

Detecting and Interpreting Uniform and Non-Uniform Differential Item Functioning in a Multidimensional Test

Damon U. Bryant and Dahlia S. Forde  
University of Central Florida

IOOB

Graduate Student Conference

March, 2002

# Introduction

## Purpose

There are several objectives of this investigation

- 1) To demonstrate how logistic regression can be used to detect and interpret Differential Item Functioning (**DIF**) in tests,
- 2) To show how non-uniform DIF influences the reliability estimates of scores (internal consistency), and
- 3) To show how logistic regression can be used to detect DIF in a multidimensional test.

# Introduction

- Tests are often used to make important decisions about selection, training, and promotion (Muchinsky, 1994; Vincent, 1996).
- Federal laws require that certain groups (e.g., females, minorities, or disabled persons) be treated fairly when selection or promotion decisions are made on the basis of test performance (McAllister, 1993).
- Organizations are often unaware of test bias, which is illegal according to the Civil Rights Act of 1964.
- Leads to costly litigation, judgments, or settlements for organizations, public and private.

# Introduction

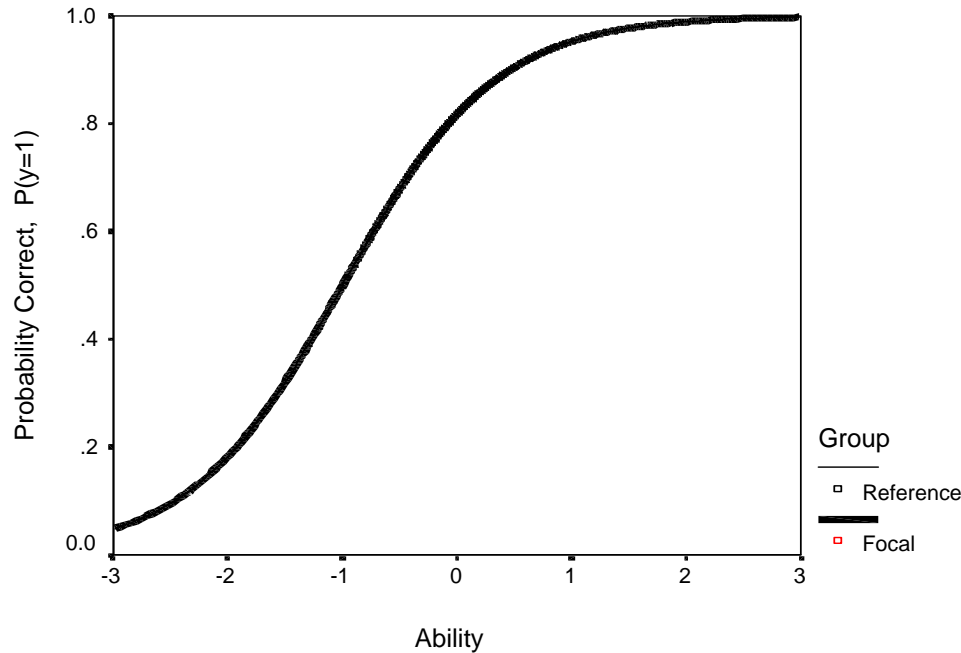
Test bias is now generally determined by evaluating the context in which the test is used, the content, and the differential functioning of items (Cole & Moss, 1989; Goldstein, 1996).

Differential item functioning (DIF) is the term commonly used to describe items that “ 'function differently' for two or more groups if the probability of a correct answer to a test is associated with group membership for examinees of comparable ability” (Camilli, 1993, p. 397-398).

# Introduction

Item that exhibits **No DIF**

(Ideal Situation)

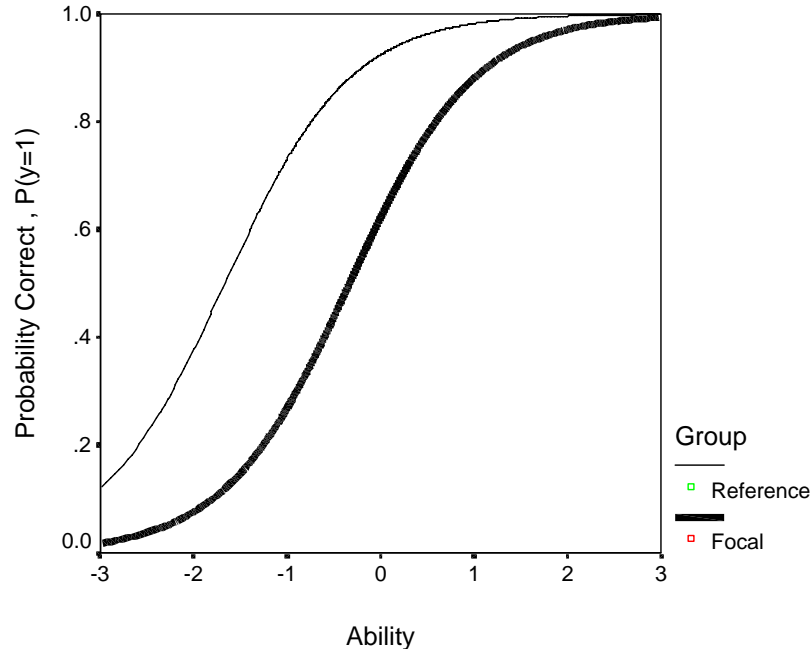


Reference group is the one to which all groups are compared (e.g., non-minorities or men).

Focal group is the one of interest to the investigator (e.g., minorities or women).

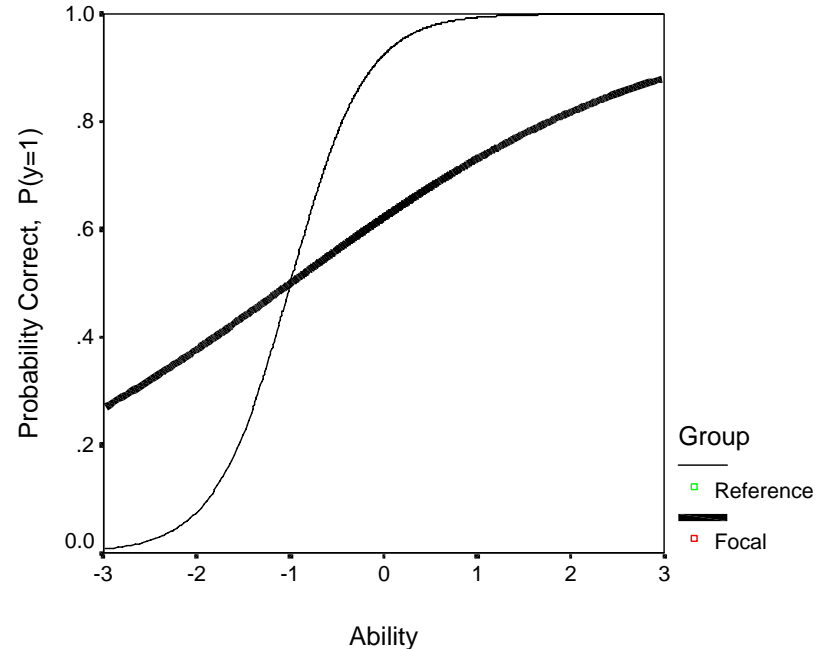
# Introduction

Two types of DIF, uniform and non-uniform (Mellenbergh, 1982)



## Uniform DIF

Significant difference in difficulty of an item between groups



## Non-Uniform DIF

Significant difference in psychometrically desirable item discrimination

# Introduction

**Item Response Theory** (Hambleton & Swaminathan, 1985; Lord, 1980; Raju, van der Linden, & Fler, 1995). Estimates difficulty, discrimination, and guessing parameters.

*Disadvantage:* Requires extremely large samples and has no tests of statistical significance

**Mantel-Haenszel Statistic** (Holland & Thayer, 1988; Mantel & Haenszel, 1959). Produces a common odds ratio across levels of the stratified variable of interest. Very useful in small samples.

*Disadvantage:* Unable to detect non-uniform DIF

**Logistic Regression** (Swaminathan & Rogers, 1990) Able to detect uniform and non-uniform DIF in large and small samples with associated significance test.

# Introduction

Simulation studies suggest that logistic regression is more effective in detecting non-uniform DIF (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990; Whitmore & Schumacker, 1999)

Whitmore and Schumacker (1999) have shown that under different conditions of sample size, item difficulty, test length, and varying underlying ability for groups, logistic regression was as effective as IRT-based ANOVA in detecting uniform DIF and outperformed ANOVA in detecting non-uniform DIF.

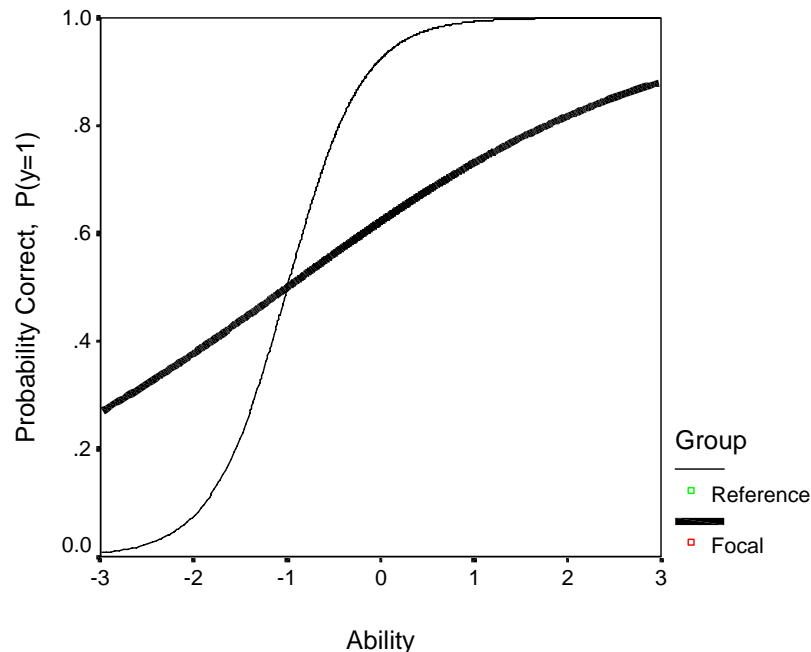
Logistic regression is as powerful as the Mantel-Haenszel procedure in detecting uniform DIF and more efficient than the Mantel-Haenszel method in detecting non-uniform DIF (Rogers and Swaminathan, 1993).

# Introduction

The extent to which logistic regression detects non-uniform DIF in non-simulated data isn't well established due to modeling problems.

**Hypothesis 1:** Logistic regression detects non-uniform differential item functioning in non-simulated data.

**Hypothesis 2:** Logistic regression detects more non-uniform DIF than the Mantel-Haenszel statistic.



# Introduction

**Item discrimination** is related to **reliability** of scores on a test through its contribution of systematic variance (Ebel, 1967; Nunnally & Bernstein, 1994).

The **reliability** of scores is related to **criterion-related validity**; this relationship is established by the formula for the correction of attenuation due to unreliability in measurement (Nunnally & Bernstein, 1994).

Therefore, **item discrimination** is related to **criterion-related validity**.

If the aforementioned is true, then non-uniform DIF, by definition, is related to the reliability of scores and criterion-related validity.

**Hypothesis 3:** Non-uniform DIF creates a difference in reliability estimates between groups for those items that have significantly better discriminations for one group relative to the other.

In other words, the reliability estimate for the non-uniform DIF items should be higher for the group with greater item discriminations as compared to the group with lower item discriminations.

# Introduction

Some researchers argue that items exhibit DIF because the test is multidimensional, i.e., it measures more than one ability (Ackerman, 1994).

Research by Mazor, Kanjee, and Clauser (1995) indicated that when a nuisance dimension, e.g., verbal ability, was controlled using logistic regression, fewer items exhibited uniform DIF in an achievement tests of chemistry and history.

Evidence from another investigation by Clauser, Nungester, Mazor, and Ripley (1996) provides support for the notion that matching on several relevant abilities simultaneously resulted in fewer items identified as exhibiting uniform DIF in a medical school certification exam as compared to the Mantel-Haenszel statistic.

Both investigations, however, did not evaluate the extent to which non-uniform DIF was present in items measuring more than one ability.

**Hypothesis 4:** Logistic regression is able to detect uniform and non-uniform DIF in a multidimensional test.

## **Method**

### *Participants*

- Three hundred and eighty-five (385) undergraduate psychology students at an urban university in the United States of America
- Two hundred and thirty (230) identified themselves as female, and 155 identified themselves as male.
- Participants were given extra credit toward their course grade.

## Method

### *Measure*

**The Work Skills Test** (Prien, Wooten, & Prien, 2000).

- Measures the ability to discriminate among objects and to understand and apply mechanical relationships and physical laws in practical situations.
- Contains sixty-five (65) multiple-choice items that are dichotomously scored: 1 for correct and 0 for incorrect
- High scores are indicative of high mechanical ability, and low scores are indicative of low mechanical ability
- Items drawn from both male and female experiences
- Reliability estimate using Kuder-Richardson formula 20 (KR-20) is .77.

## Method

### *Procedures*

### **Phase I**

- Detect DIF in a test assumed to be one dimensional (Swaminathan & Rogers, 1990).
- Compare items to those detected using the Mantel-Haenszel statistic.

### *Logistic Regression*

$$\pi(y = 1 | g, \theta) = \frac{e^z}{1 + e^z}$$

$$z = \ln [\text{odds}(g, \theta)] = \beta_0 + \beta_1 g + \beta_2 \theta + \beta_3 (g * \theta)$$

$g$  = Group (Effects coded variable)

DV =  $\pi(y = 1)$

$\theta$  = Ability (Total score in z-score form)

$g * \theta$  = Group by Ability interaction

## Method

### *Procedures*

### **Phase II**

- If DIF is detected
- Conduct non-uniform DIF reliability analyses (Spearman-Brown)
- Conduct principal components analysis to determine if test is multidimensional and confirm multidimensionality with DIMTEST procedures (Nandakumar & Stout, 1993).
- If test is multidimensional, detect DIF extending methods presented by Swaminathan and Rogers (1990), which is consistent with Cattell (1960)



# Results

## Non-Uniform DIF Reliability Analyses

Derivation of Spearman-Brown Prophecy formula (Spearman, 1910; Brown, 1910)

<u>Items</u>	<u>Overall</u>	<u>Women</u>	<u>Men</u>
2, 6, 55, & 64			
Observed Estimate	.15	.22	.03
Spearman-Brown	.80	.86	.41

### **H3:S**

Note: These items had significantly better item discriminations for women as compared to men.

Reliability estimates for Women and Men are significantly different ( $p < 05$ ).

# Results (Phase II)

## Item Loading on Rotated Components

<u>Item #</u>	<u>1 (MP)</u>	<u>Item #</u>	<u>2 (OD)</u>	<u>Item #</u>	<u>3 (LA)</u>
<b>62</b>	.79	<b>30</b>	.70	<b>21</b>	.56
<b>60</b>	.73	<b>23</b>	.69	<b>5</b>	.48
<b>64</b>	.73	<b>39</b>	.65	<b>51</b>	.47
<b>61</b>	.70	<b>17</b>	.59	<b>18</b>	.45
<b>63</b>	.67	<b>4</b>	.51	<b>26</b>	.41
<b>65</b>	.67	<b>8</b>	.44	<b>3</b>	.40
<b>59</b>	.63	<b>45</b>	.45	<b>9</b>	.38
<b>57</b>	.57	<b>14</b>	.43	<b>2</b>	.37
<b>56</b>	.54	<b>53</b>	.40	<b>13</b>	.36
<b>58</b>	.41			<b>11</b>	.34
				<b>10</b>	.32

# Results

Items loading on the three different dimensions were combined to create three different scores:

Mechanical Planning (MP,  $\theta_1$ ) - 10 items      KR-20 = .83

Object Discrimination (OD,  $\theta_2$ ) - 9 items      KR-20 = .71

Mechanical Application (LA,  $\theta_3$ ) - 11 items      KR-20 = .63

## DIMTEST (Nandakumar & Stout, 1993)

Mechanical Planning:      T = 11.52      p < .001

Object Discrimination:      T = 3.87      p < .001

Mechanical Application:      T = 2.27      p < .01

## Results

### Summary of Items Exhibiting Uniform and Non-uniform DIF in the Multidimensional Test

Logistic Regression with 3 ability estimates:

Uniform DIF  
3, 10,18, & 58

Non-uniform DIF  
2, 20, 60

**H4 : S** (7 items)  
65-item test, KR-20 = **.77**      Validity: BMCT (r = **.67** )

From multidimensional items, developed 25-item test that meets the unidimensionality assumption (Reckase, Ackerman, & Carlson, 1988).      DIMTEST:  $T = 1.06$ ,  $p > .05$

25-item test: KR-20 = **.84**      Validity: BMCT (r = **.62**)  
Increased reliability while reducing test length by **62%**.

# Discussion

This investigation provides support for the following positions:

- 1) Logistic regression detects DIF in non-simulated data.
- 2) Logistic regression detects more non-uniform DIF than the Mantel-Haenszel statistic.
- 3) Logistic regression is able to detect uniform and non-uniform DIF in multidimensional tests.
- 4) Although there were no mean differences in scores, DIF still was found.

# Discussion

Researchers and practitioners in educational and employment organizations should be aware of the inadequacies of the Mantel-Haenszel statistic when it is used in the development and evaluation phases of test construction.

Due to recent reforms in education and employment laws, there are serious social and legal implications for decisions made on the basis of tests that may be biased.

Greater consideration should be given to other viable alternatives that are more flexible and able to detect DIF in tests that are assumed to be unidimensional or multidimensional.

**As demonstrated in this empirical investigation, logistic regression is one alternative!**