Generation DIF

- 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.
- **Whatever your purpose (one of the 5 or some combination of them) you will want to know about DIF methods … so lets turn to describing some of the DIF methods.**

23

23

Third Generation DIF

- **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 non-measurement 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)**
- **Therefore, in the next section I will start with binary item response data and then move to the ordinal item response data -- also called 'polytomous', 'rating scale', or 'Likert' item response formats in the following section of the presentation.**

24

24

Agenda

Done:

1. **What is measurement invariance, DIF, and scale-level invariance?**
2. **Construct versus item or scale equivalence and the three generations of DIF**

To Do:

3. **Description of DIF methods**
4. **DIF for Ordinal / Likert / Rating Scale Items**
5. **Scale Level Effect of DIF: Graphical Method**
6. **Recommendations**

25

25

Overview of DIF Methods

- **In this section we will:**
  - **discuss some of the conceptual features of DIF;**
  - **discuss some of the statistical methods for DIF detection (there are many DIF methods we do not include, but from the ones I include you can master the others);**
  - **consider an example with an eye toward showing you how to conduct these analyses in SPSS.**

26

26

Conceptual Matters I

- It is important to note that in this presentation we have focused on 'internal' methods for studying potential item bias; i.e., within the test or measure itself. It is important for the reader to note that there is also a class of methods for studying potential item bias wherein we have a predictor and criterion relationship in the testing context.
- For example, one may have a test that is meant to predict some criterion behavior. Item (or, in fact, test level) bias then focuses on whether the criterion and predictor relationship is the same for the various groups of interest.

27

27

Conceptual Matters I

- 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?

28

28

Conceptual Matters II: What is DIF?

- DIF is the statistical term that is 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.
- In the context of social, attitude or personality measures, we are not talking about “knowledge”, per se, but different endorsement (or rating) after matching on the characteristic or attitude or interest.

29

29

Conceptual Matters II b: What is DIF?

- The key feature of DIF detection is that one is investigating differences in item scores after statistically matching the groups.
- In our case the “groups” may represent different language or adaptations of the test or measure, groups that you want to compare on the measure of interest, or groups legislated for investigation for bias.

30

30

Conceptual: DIF versus Item Bias

- DIF is a necessary, but insufficient condition for *item bias*.
- Item bias refers to unfairness of item toward one or more groups
- Item bias is different than *item impact* which describes 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.

31

31

Statistical frameworks for DIF

- At least three frameworks for thinking about DIF have evolved in the literature: (1) modeling item responses via contingency tables and/or regression models, (2) item response theory, and (3) multidimensional models (see Zumbo, 2007a for details).
- We will focus on the first framework.

32

32

One, of several, ways of classifying DIF methods

		Matching variable	
		Observed score	*Latent variable*
Item format	*Binary*	MH LogR	Conditional IRT Multidimensional SEM
	Polytomous	Ordinal LogR	Conditional IRT Multidimensional SEM

33

33

Our focus will be here and we will include MH methods as well.

		Matching variable	
		Observed score	*Latent variable*
Item format	*Binary*	MH LogR	Conditional IRT Multidimensional SEM
	Polytomous	Ordinal LogR	Conditional IRT Multidimensional SEM

34

34

Statistical Definition I

- A statistical implication of the definition of DIF (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.

35

35

Statistical Definition II

- That is, DIF exists when assuming that the same items are responded by two different groups, 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.

36

36

Statistical Definition III

- 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 'non-uniform DIF') after statistically matching on the test score.
- Conceptually, this is akin to ANCOVA or ATI; therefore, all of the same caveats about causal claims in non-experimental designs apply. (DIF is a non-experimental design)

37

37

Two classes of statistical methods, we focus on one.

- Two broad classes of DIF detection methods: Mantel-Haenszel (MH) and logistic regression (LogR) approaches.
- We will focus on LogR because it is the most general of the two methods allowing us to (a) investigate uniform and non-uniform DIF, (b) easily allow for additional matching variables, (c) easily extend our methods to rating or Likert scales, and (d) builds on the knowledge-base of regression modeling more generally defined .

38

38

DIF Statistical Analysis

- The LogR class of methods (Swaminathan & Rogers, 1990; Zumbo 1999) entails conducting a regression analysis (in the most common case, a logistic regression analysis as the scores are binary or ordinal LogR for Likert item scores) 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.

39

39

Logistic Regression

$$P(U=1)=\frac{e^{z}}{1+e^{z}}$$

$$Z=T_0+T_1\theta+T_2G$$
Tests Uniform DIF

$$Z=T_0+T_1\theta+T_2G+T_3\theta G$$
Tests Non-uniform DIF

40

Logistic Regression II

- In the DIF detection regression methods, regression model building and testing has a natural order.
- That is,
  - (i) one enters the conditioning variable (e.g., the total score) in the first step and then
  - (ii) the variables representing the group and the interaction are entered into the equation and
  - (ii) one tests whether step two variables that were added statistically contribute to the model via looking at the 2-degree of freedom test and a measure of effect size.

41

41

Example

- A synthetic sample of military officers in Canada (532 monolingual English speaking; 226 monolingual French speaking).
- This is a general aptitude test that measures verbal, spatial, and problem-solving skills, with 15, 15, and 30 items respectively.
- The question at hand is the adequacy of the translation of the problem-solving items.

42

42

Example (cont'd.)

- What do we match on to compare item performance, item by item of the problem-solving scale?
  - The standard answer is the total score of the problem-solving.
- We are going to match on both the total score on the problem-solving scale and on the total of the spatial-abilities scale (hence matching on general intellectual functioning and not just problem-solving)
  - Because it is not a verbal measure, per se, the spatial-abilities scale is not effected by language it is administered.

43

43

Example (cont'd) Item #1

Lets start with the typical DIF analysis and condition only on the total scale of problem-solving (we are, after all, studying a problem-solving item)

- Used indicator coding for language (English =1, French =0)
- Chi-Squared, 2-df, 61.6, p<=.001.
- So, statistically we have signs of a DIF item.

44

44

Output from SPSS

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	problem	.218	.057	14.712	1	.000	1.244
	lang(1)	2.317	.825	7.883	1	.005	10.148
	lang(1) by proble	-.001	.061	.000	1	.983	.999
	Constant	-4.850	.718	45.576	1	.000	.008

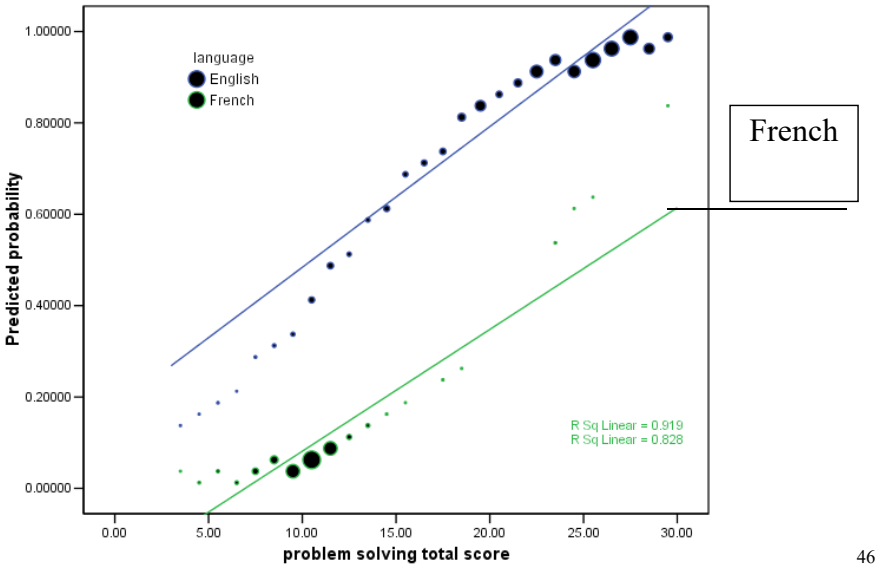
a. Variable(s) entered on step 1: lang, lang * problem .

(Conditional) Odds ratio

45

45

Plot of the "predicted probabilities of LogR analysis

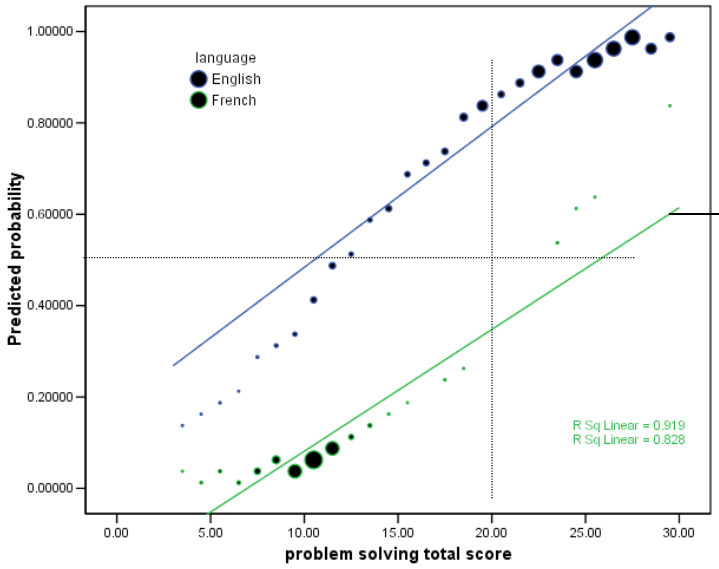


French

46

46

Plot of the "predicted probabilities of LogR analysis



French

47

47

This matching assumption

- **Please note that the matching assumption highlighted and described in the previous few slides applies to all DIF methods, not just Logistic regression, MH, or the latent variable variable approaches.**

48

48

Example (cont'd) Item #1

Now lets do the multiple conditioning example with

- Used indicator coding for language (English =1, French =0)
- Chi-Squared, 2-df, 53.6, p<=.001.
- So, statistically we have signs of a DIF item.

49

49

Example (cont'd) Item #1

Variables in the Equation --- Note the results for "Lang"

	Wald	df	Sig.	Exp(B)
PROBLEM	15.489	1	.000	1.251
SPATIAL	1.168	1	.280	.953
LANG(1)	6.759	1	.009	8.704
LANG(1) by Problem	.006	1	.939	1.005
Constant	31.198	1	.000	.012

50

50

Example (cont'd) Item #1

- We conclude the item #1 displays DIF in this synthetic sample with an partial odds-ratio of over 8 times in favor of the English version – i.e. the French version of that item was over 8 times more difficult than the English version, while matching on general intellectual functioning.
- This analysis would be conducted for each of the 30 problem-solving items

51

51

Logistic Regression: Assessment on Balance

- **Strengths**
  - **Multivariate matching criteria (multidimensional matching).**
  - **Can test for both uniform and non-uniform DIF**
  - **Significance test (like IRT likelihood ratio)**
  - **Effect size (T_2)**
  - **Popular software (SAS, SPSS).**
- **Limitations (relative to IRT)**
  - **Observed score metric**
  - **May need to purify matching score for scales with few items.**

52

52

Logistic Regression: Assessment on Balance

- **Note that, at this point in the session, I do not talk about the various effect size indices available for quantifying the magnitude of DIF, and how to use them in our DIF practice.**
- **I will, however, talk about effect sizes and their use a little later in this presentation. Please note that whatever I say about the use of DIF effect sizes in the later section (on rating scale or Likert items) applies to the binary case as well.**
  - **The rating scale or Likert items are a natural place to talk about effect size but the point applies to binary items as well.**

53

53

DIF: MH Method (optional)

- **Mantel-Haenszel: another conditional procedure like LogR**
  - **Like standardization, typically uses total test score as conditioning variable**
- **Extension of chi-square test**
- **3-way Chi-square:**
  - **2(groups)-by-2(item score)-by-K(score levels)**
  - **The conditioning variable is categorized into k levels (i.e., bins)**

54

54

Mantel Haenszel Formulas

Item Score

Group	1	0	Total
Reference	A_j	B_j	N_{rj}
Focal	C_j	D_j	N_{rj}
Total	M_{1j}	M_{0j}	T_1

$$H_0: \frac{A_j / C_j}{B_j / D_j} = 1 \qquad H_A: A_J / C_J = \alpha \, B_J / D_J$$

$$MH\chi^2 = \frac{\left[\left|\sum_j A_j - \sum_j E(A_j)\right| - .5\right]^2}{\sum_j Var(A_j)}$$

55

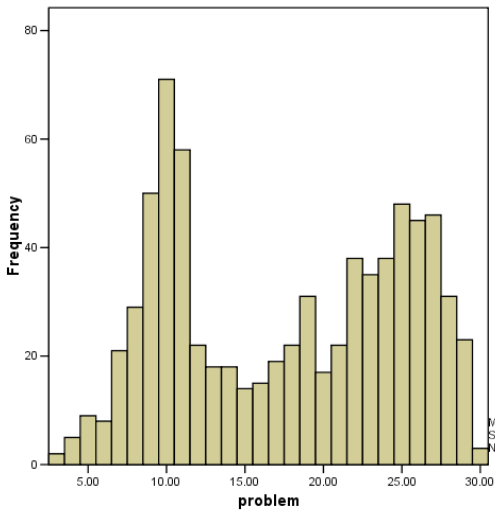
Lets look at the item #1 from LogR

- **Cannot easily do multiple matching variables so lets look at the conventional matching situation.**
- **We need to divide up (bin) the total scale score so that we can do the MH. Because there is not a lot of range of scores lets do a fairly thick matching and divide it into two equal parts. {I tried other groupings and ended-up with missing cells in the 3-way table}.**

56

56

Histogram of the scale score



Statistics

problem

N	Valid	758
	Missing	0
Percentiles	25	10.0000
	50	18.0000
	75	25.0000

Cut offs for grouping; at median

57

57

SPSS result: 3-way table of DIF on item #1 from above

grouping * problem solving * language Crosstabulation

Count

language			problem solving 0	problem solving 1	Total
English	grouping	1.00	68	93	161
		2.00	27	344	371
	Total		95	437	532
French	grouping	1.00	205	15	220
		2.00	3	3	6
	Total		208	18	226

Still problematic

58

58

SPSS result: MH test of DIF on item #1 from above

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	103.266	1	.000
Mantel-Haenszel	100.613	1	.000

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

59

59

DIF: MH Method (optional)

- **Strengths:**
  - **Powerful**
  - **Has statistical test and effect size**
  - **Popular software (SAS, SPSS)**
- **Limitations**
  - **Does not test for non-uniform DIF**
  - **Unlike LogR, you need to put the matching score into bins (i.e., levels) and this is somewhat arbitrary and may affect the statistical decision regarding DIF.**

60

60

Important to keep in mind …

- You need to assume that the items are, in essence, equivalent and that the grouping variable (or variable under study) is the reason that the items are performing differently.
- Need to assume that the matching items are equivalent across languages
  - Screened items
  - Non-verbal items
- No systematic bias because DIF will not detect this.

These assumptions must be defended!

61

61

Measurement of the construct is equivalent across groups

- Scale-level bias
  - Are there substantive differences across groups in the operationalization or manifestation of the construct?
- Scale-level equivalence does not guarantee item level equivalence (or vice versa)
  - This was shown by Zumbo (2003) in the context of translation DIF for language tests.

62

62

Agenda

Done:

1. **What is measurement invariance, DIF, and scale-level invariance?**
2. **Construct versus item or scale equivalence and the three generations of DIF**
3. **Description of DIF methods**

To Do:

4. DIF for Ordinal / Likert / Rating Scale Items
5. Scale Level Effect of DIF: Graphical Method
6. Recommendations

63

63

Ordinal (Likert / Rating Scale) Items

- **Let now turn to:**

  Statistical and Graphical Modeling to Investigate Differential Item Functioning for Rating Scale and Likert Item Formats

- *Based on work with Michaela Gelin and Suzanne Slocum.*

64

64

Ordinal (Likert / Rating Scale) Items

- **The purpose of this section is to describe a statistical modeling approach and introduce a graphical display for studying DIF for rating scale or Likert-type items.**
- **Recognizing that it more effective to illustrate a statistical technique in the context of an example rather than describe in abstract terms, the remainder of this section will use the Center of Epidemiology Scale of Depression (CES-D) to motivate the presentation.**

65

65

Differential Item Functioning of the CES-D

- **The measurement of depression in the general population is increasingly more common in educational, psychological, and epidemiological research. One of the most widely used instruments is the Center of Epidemiology Scale of Depression (CES-D) (Radloff, 1977).**
- **It is a commonly used community / population health self-report measure of depressive symptomotology.**
- **The CES-D is comprised of 20 items that ask respondents to rate, on a 4-point Likert scale, the frequency in which they experience depressive symptoms. A total score is computed by summing the 20 item responses. The CES-D instructions and items 17 and 20 of the scale are shown in Figure 1. We will be using items 17 and 20 to illustrate the DIF methods.**

66

66

Instructions and Example Items from the CES-D

Figure 1.
Instructions and Example Items from the CES-D
The CES-D
For each statement, circle the number (see the guide below) to indicate how often you felt or behaved this way **during the past week**.

0 = rarely or none of the time (less than 1 day)
1 = some or a little of the time (1-2 days)
2 = occasionally or a moderate amount of time (3-4 days)
3 = most or all of the time (5-7 days)

	Less than 1 day	1-2 days	3-4 days	5-7 days
2. I did not feel like eating; my appetite was poor.	0	1	2	3
17. I had crying spells.	0	1	2	3

67

67

Participants for our example

- **The data are from a community-based survey conducted in Northern British Columbia as part of the activities of the Institute for Social Research and Evaluation at the University of Northern British Columbia.**
- **The participants are 724 adults between the ages of 17 and 92 years (357 females; 367 males).**
  - **Males in the study are statistically significantly older (female: 44 years old; male: 47 years old, t(722)=3.1, p=0.002, eta-squared=0.013), although the effect size is not very large.**
- **Missing values for item responses were imputed by using an EM algorithm in the software PRELIS.**

68

68

DIF Example with the CES-D

- **How does DIF apply in the context of the CES-D?**
  - **An example would be gender DIF.**
- **It is widely reported in the literature that there are gender differences in the CES-D total scores, which leads us to the question:**
  - **Are these differences due to measurement artifact?**
  - **By measurement artifact we mean that the difference is due to the way that depression is being measured. That is, men and women are responding differently to one or more of the items, and this difference in item responses is an *artifact* of the measurement process – item bias.**

69

69

Measurement Artifact

- **There are several important methodological points that need to be highlighted when performing DIF analyses.**
- **The first is that the focus of the DIF analysis is to study each item individually.**
- **And the second is that one needs to perform a statistical matching to make the comparison between genders meaningful.**
- **DIF would mean that the typical item responses are different for males and females after matching on total scores; that is taking into account depressive symptoms.**
- **See Zumbo (2007a) for details on matching and the variety of DIF models available.**

70

70

How do we test for DIF?

- **In this paper, we will focus on two classes of approaches:**
  - **the first is the statistical modeling for DIF and the second is a graphical display.**
  - **When generating a statistical model for DIF, as Zumbo (2007a) states, many of the DIF techniques are akin to ANCOVA, wherein we statistically condition (match) on a variable of interest and then investigate group differences.**
  - **Over and above what we are trying to measure, is there a group difference for each item? If so, then there is DIF.**

71

71

How do we test for DIF?

- **There are several types of regression modeling one can utilize when investigating potential DIF.**
- **One can either use logistic regression, ordinal logistic regression, or ordinary regression depending on the dependent variable in the model and the assumptions one is willing to make about the error term.**
- **That is, logistic regression is used for binary item scores, ordinal logistic regression is used for rating scale or Likert items, and ordinary least-squares regression is used for continuous or Likert items that have many scale points (e.g., more than 6 scale points).**

72

72

How do we test for DIF?

- **In our example with the CES-D, we used ordinal logistic regression because of the 4-point Likert scale item response format.**
  - **Herein I am making several simplifying assumptions to write the ordinal logistic regression model.**

73

73

How do we test for DIF?

- The binary logistic regression methods we have discussed so far in this presentation apply when we have a categorical response of the simplest possible form - dichotomous.
- It is natural to consider methods for more categorical responses having more than two possible values.
- A variety of methods have been developed for covering the various possibilities. The best known and most highly developed are methods for ordinal response variables.
  - Note that that a categorical variable is considered ordinal if there is a natural ordering of the possible values, for example strongly disagree, disagree, agree, strongly agree.
- A number of proposed models for this type of data are extensions of the logistic regression model.
  - The most well known of these ordinal logistic regression methods methods is called the proportional odds model

74

74

How do we test for DIF?

- The basic idea underlying the proportional odds model is re-expressing the categorical variable in terms of a number of binary variables based on internal cut-points in the ordinal scale.
- Ordinal logistic regression is but one method currently available for investigating
- DIF for items commonly found in personality and social psychological measures. I selected ordinal logistic regression rather than a generalized Mantel-Haenszel (M-H; Agresti, 1990) or logistic discriminant function analysis (Miller & Spray, 1993) because:
  - (a) using ordinal logistic regression has the advantage of using the same modeling strategy for binary and ordinal items,
  - (b) this common statistical model for binary and ordinal items should ease the process of implementation in an organization where DIF analyses are not yetcommon, and
  - (c) the Zumbo-Thomas DIF effect size method can be extended to ordinal logistic regression hence, unlike the other methods (e.g., generalized M-H or logistic discriminant function analysis), one has a test statistic and a natural corresponding measure of effect size.

75

75

How do we test for DIF?

- One can interpret logistic regression as a linear regression of predictor variables on an unobservable continuously distributed random variable, y^*.
- Thus, a linear form of a typical regression equation can be re-expressed as

$$y^* = b_0 + b_1 TOT + b_2 GENDER + b_3 TOT * GENDER_i + \varepsilon_i ,$$

where the errors are, for the logistic model, with mean zero and variance $\pi^2/3$.

76

76

How do we test for DIF?

- From this and some additional conditions, one can get an R-squared like quantity for ordinal logistic regression (see, Thomas, Zhu, Zumbo, & Dutta, 2008; Zumbo, 1999).
- Also please note that this ordinal logistic regression model makes some important assumptions
  - It operates on the principle of cumulative information along the latent variable.
    - That is, for example, for a 3-point response an ordinal logistic regression model describes two relationships: the effect of X (in our case the total score for the scale) on the odds that Y 1 instead of Y > 1 on the scale, and the effect of X on the odds that Y 2 instead of Y > 2.
    - Of course, for our three point scale, all of the responses will be less than or equal to three (the largest scale point) so it is not informative and hence left out of the model. The model requires two logistic curves, one for each cumulative logit.
  - At any given point on the X-axis the order of the two logistic curves is the same.

77

77

How do we test for DIF?

- In summary, if we had for example a 3-point ordinal (Likert or rating) scale for an item, a ordinal logistic regression models the odds that someone will select the scale point 2 (or less) on the scale in comparison to selecting a response higher on the scale.
- Furthermore, the regression model does this for each point on the scale simultaneously.
- What a researchers ends up with is a regression equation having more than one intercept coefficient and only one slope.
  - The common slope assumption could be tested with a nominal multinomial logit model.

78

78

Statistical Modeling

- DIF methods allow one to judge whether items are functioning the same in various groups. One needs to match these groups on the ability of interest prior to examining whether there is a group effect. One is interested in stating the probability model in studying main effects of group differences (uniform DIF) and interaction of group by ability (non-uniform DIF).
- In order to conduct these analyses we used SPSS because it is widely available and used in the social and health sciences.
- See Zumbo (1999) for a detailed description.

79

79

Statistical Modeling

- As described by Zumbo (1999), the ordinal logistic regression (OLR) procedure uses the item response as the dependent variable, with the grouping variable (denoted as GRP), total scale score for each examinee (denoted as TOTAL) and a group by total interaction as independent variables.
- This can be expressed as a linear regression of predictor variables on a latent continuously distributed random variable, y*.

The ordinal logistic regression equation is

$$y^* = b_0 + b_1 TOTAL + b_2 GRP_i + b_3 TOTAL * GRP_i$$

80

80

Statistical Modeling

- Zumbo's OLR method has a natural hierarchy of entering variables into the model in which the conditioning variable (i.e. the total score) is entered first.
- Next, the grouping variable (e.g., gender) is entered. This step measures the effect of the grouping variable while holding constant the effect of the conditioning variable (uniform DIF).
- Finally, the interaction term (e.g., TOTAL*GENDER) is entered into the equation and describes whether the difference between the group responses on an item varies over that latent variable continuum (non-uniform DIF).
- Each of these steps provides a Chi-squared test statistic, which is used in the statistical test of DIF.

81

81

Statistical Modeling

- Just as a Chi-squared statistic is computed for each step in Zumbo's OLR DIF method, the corresponding effect size estimator is computed for each step.
- This corresponding effect size value is calculated as an R-squared from the OLR that can be applied to both binary and ordinal items.
- As Zumbo notes, because the R-squared is computed for the OLR, it has properties akin to the R-squared measures we commonly find in ordinary least-squares regression.
- Using these R-squared values, the magnitude of DIF can be computed by subtracting the R-squared value for the first step from that of the third step.

82

82

Statistical Modeling: Using Effect Size Indices

- Finally, in order to classify an item as displaying DIF, one must consider both the Chi-squared test of DIF and the corresponding effect size measure.
- First, the two degrees of freedom Chi-squared test for DIF (i.e., testing for the gender and interaction effects simultaneously) must have a p-value less than or equal to 0.01.
- Second, the corresponding effect size measure must have a R-squared value of at least 0.035
  - that is, Jodoin and Gierl's (2001) effect size criteria will be used to quantify the magnitude of DIF: R^2 values below 0.035 for negligible DIF, between 0.035 and 0.070 for moderate DIF, and above 0.070 for large DIF.

83

83

Statistical Modeling: Using Effect Size Indices

- Furthermore, if DIF exists for an item, the steps computed in the calculation of DIF using Zumbo's (1999) ordinal logistic regression will be reviewed to determine if the DIF is uniform or non-uniform.
- Uniform DIF is determined by comparing the R-squared values between the second and first steps to measure the unique variation attributable to the group differences over-and-above the conditioning variable.
- If uniform DIF is found, the odds ratio will be used to interpret the direction of the DIF (i.e., are females or males more likely to endorse the item?).
- More precisely, the odds ratio can be used to determine the odds of one group responding higher to an individual item than those in the corresponding group, after matching on overall depressive symptomatology. The process is repeated for non-uniform DIF.

84

84

Statistical Modeling: Using Effect Size Indices

- The reader should note that the DIF test is by definition a two-degree of freedom test that involves an interaction and main effect.
- If one finds that this two-degree of freedom test is statistically significant than she/he may want to follow-up with the single-degree of freedom test to see whether the effect is predominantly uniform DIF.

85

85

Statistical Modeling: Using Effect Size Indices

- This seems like a good point to bring up the matter of how we use effect size information in our DIF decision making.
  - As part of our overall scientific judgement.
  - As part of the statistical hypothesis testing criterion.
- Let me speak to each of these and what might be the implications of their use.

86

86

Statistical Modeling: Using Effect Size Indices

- Effect size methods have historically been used to help researchers cope with the fact that null hypothesis significance testing, on its own, does not tell us about the "practical", "day-to-day", or "clinical" significance of a statistical finding.
- Effect sizes were meant to aid the researcher in deciding whether their statistical hypothesis test result was "worth" discussing further; i.e., giving the scientist and the reader of a research report some sense of the "magnitude of the effect".
- What I am describing here is a "extra-statistical" use of the effect size as a bit of extra-statistical information to help in the "scientific judgement" about the statistical effect – as opposed to the micro-level statistical decision itself, which is usually based on a p-value or a confidence interval.

87

87

Statistical Modeling: Using Effect Size Indices

- **Curiously, there is a growing tendency in the recent DIF literature and DIF practice to actually use the effect size in a subtly, but very important, different way.**
- **As I was describing a few slides ago, some DIF researchers (both practitioners using DIF in day-to-day work and DIF researchers) are now creating what I will call a "blended" statistical test of DIF; a test that incorporates both the p-value and the effect size (with a corresponding criterion) in the decision making step of deciding whether there is or is not DIF.**
- **In short, these new applications of DIF blend what was once "extra-statistical" information that was once tailored for the "scientific judgement" into the "statistical judgement" of whether this is DIF or not.**

88

88

Statistical Modeling: Using Effect Size Indices

- **So, in earlier uses of effect size one would make**
  - **a statistical decision based on a p-value or confidence interval and**
  - **then, after having made the decision to reject or not reject the statistical hypothesis of DIF,**
  - **and once they have decided to reject the statistical hypothesis of no DIF, examine the effect size to see if the statistical conclusion of DIF was worth scientific discussion.**

89

89

Statistical Modeling: Using Effect Size Indices

- **In some current uses of DIF the statistical conclusion is based on both the p-value and the effect size criterion being met.**
- **As I noted earlier there is sometimes two components to the statistical decision of DIF:**
  - In order to classify an item as displaying DIF, one must consider both the Chi-squared test of DIF and the corresponding effect size measure.
    - First, the two degrees of freedom Chi-squared test for DIF (i.e., testing for the gender and interaction effects simultaneously) must have a p-value less than or equal to 0.01.
    - Second, the corresponding effect size measure must have a R-squared value of at least 0.035
- I will call this two-part decision function the “blended decision rule”.

90

90

Statistical Modeling: Using Effect Size Indices

- Important points to keep in mind when interpreting DIF studies or when you read DIF simulation studies are:
  - Traditionally, the use of effect size as a extra-statistical decision making bit of information that informs the scientific conclusion relied on the statistical test to decide on the decision of whether there is DIF present.
  - This "blended" approach is actually a different statistical test and hence any of the Type I error rate, false positives, and statistical power will probably be quite different for the conventional uses of effects as compared to this "blended" use of effect size. In short, the blended method is a new and different method than the conventional approach.

91

91

Statistical Modeling: Using Effect Size Indices

- Also, please note that this "blended" strategy depends entirely on what we set as the cut-off or criterion for deciding on DIF. For example, in the blended case we are discussing herein we set that the corresponding effect size measure must have a R-squared value of at least 0.035. If we had set some other value than 0.035 we would have a different Type I error rate and power of the blended test.
- Furthermore, my preliminary research and analytical results, suggests that these blended tests are actually quite conservative in terms of their Type I error rate and statistical power.
- Mine is not a criticism of these blended methods, per se, because they may actually reflect the true Type I error rates in practice -- given that we probably use a blended approach in a lot of practice now a days.

92

92

Statistical Modeling: Using Effect Size Indices

- I just want us to recognize that the blended test is different than just using a p-value or confidence interval, and that the resultant operating characteristics of this blended test might be quite different than the original p-value on its own.
  - This latter point is obvious because if the effect size did not add any information to the statistical decision then why would we use it at all.
  - Also, the conservative nature of the blended test is also obvious because we will often set conservative cut-offs for the effect size to counter-act the statistical power of the p-value; which is why we included the effect size in the first place.

93

93

Now, back to our example: *Results from Ordinal Logistic Regression*

- **We will provide the detailed computations for one of the items as a model for how to conduct DIF analyses.**
- **After conducting ordinal logistic regression on item 17: *I had crying spells*, the following model-fitting Chi-squares and Nagelkerke R-squares were found:**

  Step #1: Chi square = 190.63, df=1, R-squared=.338

  Step #2: Chi square = 251.20, df=2, R-squared=.429

  Step #3: Chi square = 254.77, df=3, R-squared=.434

94

94

Results from Ordinal Logistic Regression

- **This resulted in a two-degree freedom Chi-Square test (Step 3 – Step 1) of 64.14 with a p-value of 0.0001 and a Nagelkerke R-squared (Step 3 – Step 1) of .096. Consequently, item 17 shows DIF with a large effect size.**
- **Item 20: *I could not get going,* shows no DIF and a trivial to small effect size: model-fitting Chi-square = 0.25, df=2, p = 882, and Nagelkerke R-Squared = 0.0001.**

95

95

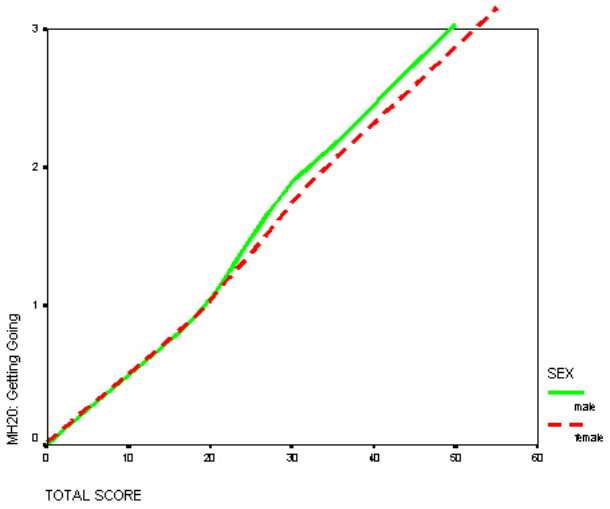
Graphical Representation of DIF

- **The approach we will present is a non-parametric regression graphical visualization tool for DIF. The graphical representation of DIF is based on examining the relationship between total score and item responses for each group separately but simultaneously on the same graph.**
- **A graphical display of items 20 and 17 are depicted in Figures 2 and 3. Note that (a) with reference to item response theory, each of these lines is, in essence, an item response function, and (b) we are looking for difference between the lines.**

96

96

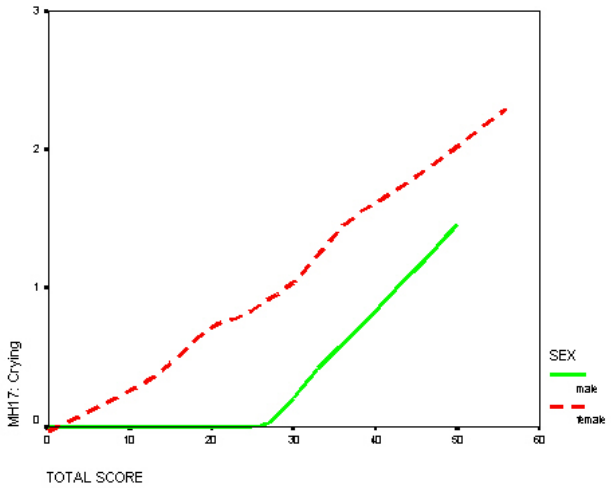
Figure 2: Graphic Model of no DIF: Item 20



97

97

Figure 3. Graphic Model of DIF: Item 17



98

98

Graphical Representation of DIF

- **As can be seen by Figure 2, there are no signs of DIF, indicating that item 20 is performing the same for males and females.**
- **As for item 17 in Figure 3, there are signs of DIF, which can be seen by the large area between the curves.**
  - **This is indicating that item 17 is performing differently for males and females. In fact, males are not endorsing this item until one has a total scale score of 30.**

99

99

Graphical Representation of DIF

- **As Zumbo (2007a) notes, unlike the regression modeling approach, here we are not "covarying" out the total score.**
- **This graphical IRT approach is focusing on the area between the curves (in this case, the LOWESS curves) of the two groups. This is conceptually different than the logistic regression approach in that we are utilizing area between curves in determining DIF.**
- **This method does not technically match on the conditioning variable (total score).**

100

100

Statistical modeling: Conclusions

- The purpose of DIF is to rule out item bias.
  - Item 20 showed no DIF, so there is no item bias. Item 17, however, did show DIF. Therefore; this item is displaying either bias or impact.
  - By examining the R-squared results in Steps 1 through 3, one can see that the R-squared increases markedly from Steps 1 to 2, but far less so for Steps 2 to 3 in the ordinal logistic section above.
  - This is interpreted to mean that the observed DIF is mostly uniform DIF.
  - The presence of uniform DIF is supported by examining Figure 3, wherein the item response lines appear parallel for total scores greater than 30 – at less than 30, males are not endorsing the item so parallelism is not a question.
  - This item is an interesting case of DIF because of the low endorsement at the low end of the scale.

Statistical modeling : Conclusions

- In the end, it should be noted, as Zumbo (2007a) reminds us, that one of the limitations of DIF analyses is that we are investigating differences between non-randomized groups. Like all such analyses, we therefore need to be cautious of interpretations of what is the source or “cause” of DIF.
- At this point all we know is that we have flagged item 17 as DIF, and Figure 3 shows us that males and females with the same level of depressive symptomatology (total score) may have different thresholds and hence endorse the item differently.
- In other words, is it that the difference lies in the reporting of depression (men and women report depression differently – item bias) or is it that women are truly more depressed than men (item impact)?

Statistical modeling : Conclusions

- The graphical nonparametric approach to DIF is an important complement to the statistical modeling because, like all exploratory and data based nonparametric methods, it provides the analyst with a more robust view of the statistical problem and helps validate some of the assumptions -- such as, in our case, the monotonic increasing relationship between the total score and the item response.
- The nonparametric regression strategy (in this case, LOWESS smoothing) is robust in the same sense that the median is a more robust estimate of the center of a distribution as compared to the mean. Overall, the two methods together provide complementary evidence as to the presence of DIF.

103

103

Statistical modeling : Conclusions

- **In closing, two methodological notes are important.**
  - **First, instead of matching on total score, one could match on factor scores (i.e., latent variable score). The advantage of this is that measurement error is less influential with a latent variable.**
  - **Second, for short scales, one may want to match on a 'rest score', which is the total score excluding the item under investigation.**
    - **Each item would then, of course, have a different total score. This is called item-matching purification in the measurement literature. Finally, it is important to note that DIF analyses are conceptually different than scale-level analyses involving multi-group factor analysis. As Zumbo (2003) showed, scale level analyses ignore item-level DIF.**

104

104

Agenda

Done:

1. **What is measurement invariance, DIF, and scale-level invariance?**
2. **Construct versus item or scale equivalence and the three generations of DIF**
3. **Description of DIF methods**
4. **DIF for Ordinal / Likert / Rating Scale Items**

To Do:

5. Scale Level Effect of DIF: Graphical Method

6. Recommendations

105

105

Scale Level Effect of DIF: Graphical Method

- I will describe a new method for investigating/displaying scale-level (or test-level) of the impact of item level concerns (i.e.,DIF).
- I am developing a graphical method that parallels a technique used in nonparametric item response modeling
  - so along the way I will provide a bit of information on nonparametric IRT methods.
- Lets contextualize this material within an example. So far we have seen language test data, achievement test data, psychological variables like depression
  - so lets now turn to a cross-national study of subjective wellbeing measure.

106

106

Scale Level Effect of DIF: Graphical Method

Nonparametric IRT Differential Item Functioning and Differential Test Functioning (DIF/DTF) analysis of the Diener Subjective Well-being scale

- Cross-national DIF;
  - China is group labelled “1” (N=537), and
  - USA is group labelled “2” (N=438).
- Satisfaction with Life Scale (SWLS)
- The SWLS is a short, 5-item instrument designed to measure global cognitive judgments of one's lives. The scale usually requires only about one minute of respondent time. The scale is not copyrighted, and can be used without charge and without permission by all professionals (researchers and practitioners). The scale takes about one minute to complete, and is in the public domain. A description of psychometric properties of the scale can be found in Pavot and Diener, 1993 Psychological Assessment

107

107

Scale Level Effect of DIF: Graphical Method

The SWLS Survey Form (composite total score of 5 items)

Below are five statements that you may agree or disgree with. Using the 1 - 7 scale below indicate your agreement with each item by placing the appropriate number on the line preceding that item. Please be open and honest in your responding.

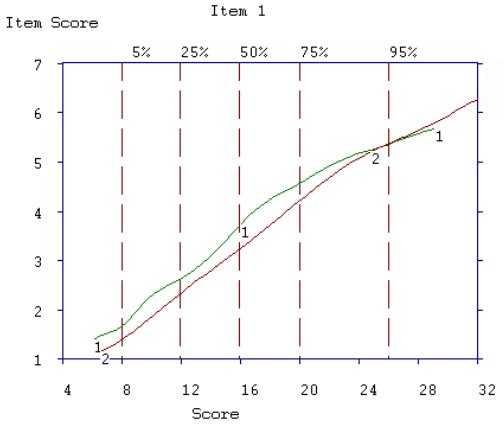
7 - Strongly agree 6 – Agree … 2 – Disagre 1 - Strongly disagree

1. In most ways my life is close to my ideal.
2. The conditions of my life are excellent.
3. I am satisfied with my life.
4. So far I have gotten the important things I want in life.
5. If I could live my life over, I would change almost nothing

108

108

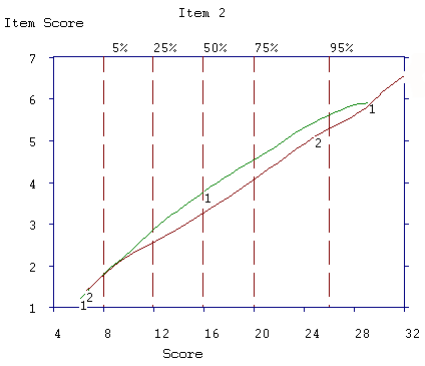
Nonparamteric Item Response Functions (two groups on same graph)

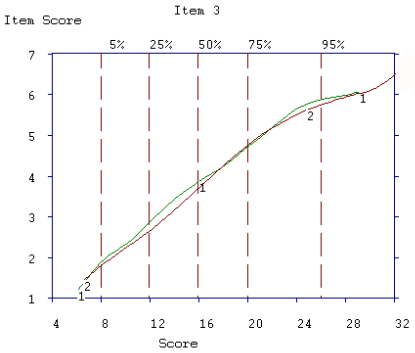


109

109

Nonparamteric Item Response Functions (two groups on same graph)

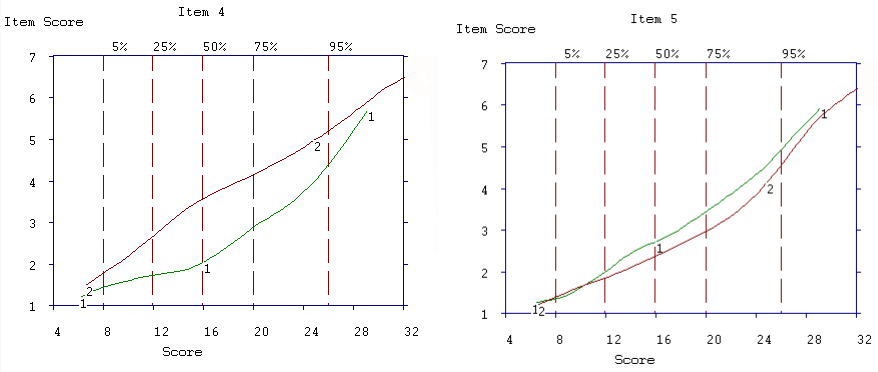




110

110

Nonparametric IRT Differential Item Functioning and Differential Test Functioning (DIF/DTF)



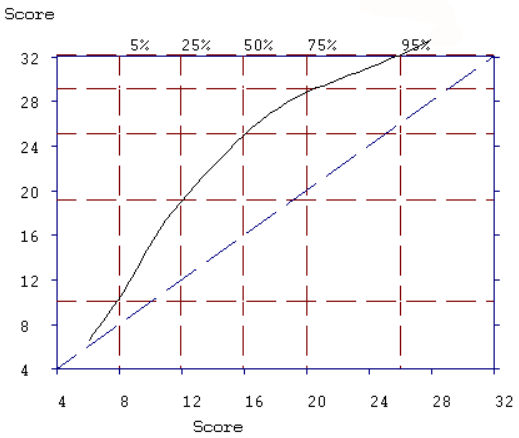
111

111

Nonparametric IRT Differential Item Functioning and Differential Test Functioning (DIF/DTF)

- **Testlevel information: The Differential Test Functioning**
- **USA on the Y-axis, China on the X-axis**

These are "test level" functions" which are composites of the item response functions. But they are two groups on the same plot so you can "trace" the item level effect to the scale level … deviation from 45 degree line.



112

112

Scale Level Effect of DIF: Graphical Method

Differential Test Functioning with Ordinal Logistic Regression

- I can create similar item graphs using ordinal logistic regression (these would be analogous graphs for the parametric DIF analysis).
- To create the DIF plot for each item you would save the predicted category score from the model with both the uniform and non-uniform DIF.
- You would then create a scatterplot with the predicted item response on the Y-axis and the total score (or, if you used the item-corrected total score you would use that instead) on the X-axis.
- Recall that you would want to use the grouping variable as a marker variable so that you can get the separate curves for the groups on the same graph. You can then use non-parametric smoothing to trace that item line – or plot the ordinal logistic curve by plotting the function itself from the results.

113

113

Scale Level Effect of DIF: Graphical Method

- **One can then create a Differential Test Function by**
  - **computing the predicted item score for each item separately, as described above,**
  - **computing a total predicted item score, and then**
  - **creating a scatterplot with predicted score on the Y-axis with the smoothed trace lines for the groups separately.**
- **This allows you to see the test-level impact of the item level DIF.**

114

114

Scale Level Effect of DIF: Graphical Method

- **In the graph that follows you can see the test level effect of the item level DIF, the ordinal logistic regression differential test functioning plot (DTF).**
- **You can see that, depending on the scale score level, there is a rather dramatic differential test functioning between Chinese and American respondents.**

115

115

Scale Level Effect of DIF: Graphical Method: **Ordinal Logistic Regression Differential Test Functioning Plot**

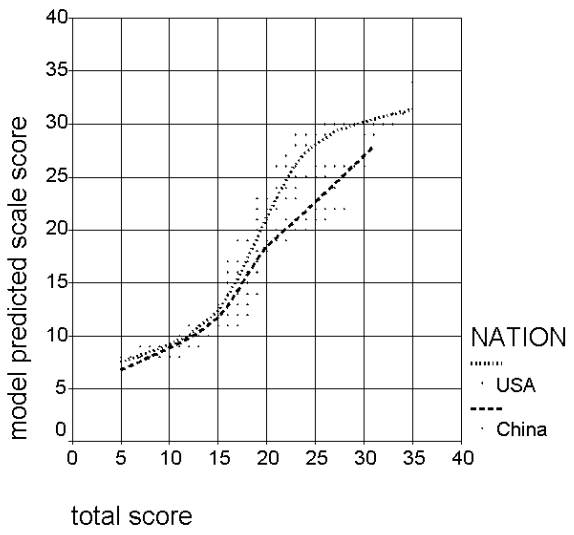
Summary stats of SWLS for the 2 countries:

China:

25th percentile=12,
median=16,

75th perc.=20.

USA:
25th percentile=19,
median=25,
75th perc.=29.



116

116

Scale Level Effect of DIF: Graphical Method

- In comparing the non-parametric DTF plot and the parametric (Ordinal Logistic Regression, OLR) DTF plot, you see similar conclusions.
- This is gratifying and tells me that the effect one sees in the OLR DTF plot represents the data.
- I like both plots but the OLR DTF plot is, by definition, subject to model-data fit in a way that the nonparametric plot is not ... the counter, of course, is that the nonparametric plot may actually over-fit and accentuate small patterns in the data (patterns that are smoothed out in the parametric approach).
- You can see why I like both!
  - Note that these plots are descriptive and do not have a significance test associated with the effects

Agenda

Done:

1. What is measurement invariance, DIF, and scale-level invariance?
2. Construct versus item or scale equivalence and the three generations of DIF
3. Description of DIF methods
4. DIF for Ordinal / Likert / Rating Scale Items
5. Scale Level Effect of DIF: Graphical Method

To Do:

6. Recommendations

If time permits, some new ideas on how to deal with "indicators" for questions of indicator bias or Differential Indicator Functioning.

RECOMMENDATIONS

1. **Select an acceptable DIF detection method.**
2. **For both construct equivalence and DIF studies**

   Replicate
   Physically Match (if possible)

3. **For DIF studies—purify criterion (iterative), if necessary; latent variable matching is also very productive (e.g., MIMIC).**

119

119

RECOMMENDATIONS (continued)

4. **DIF analyses should be followed up by content analyses.**
5. **Translation DIF is usually explainable, others are bit more difficult (in translation items are homologues).**
6. **Use results from Content Analyses and DIF studies to inform future test development and adaptation efforts**

120

120

RECOMMENDATIONS (continued)

7. **Realize that items flagged for DIF may not represent bias**

   --differences in cultural relevance

   --true group proficiency differences

8. **When subject-matter experts review items, they should evaluate both DIF and non-DIF items.**

121

121

RECOMMENDATIONS (continued)

9. **If you have Likert responses you can use ordinal logistic regression.**
   - **We have shown in recent collaborative research that the generalized Mantel is also a promising approach; it however, does not have the advantages of modeling (multiple conditioning variables, etc.).**
10. **Gelin & Zumbo (2003) were the first to show that DIF may depend how an item is scored. Please keep in mind that DIF conclusions are limited to item scoring and hence item usage. You need to check DIF for various scorings and hence various usages of items (e.g., prevalence verus presence scoring of the CES-D).**

122

122

RECOMMENDATIONS (continued)

11. **If the scale is short (less than 10 items) you might run into sparse data matching problems and in this case either multiple variable matching or plausible values matching might be useful.**
12. **When sample sizes are small, use nonparametric IRT (e.g., TestGraf) and the cut-offs provided by Zumbo & Witarsa (2004) for sample sizes that 25 per group.**
    http://www.educ.ubc.ca/faculty/zumbo/aera/papers/2004.html

123

123

Remember:

"As with other aspects of test validation, DIF analysis is a process of collecting evidence. Weighting and interpreting that evidence will require careful judgment.

There is no single correct answer"

(Clauser & Mazor, 1998, p. 40).

124

124

Agenda

Done:

1. **What is measurement invariance, DIF, and scale-level invariance?**
2. **Construct versus item or scale equivalence and the three generations of DIF**
3. **Description of DIF methods**
4. **DIF for Ordinal / Likert / Rating Scale Items**
5. **Scale Level Effect of DIF: Graphical Method**
6. **Recommendations**

To Do:

**** If time permits, some new ideas on how to deal with "indicators" for questions of indicator bias or Differential Indicator Functioning.**

125

125

Now for something that is a bit different and relevant to indicators and indices

- **As promised at the beginning, if time permits, I will discuss some new ideas at the end about how to extend this work to "indicators" using a structural equation modeling framework.**
- **We here are a few ideas.**

126

126

Structural Equation Modeling Approaches

		Matching variable	
		Observed score	*Latent variable*
Item format	*Binary*	MH LogR	Conditional IRT Multidimensional SEM
	Polytomous	Ordinal LogR	Conditional IRT Multidimensional SEM

127

127

Structural Equation Modeling Approaches

- **We are interested in short scales (e.g., less than 25 items) as often found in psychological research.**
- **We are interested in using the latent variable approach because we want to condition on the "measurement error free" latent variable. This is particularly important with short scales (i.e., more items higher reliability and hence less measurement error).**
- **IRT is not feasible for us because it requires lots of items to estimate the latent variable (theta) score and we want to focus on "short" scales. SIBTEST also requires more items than we have in our target scale.**
- **This leaves us the SEM approach to DIF.**

128

128

Structural Equation Modeling Approaches

- **Unlike the observed score framework that conditions on the observed scale score, latent variable models (IRT or SEM) make use of the joint distribution of the items and hence one needs to model method effects, or the latent variable model will give incorrect standard errors.**
- **SEM DIF was first proposed by Muthén (1989).**
- **When dealing with "indicators" we adapt Muthén's approach to allow us to investigate a "principal components" or formative model for indicators.**
  - **But, first, the more conventional "reflective" model.**

129

129

Structural Equation Modeling Approaches

- **The SEM DIF model is, in essence, a multiple-indicators, multiple-causes (MIMIC) model – akin to a latent variable ANCOVA.**
- **Given that the grouping variable (e.g., gender) is observed and the item responses are ordered categories, we need to write a model that jointly estimates the quantities correctly – it may be some kind of conditional model.**
- **This estimation method allows one to compute the joint covariance matrix of the predictor and the variables underlying each of the ordinal variables – akin to a polychoric correlation matrix. This matrix can be used as input for SEM, and estimation can be correctly applied.**

130

130

Structural Equation Modeling Approaches

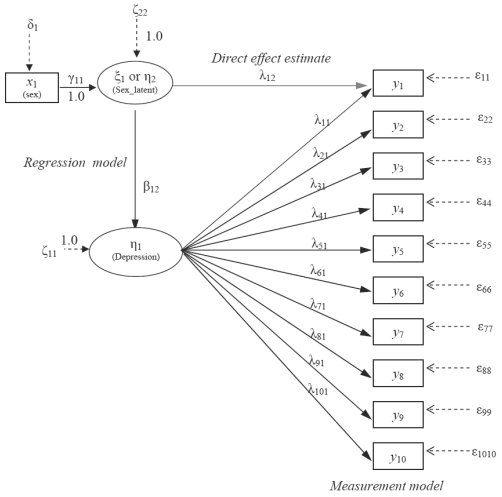
The DIF MIMIC model consists of three components:

1. Measurement model.
   - **Relates the observed indicators to the continuous latent variable η_1 (e.g., depression)**
   - **Can be expressed as** $y = \nu + \Lambda\eta + \varepsilon$
2. Regression model
   - **Relates the latent variable to the latent covariate (e.g., gender)**
   - **Can be expressed as** $\eta = \alpha + \beta\eta + \Gamma x + \zeta$
3. Direct effects model.
   - **Detects measurement invariance in an item response associated with group membership**

131

131

Structural Equation Modeling Approaches

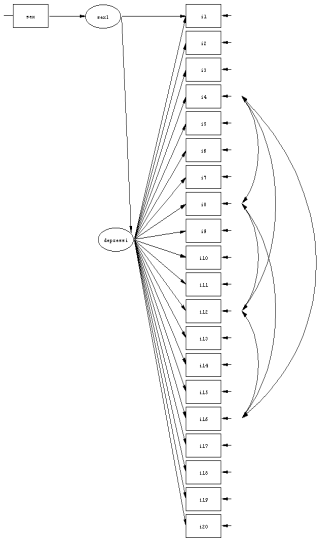


132

132

Structural Equation Modeling Approaches

- **On can even model complex item response dependencies to test for complex DIF -- e.g., via the MIMIC model with method effects modeled for the 4 positively worded items of the CES-D.**



133

133

Structural Equation Modeling Approaches

- Ongoing research is investigating the operating characteristics of the DIF MIMIC approach estimated via Jöreskog's recently proposed methods and also via Muthen's methods.

Advantages of the DIF MIMIC approach:

- Error-free covariate with short scales
- Different item response formats: (a) binary, (b) ordinal, and (c) mixed item formats
- Multidimensional and complex scales
- Multilevel or hierarchical structured data

134

134

Structural Equation Modeling Approaches

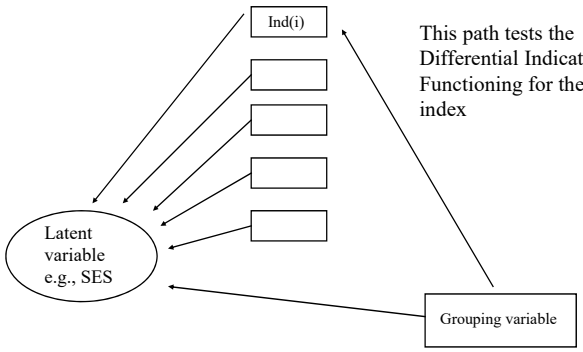
- **To deal with the "indicators" approach then model a "formative" SEM model wherein the "arrows" go from the indicators (items or measures) to the latent variable in the "measurement model" part, and then use the MIMIC approach.**
- **This should work … we need to test it!**

η_1

135

135

New method for differential indicators analysis



Example of an index with 5 manifest indicators; investigate whether, for example, gender differentially effects the system. Is the system gender fair?

Note: We could also use a latent class method for paths and latent variance, with predictors of the latent class membership.

136

136

A peek into (my) future

- My excitement these days is as high as ever about/for DIF. Here are some of the projects I am vigorously pursuing.
  - Latent class modeling with latent classes linked to the thresholds, loadings, and latent variable variance, and simultaneously model predictors of (latent) class membership as an inductive explanatory DIF modeling strategies.
  - The use of SEM approaches to develop a class of DIF-like methods for indicators.
  - I have proposed 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 non-uniform 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.

137

137

A peek into (my) future

- DIF when there is no internal matching variable. The situation of no internal matching variables (or short scales with only 1 or 2 scale/test items) can occur in (a) in widely used operational testing contexts with performance assessments, or (b) with single-item measures in social, health or educational surveys– emerging from a sociological and sample survey tradition rather than multi-item measures in psychometrics.
  - There are two strategies for resolving this issue.
  1) A variation on a multiple matching variables strategy;
  2) The statistical science literature (primarily in biostatistics) has shown that in some settings propensity scored matching methods perform better than multiple matching (also called covariance matching or analysis of covariance).

  Multiple matching variables (multiple covariates) is computationally easier to implement in DIF analyses because the propensity score matching requires us to resolve an issue of the complex data structure from the resultant match samples; however, propensity score matching may be more effective and powerful. Although promising, this needs study to examine implementation and computational feasibility.
- **…. and on, and on, and on …** maybe some collaborative work with be initiated with the closing of today' session.

138

138

References

Breithaupt, K., & Zumbo, B. D. (2002). Sample invariance of the structural equation model and the item response model: a case study. *Structural Equation Modeling, 9*, 390-412.

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

Gelin, M. N., & Zumbo, B. D. (2003). DIF results may change depending on how an item is scored: An illustration with the Center for Epidemiological Studies Depression (CES-D) scale. *Educational and Psychological Measurement, 63*, 65-74.

Hubley, A. M., & Zumbo, B. D. (1996). A dialectic on validity: Where we have been and where we are going. *The Journal of General Psychology, 123,* 207-215.

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.

Lu, I. R. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT In Structural Equation Models: A Comparison With Regression Based On IRT Scores. *Structural Equation Modeling*, *12*, 263-277.

Maller, S. J., French. B. F., & Zumbo, B. D. (2007). Item and Test Bias. In Neil J. Salkind (Ed.), *Encyclopedia of Measurement and Statistics* (Vol. II, pp. 489-493). Thousand Oaks, CA: Sage Publications..

Rupp, A. A., & Zumbo, B. D. (2003). Which Model is Best? Robustness Properties to Justify Model Choice among Unidimensional IRT Models under Item Parameter Drift. {Theme issue in honor of Ross Traub} *Alberta Journal of Educational Research, 49*, 264-276.

Rupp, A. A., & Zumbo, B. D. (2004).. A Note on How to Quantify and Report Whether Invariance Holds for IRT Models: When Pearson Correlations Are Not Enough. *Educational and Psychological Measurement, 64*, 588-599. {Errata, (2004) *Educational and Psychological Measurement, 64,* 991}

Rupp, A. A., & Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement, 66,* 63-84.

139

References

Thomas, D. R., Zhu, P., Zumbo, B. D., & Dutta, S. (2008). On Measuring the Relative Importance of Explanatory Variables in a Logistic Regression. *Journal of Modern Applied Statistical Methods, 7, 21-38*.

Zumbo, B.D. (2007a). 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. (2007b). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics*, (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Zumbo, B. D. (2005). Structural Equation Modeling and Test Validation. In Brian Everitt and David C. Howell, *Encyclopedia of Behavioral Statistics,* (pp. 1951-1958). Chichester, UK: John Wiley & Sons Ltd.

Zumbo, B. D. (Ed.) (1998). *Validity Theory and the Methods Used in Validation: Perspectives from the Social and Behavioral Sciences*. Netherlands: Kluwer Academic Press. {This is a special issue of the journal *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement, Volume 45, No. 1-3, 509 pages*}

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.

140

References

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., & Hubley, A. M. (2003). Item Bias. In Rocío Fernández-Ballesteros (Ed.). *Encyclopedia of Psychological Assessment*, pp. 505-509. Sage Press, Thousand Oaks, CA.

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.

Zumbo, B. D., & Rupp, A. A. (2004). Responsible Modeling of Measurement Data For Appropriate Inferences: Important Advances in Reliability and Validity Theory. In David Kaplan (Ed.), *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press.

Zumbo, B. D., Sireci, S. G., & Hambleton, R. K. (2003). *Re-Visiting Exploratory Methods for Construct Comparability and Measurement Invariance: Is There Something to be Gained From the Ways of Old?* Annual Meeting of the National Council for Measurement in Education (NCME), Chicago, Illinois.

Zumbo, B. D., & Witarsa, P.M. (2004). *Nonparametric IRT Methodology For Detecting DIF In Moderate-To-Small Scale Measurement: Operating Characteristics And A Comparison With The Mantel Haenszel*. Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.

141

141

Thank You For Your Attention!

Professor Bruno D. Zumbo

University of British Columbia

bruno.zumbo@ubc.ca

http://www.educ.ubc.ca/faculty/zumbo/zumbo.htm

142

142