Nonparametric IRT Methodology For Detecting DIF In Moderate-To-Small Scale Measurement:

Operating Characteristics And A Comparison With The Mantel Haenszel

Bruno D. Zumbo

Petronilla Murlita Witarsa

The University of British Columbia

Introduction to the Problem

- Introduction to the problem.
  - In DIF work focus has been on large-scale wherein there are lots items and lots of examinees.
  - We are focusing on the sort of measurement work done in educational Psychology research, small or non-repeating surveys, pilot studies, and some large college classes (e.g., intro psych).

Why an IRT DIF method?

- Two broad classes of DIF detection methods
  - Modeling contingency tables or modeling logistic regression models
  - IRT methods
- The essential difference is the "what" and "how" the matching or conditioning is performed.
  - 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.
  - Comparing the IRT parameter estimates or IRFs [item response functions] 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 IRFs across the distribution of the continuum of variation, theta.

Why a nonparametric IRT method?

- Two reasons:
  - The interest on relatively small sample sizes and relatively few items in the scale or measure, made it so that we could not use most conventional parametric IRT models.
  - We also wanted an approach that has a very "exploratory data analysis" data driven orientation because we had no reason to believe that the item response functions would be simple parametric functions -- such as a 1-parameter or Rasch model, which is sometimes recommended for moderate-to-small-scale testign.
- Hence, why we used Ramsay's nonparametric IRT method.

Hypothesis Testing IRT DIF

- We describe a statistic (beta) based on nonparametric item response theory as well as:
  - a formal hypothesis test of DIF based on the assumed sampling distribution of beta, and
  - a hypothesis testing strategy that does not use a sampling distribution, per se, but rather a cut-off value for testing for DIF (Roussos & Stout criterion).

Hypothesis Testing IRT DIF

- With our knowledge of beta there are two approaches to testing the "no DIF" hypothesis.
  - Perform a formal hypothesis test making use of the purported sampling distribution of beta (never been studied).
  - A less formal hypothesis tes: Compute beta and compare its value to a criterion (not making use of the sampling distribution of beta); Roussos and Stout (1996) proposed the following cut-off indices: (a) negligible DIF if $|\beta| < .059$, (b) moderate DIF if $.059 \leq |\beta| < .088$, and (c) large DIF if $|\beta| \geq .088$. Gotzmann (2002) recently investigated the use of these cut-off indices with large-scale testing (sample sizes of 500 or greater per group) and found that these cut-offs result in Type I error rates less than or equal to 5%. In using the Roussos-Stout cut-offs, Gotzmann declared an item as displaying DIF if the $|\beta|$ was greater than or equal to 0.059.

Research Questions

- Little to nothing is known about the performance of either of the hypothesis testing strategies in moderate-to-small-scale testing contexts.
- We conducted two simulation studies in the context of moderate-to-small-scale testing:
  - Study 1 was aimed at studying (a) the properties of the sampling distribution of the beta statistic under the null hypothesis of no DIF, (b) the Type I error rate of using the Roussos-Stout cut-off value, and (c) to allow comparison to a known method, we also studied the Mantel Haenszel DIF detection method.
  - Study 2. Based on the results of Study 1, a simulation study was conducted to compare the statistical power of the methods for which the Type I error rate as maintained at nominal levels.

Study 1: Methodology

- Used a similar methodology as that used by Muniz, Hambleton, and Xing (2001).
- The following variables were manipulated in the simulation study:
  - Sample sizes. 500/500, 200/100, 200/50, 100/100, 100/50, 50/50, 50/25, and 25/25 examinees in pairs, respectively. Five of the above combinations were the same with that used in the study by Muniz et al. The additional sample size combinations, 200/100, 50/25, and 25/25 were included so that an intermediary between 500/500 and 200/50, and smaller sample size combinations were included. In addition, as Muniz et al. suggested, these sample size combinations reflect the range of sample sizes seen in practice in, what we would refer to as, moderate-to-small-scale testing.

Study 1: Methodology

Statistical characteristics of the studied test items.

- We simulated a 40 (binary) item test using a 3-parameter (parametric) IRT model.
- The item parameters for the first 34 items came from the 1999 TIMSS math test for grade eight. Descriptive statistics for these items are:

  discrimination: mean=.95, range= .42 – 1.59

  difficulty: mean=-.03, range= -1.91 – 1.13

  guessing: mean= .23, range= .06 - .43

Study 1: Methodology

- The last six items were the items for which DIF was investigated – i.e., the studied DIF items. The *a* refers to the item discrimination parameter, *b* the item difficulty parameter, and *c* the pseudo-guessing parameter.
- The following table (next slide) lists the values of item parameters for the six studied items.

Study 1: Methodology

Item #	*a*	*b*	*c*
35	0.50	-1.00	.17
36	1.00	-1.00	.17
37	0.50	0.00	.17
38	1.00	0.00	.17
39	0.50	1.00	.17
40	1.00	1.00	.17

As in Muniz, Hambleton, and Xing (2001)

two levels of the a-parameter (0.5 and 1)

three levels of the b-parameter (-1, 0, 1)

the c-parameter is constant at 0.17

Study 1: Methodology

- Therefore, there were three factors varied in the simulation: (a) sample size combination, (b) item difficulty, and (c) item discrimination.
- The design was an 8x3x2 completely crossed design.
- 100 replications in each cell of the design.
- For each replication, for each of the six studied items the dependent variables were: (a) TestGraf's beta, (b) TestGraf's standard error of beta, and (c) the Mantel Haenszel statistic.

Study 1: Results

- Because there were a lot of results I will summarize them below.

  Formal Hypothesis Testing of DIF via TestGraf Beta:

  – The sampling distribution of beta is, as postulated, Gaussian the beta produced by TestGraf is an unbiased estimate of the population beta even at small sample sizes, -- i.e., the mean of the sampling distribution is the population value of zero under the null distribution of no DIF.

  – However, the standard error of beta produced by TestGraf was smaller than it should be. Of course, this under-estimated standard error resulted in the Type I error rate of the statistical test of beta (described as the formal statistical test above) is substantially inflated above the nominal levels.

Study 1: Results

Less formal test of DIF using Beta and a cut-off

- The Roussos-Stout cut-off of |beta| < .059 for detecting DIF resulted in error rates, under the null hypothesis, as high as .37.
- That is, under the simulated condition of no DIF, this cut-off approach would lead the researcher to declare that there is at least moderate DIF 37% of the time if the sample size was less than 500 per group.
- For 500 examinees per group, the Roussos-Stout cut-off of |β| < .059 resulted in acceptable Type I error rates ranging from zero to three percent depending on the item's discrimination and difficulty.

Study 1: Results

The MH test of DIF

- The Mantel-Haeszel DIF test maintained its Type I error rate at or below the nominal rate.

What to do given the results?

- Because neither the Roussos-Stout cut-off nor the formal hypothesis test of beta maintained their Type I error rates, it seemed natural to compute new cut-offs for beta (in the context of moderate-to-small scale testing we simulated) based on the 90^{th}, 95^{th}, and 99^{th} percentiles of the null DIF distribution of beta.
- These new cut-offs may replace the Roussos-Stout values for moderate-to-small scale testing, and particularly when one has less than 500 examinees per group.

Study 1: Results

Cut-off indices for β in identifying TestGraf DIF across sample size combinations and three significance levels irrespective of the item characteristics

	Level of Significance α		
N_1 / N_2	.10	.05	.01
500/500	.0113	.0161	.0374
200/100	.0249	.0373	.0415
200/50	.0460	.0540	.0568
100/100	.0308	.0421	.0690
100/50	.0421	.0579	.0741
50/50	.0399	.0455	.0626
50/25	.0633	.0869	.1371
25/25	.0770	.0890	.1154

Study 2: Comparative Statistical Power

- Based on the results of the first study, an additional simulation study was conducted to investigate the statistical power of the DIF tests that maintain their Type I error rate at, or below, nominal levels. The simulation design for the power component was the same as the Type I error rate except for non-zero population DIF.
- A comparison was made of the statistical power of the (a) Mantel-Haenszel, and (b) informal test of TestGraf's beta against the new cut-offs from Study 1.

Study 2: Methodology

- The design is the same as for Study, an 8x3x2 (sample size, item difficulty, item discrimination).
- In order to study the statistical power, the non-zero DIF was introduced through *b* value differences in the DIF items between the two test sets for generating the reference and focal population groups.
- Three levels of *b* value differences were applied: 0.5 for small DIF, 1.0 for medium DIF, and 1.5 for large DIF. These are standard values seen in the literature.

Study 2: Results

- In all cases the statistical power of the TestGraf beta, using our new cut-offs appropriate for moderate-to-small-scale testing, is substantially higher.
- The power superiority of the TestGraf beta is most noteworthy for the smaller sample sizes, and for small DIF effect size (differences in power ranging from .2 to .5 … this is quite noteworthy).

Overall Summary

- In the first simulation study we investigate TestGraf beta's standard error and operating characteristics. We found that although the beta DIF statistic produced by the nonparametric IRT software TestGraf is unbiased, the standard error of that statistic is negatively biased resulting in an inflated Type I error rate.
- Likewise, the Roussos-Stout cut-offs for beta produced inflated Type I error rates. Given that the formal test and Roussos-Stout's cut-offs resulted in an inflated Type I error rate, new cut-offs values are proposed based on our simulation results.
- In the second study statistical power from using these new cut-off values for beta are compared to the power of using the Mantel Haenszel (MH) DIF statistic. We found that the procedure based on the new cut-off values had substantially more power than the MH.