

# Validity, Fairness, and Bias in Standardized Testing

by

Damon Ulysses Bryant, Ph.D.

Visiting Assistant Professor of Organizational Behavior

Tulane University

[dbryant@tulane.edu](mailto:dbryant@tulane.edu)

## Overview

- Purpose and research questions
- Validity and fairness in cognitive ability testing
- Theory and methods for investigating differential item functioning (DIF) and predictive bias
- Investigation of the potential for bias

# Purpose and Research Question

## Purpose

- To evaluate the extent to which validity judgments of test score interpretation, bias, and fairness can be made in regard to the Florida Comprehensive Assessment Test.

## Research Question

- Is the FCAT Mathematics section a fair assessment tool of mathematics achievement?

# Validity in Cognitive Ability Testing

- *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999; APA, 1966, 1974)
- *Principles for the Validation and Use of Personnel Selection Procedures* (APA, 1980, SIOP, 2003)
- Represent consensus of practices

# Validity in Cognitive Ability Testing

- One of the most important aspects of testing
- “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA, APA, & NCME, 1999, p. 9).
- Evolved from a view based on content, construct, and criterion-related validity, to one of a unitary concept (Messick, 1989; 1995b)

# Validity in Cognitive Ability Testing

Forms of Validity Evidence:

- Test Content
- Substantive Aspects of Construct Domain
- Internal Structure of the Test
- Generalizability of Test Score Interpretations
- External Relation with Other Variables
- Social Consequences of Testing

# Fairness in Cognitive Ability Testing

According to the *Standards and Principles*, fairness has several meanings:

- Lack of Bias
- Equitable Treatment
- Equality of Outcomes for All
- Opportunity to Learn

This study will focus on the lack of bias

# Fairness in Cognitive Ability Testing

*Standards* and *Principles* identify two types of bias studies:

- Item Bias (Differential Item Functioning) Studies-  
provide evidence of a consistent internal structure of a test across sub-populations and do not require an external criterion
- Predictive Bias Studies-  
give information about the relation between the construct as measured by the test and an external criterion of interest, which is also assumed to be bias-free

# Fairness in Cognitive Ability Testing: Differential Item Functioning

- Concerned with measurement bias at the item level.
- Referred to as differential item functioning (DIF) analyses
- Background
- Definition
- Methods
- Evidence

# Fairness in Cognitive Ability Testing: DIF Background

- Advanced under Item Response Theory (IRT)

IRT

Assumptions

Monotonicity

Local Independence

Unidimensionality

Models

1- Parameter Logistic

2- Parameter Logistic

3- Parameter Logistic

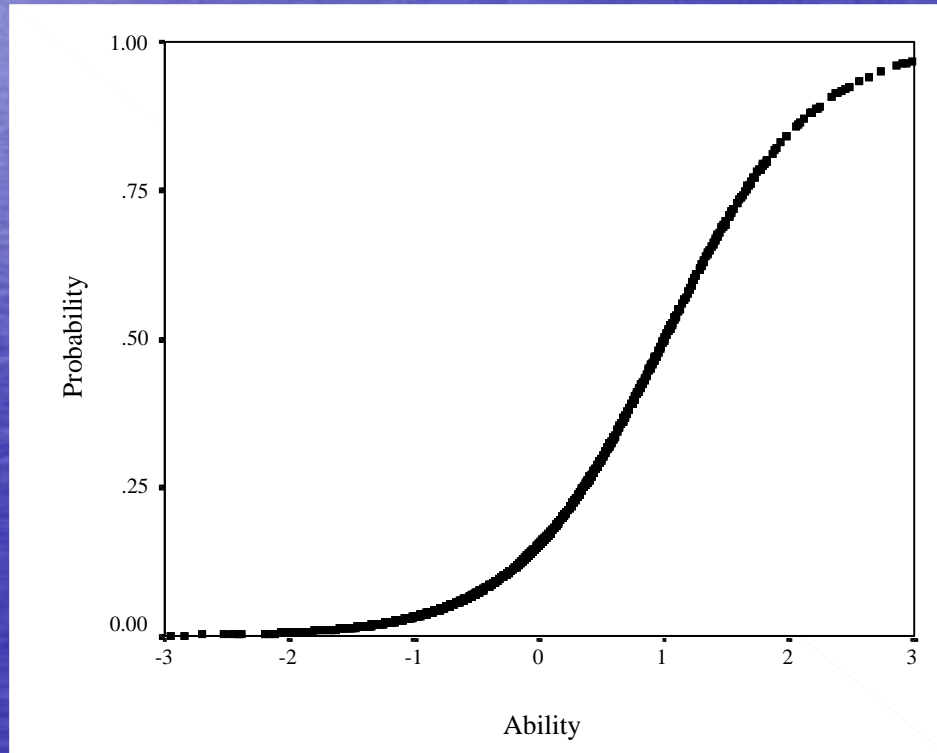
# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Assumptions

## Monotonicity

- As ability increases, the chance or probability of getting an item right increases.
- The function is non-decreasing with increases in the ability or latent trait ( $\theta$ ) being measured by the test (Lord & Novick, 1968).

# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Assumptions

## Monotonicity



# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Assumptions

## Local Independence

- When the latent trait ( $\theta$ ) is held constant, the conditional distributions of responses should be orthogonal.
- Probability or likelihood of any sequence of item responses occurring is simply the product of all individual item probabilities.

# Fairness in Cognitive Ability Testing: Item Response Theory (IRT)

## Assumptions

### Unidimensionality

- Responses to different test items are a function of one underlying trait ( $\theta$ )
- Criticized by some psychometricians (Hunter & Schmidt, 2000; Miller & Hirsch, 1992; Reckase, 1985, Reckase & McKinley, 1991)
- Researchers proposed alternative models and made a call to expand to a multidimensional framework (Bryant, 2005; Reckase & McKinley, 1991)

# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Models

- 1 – Parameter Logistic
- 2 – Parameter Logistic
- 3 – Parameter Logistic

# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Models

1- Parameter Logistic

$$P_i(\theta_j) = \{1 + \text{Exp}[-D(\theta_j - b_i)]\}^{-1}$$

Note:  $D$  is equal to a scaling constant 1.7 or

$b_i$  = difficulty of item  $i$ ,

$\text{Exp}(-L) = 2.71828^{-L}$ ,

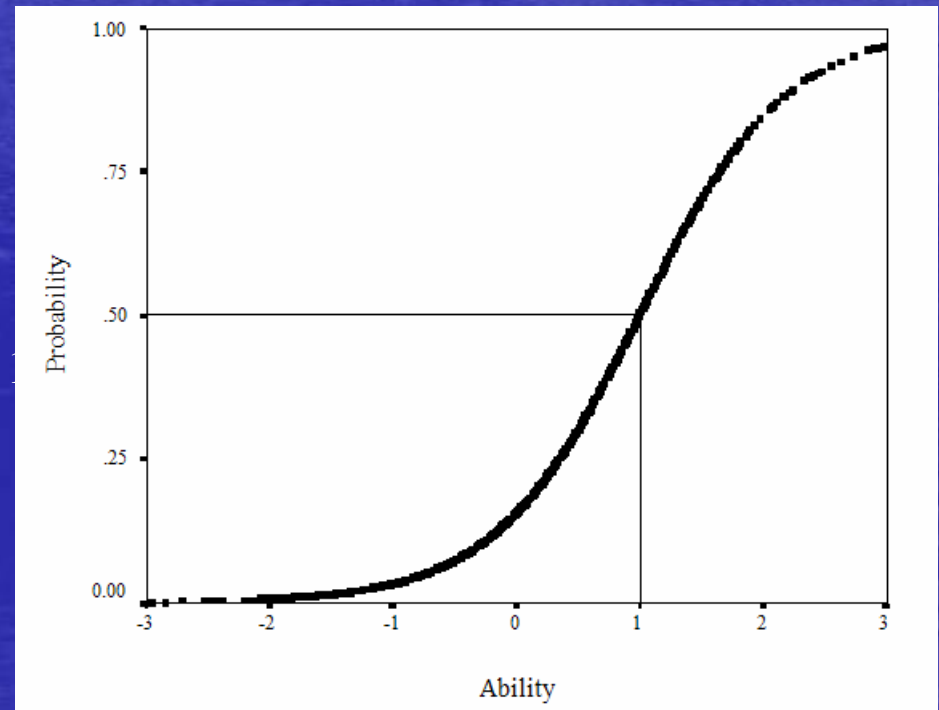
$L = D(\theta - b_i)$ , and

$Q_i(\theta) = 1 - P_i(\theta)$ .

Item Information Function

$$I_i(\theta) = D^2 P_i(\theta) Q_i(\theta)$$

**Figure 1.** Graph of 1-Parameter Logistic Model with  $b_i = 1.0$ .



# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Models

2- Parameter Logistic

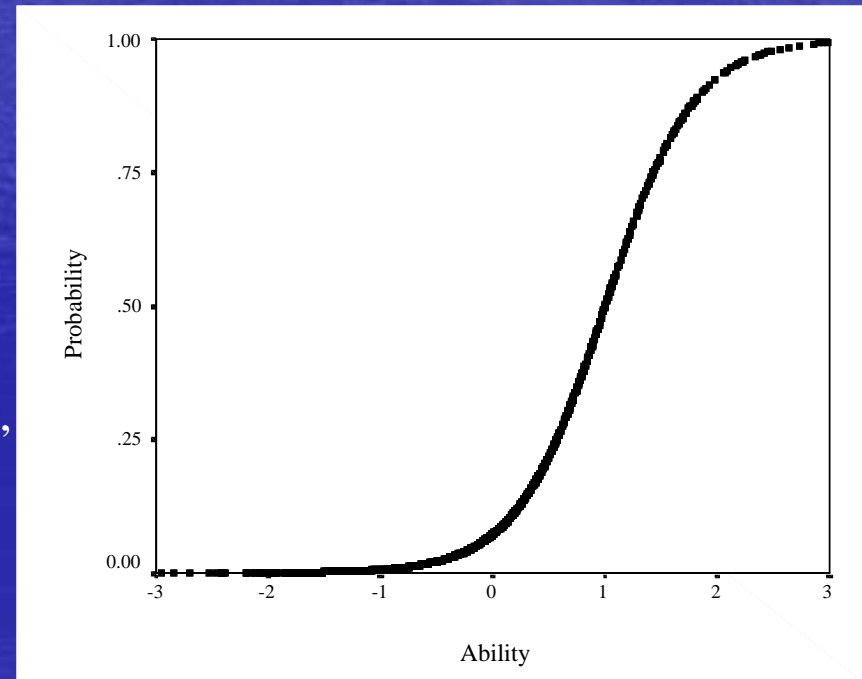
$$P_i(\theta_j) = \{1 + \text{Exp}[-Da_i(\theta - b_i)]\}^{-1}$$

Note:  $D$  is equal to a scaling constant 1.7 or 1,  
 $a_i$  = discrimination of item  $i$ ,  
 $b_i$  = difficulty of item  $i$ ,  
 $\text{Exp}(-L) = 2.71828^{-L}$ ,  $L = Da_i(\theta - b_i)$ , and  
 $Q_i(\theta) = 1 - P_i(\theta)$ .

Item Information Function

$$I_i(\theta) = D^2 a_i^2 P_i(\theta) Q_i(\theta)$$

**Figure 2.** Graph of 2-Parameter Logistic Model with  $b_i = 1.0$  and  $a_i = 1.5$ .



# Fairness in Cognitive Ability Testing: Item Response Theory (IRT) Models

## 3- Parameter Logistic

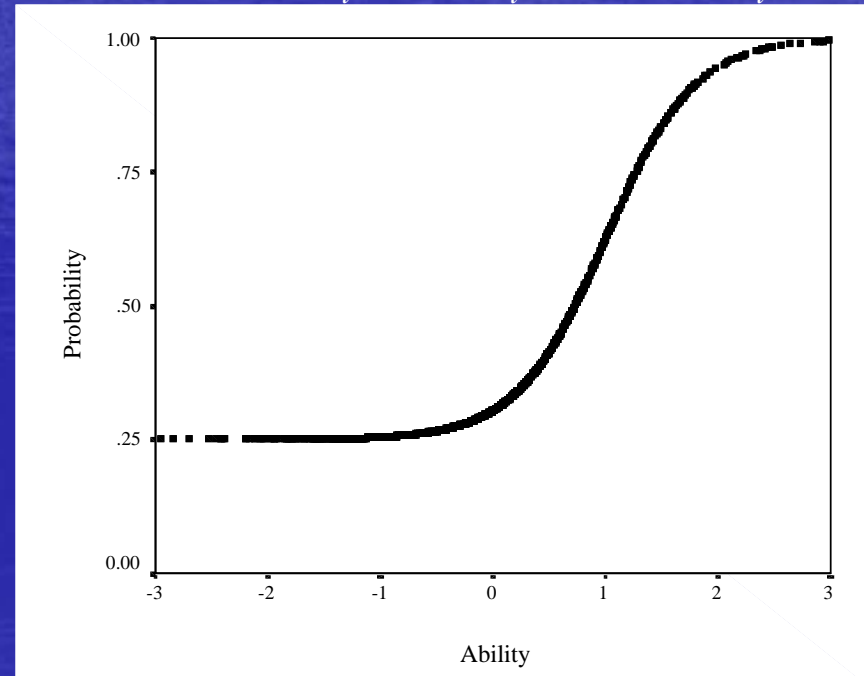
$$P_i(\theta_j) = c_i + (1 - c_i) \{1 + \text{Exp}[-Da_i(\theta - b_i)]\}^{-1}$$

Note:  $D$  is equal to a scaling constant 1.7 or 1,  
 $a_i$  = discrimination of item  $i$ ,  
 $b_i$  = difficulty of item  $i$ ,  
 $c_i$  = guessing of item  $i$ ,  
 $\text{Exp}(-L) = 2.71828^{-L}$ ,  $L = Da_i(\theta - b_i)$ , and  
 $Q_i(\theta) = 1 - P_i(\theta)$ .

## Item Information Function

$$I_i(\theta) = D^2 a_i^2 Q_i(\theta) \{P_i(\theta) [1 + \text{Exp}(-L)]^2\}^{-1}$$

**Figure 3.** Graph of 3-Parameter Logistic Model with  $b_i = 1.0$ ,  $a_i = 1.5$ , and  $c_i = .25$ .



# Fairness in Cognitive Ability Testing: Differential Item Functioning Definition

- Differential item functioning (DIF) is used to describe items on a test 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, pp. 397-398).
- Two groups are compared:  
A focal group (e.g., African-Americans or Females)  
A reference group (e.g., Anglo-Americans or Males)
- If some members of the focal group have the same ability (as indicated by total score or the latent trait) as members of the reference group but have a difference chance of getting the item right, then the item is known to exhibit DIF.

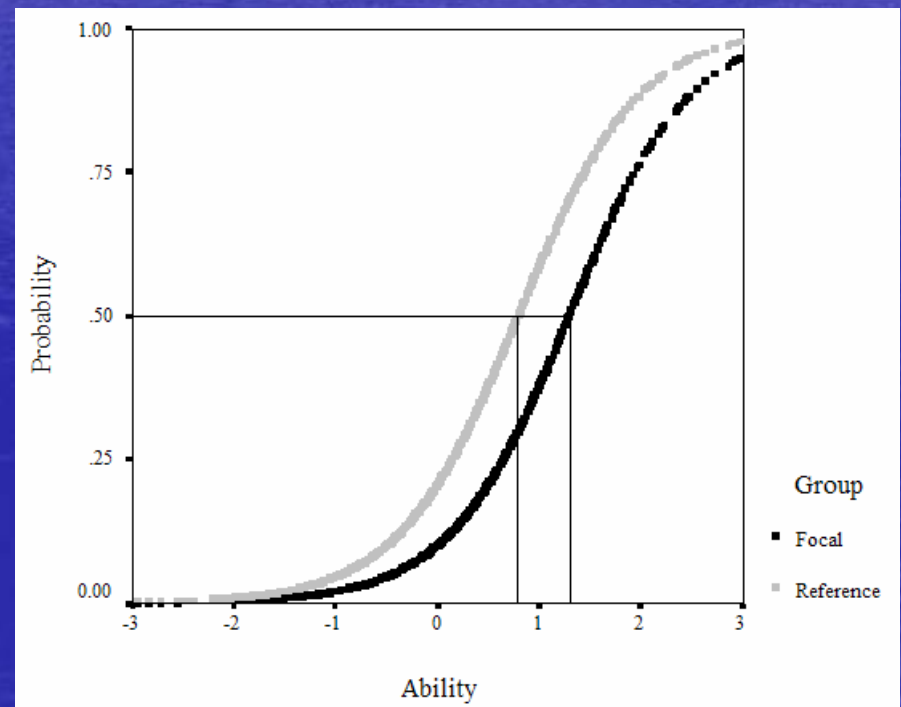
# Fairness in Cognitive Ability Testing: Differential Item Functioning Types

- Uniform DIF
- Non-Uniform DIF

# Fairness in Cognitive Ability Testing: Differential Item Functioning Types

## Uniform DIF

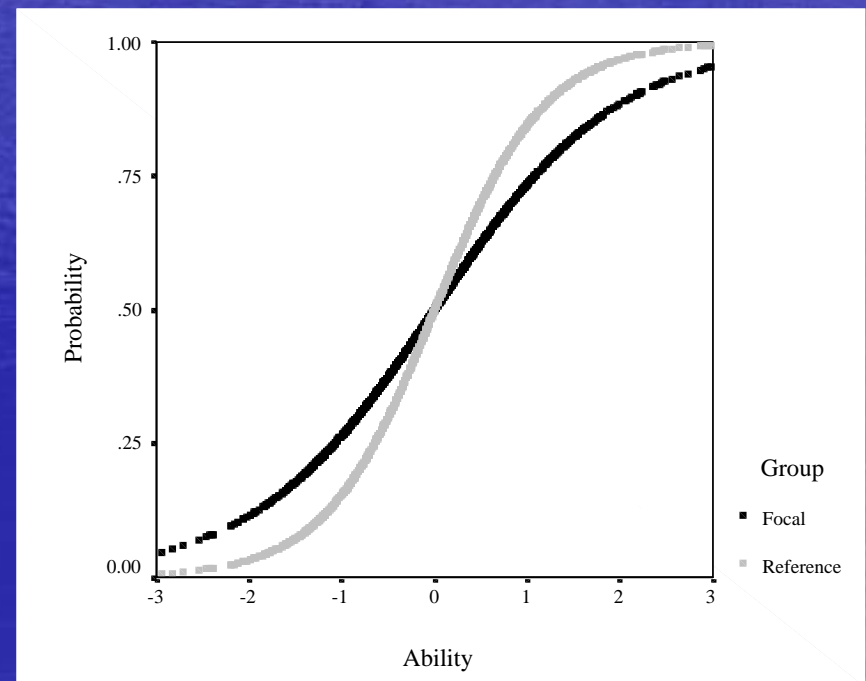
- A consistent difference in item difficulty across levels of ability (Mellenberg, 1982)
- Common odds ratios across latent trait (3PL: Parallel DIF)
- Curves are parallel across the latent trait or ability
- Shifts item probabilities
- Influences mean sub-group scores
- Violates assumptions of local independence



# Fairness in Cognitive Ability Testing: Differential Item Functioning Types

## Non-uniform DIF

- Non-uniform DIF is a between-group difference in item discriminations (Mellenberg, 1982)
- Uncommon odds ratios across latent trait (3-PL: Crossing DIF)
- Curves cross at a point on the latent trait or ability
- Influences test reliability
- Degrades item information
- Violates assumptions of local independence



# Fairness in Cognitive Ability Testing: Differential Item Functioning Types

There are two types of approaches to detect DIF:

## IRT:

Lord's Chi-Squared (Lord, 1980)

Signed and Unsigned Areas (Raju, 1988, 1990)

## Non-IRT:

Mantel-Haenszel Statistic (Holland & Thayer, 1988)

Logistic Regression (Swaminathan & Rogers, 1990)

SIBTEST (Shelly & Stout, 1996)

# Fairness in Cognitive Ability Testing: Differential Item Functioning Evidence

Logistic regression is more effective than the Mantel-Haenszel technique in detecting non-uniform DIF (Narayanan & Swaminathan, 1996; Rogers & Swaminathan, 1993; Swaminathan & Rogers, 1990)

Research on multidimensional DIF suggests that fewer items are found to exhibit DIF as compared to unidimensional methods (Clauser et al., 1996; Mazor et al., 1995; Mazor, Hambleton, & Clauser, 1998)

Uniform DIF occurs most often (Holland & Thayer, 1988, Swaminathan & Rogers, 1990)

Most methods are adequate in detecting uniform DIF

# Fairness in Cognitive Ability Testing: Differential Item Functioning Evidence

Some items show DIF because of a relevant between-group difference on a secondary trait not fully accounted for in the unidimensional composite (Shealy & Stout, 1993)

When multidimensional DIF techniques are used in conjunction with multidimensional measurement models, the number of items that show DIF is reduced (Mazor et al., 1995).

DIF has been found to occur at a rate of approximately twenty to thirty percent when relevant, secondary abilities are taken into account (Clauser et al., 1996)

# Fairness in Cognitive Ability Testing: Predictive Bias

- Concerned with the prediction of performance on a criterion of interest for different groups (AERA, APA, & NCME, 1999, SIOP, 2003).
- Background
- Definition
- Method
- Evidence

# Fairness in Cognitive Ability Testing: Predictive Bias Background

Evolved from views of:

Single-group Validity

Differential Validity

Both based on differences on correlation coefficients.

# Fairness in Cognitive Ability Testing: Differential Validity

Testing the differences in sub-group intercepts and slopes is now advocated and recognized as predictive bias studies (AERA, APA, & NCME, 1985, 1999)

Most consistent with Cleary's (1968) Model of Fairness

# Fairness in Cognitive Ability Testing: Predictive Bias Definition

Cleary (1968) states:

A test is biased for members of a subgroup of a population if in the prediction of a criterion for which the test was designed, consistent non-zero errors of prediction are made for members of the subgroups. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of a test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance. (p. 115).

# Fairness in Cognitive Ability Testing: Predictive Bias Method

## Multiple Regression

(AERA, APA, NCME, 1999; APA, 1980; SIOP, 2003; Stone, 1988; Stone & Hollenbeck, 1989)

Linear

All relevant variables included

Error-related assumption

- (a) the expected value of error is zero,
- (b) the variance of the error is constant across all levels of the predictor, i.e., homoscedasticity,
- (c) error terms are uncorrelated,
- (d) the predictors are uncorrelated with the error term
- (e) the error term is normally distributed.

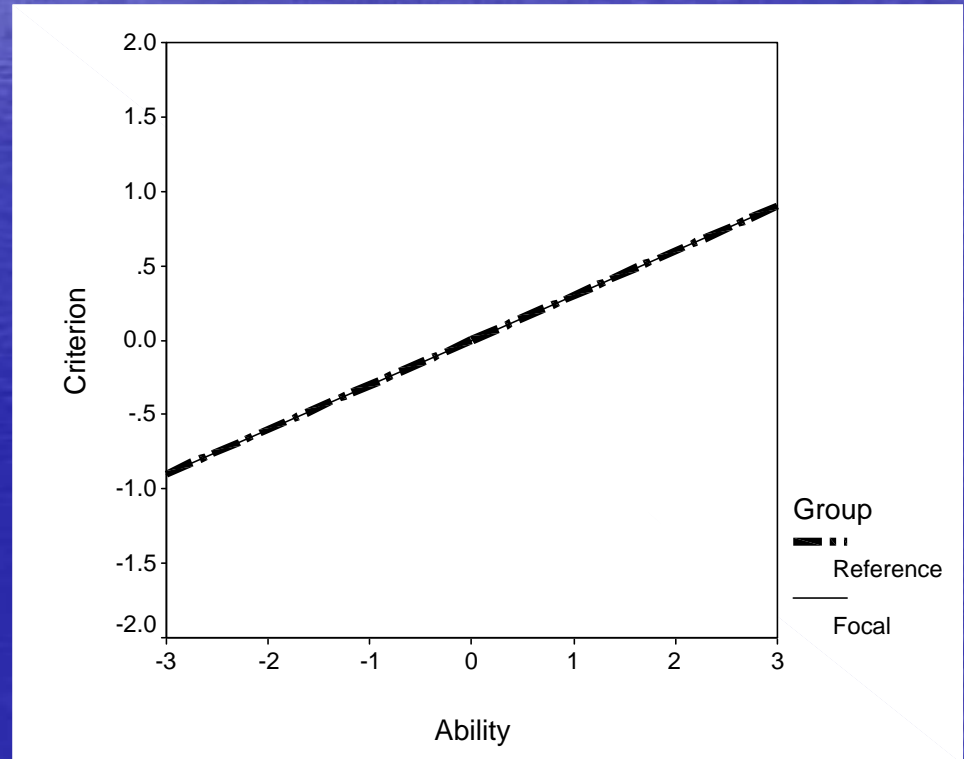
# Fairness in Cognitive Ability Testing: Predictive Bias Scenario

Several situations can occur in predicting scores on a criterion (Cleary et al., 1975) :

- (a) a similar regression line for both groups
- (b) uncommon regression lines due to sub-group intercept difference
- (c) uncommon regression lines due to a sub-group slope difference

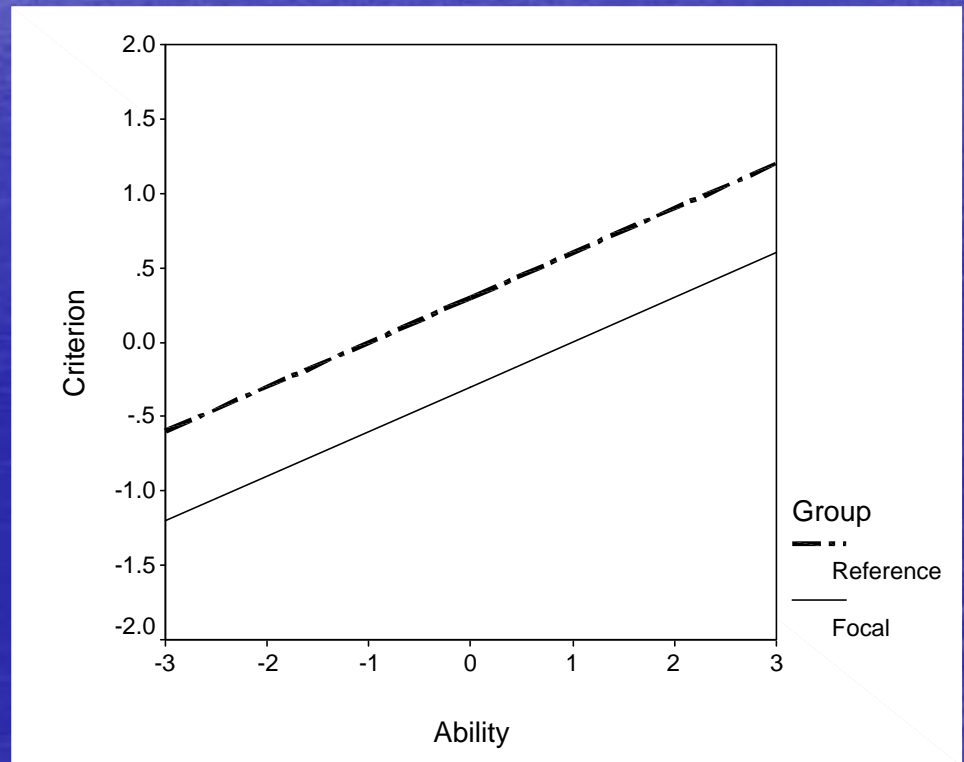
# Fairness in Cognitive Ability Testing: Predictive Bias Scenario

A similar regression line for both groups



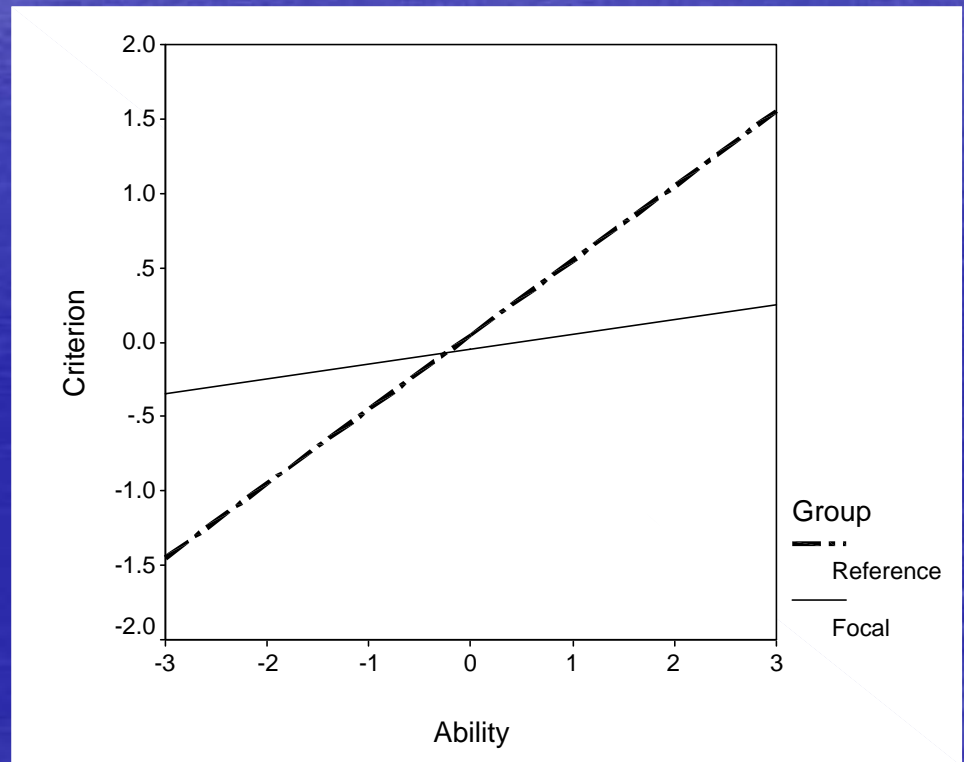
# Fairness in Cognitive Ability Testing: Predictive Bias Scenario

Uncommon regression lines due to sub-group intercept differences



# Fairness in Cognitive Ability Testing: Predictive Bias Scenario

Uncommon regression lines due to a sub-group slope difference



# Fairness in Cognitive Ability Testing: Predictive Bias Evidence

Differences in sub-group intercepts with little evidence of slope differences (Cascio, 1998; Hartigan & Wigdor, 1989; Hunter & Schmidt, 2000; Jensen, 1980; Schmidt & Hunter, 1974)

Hartigan and Wigdor (1989)

In 72 studies that had at least 50 African-American and 50 non-minorities

Two (2) studies showed evidence of slope differences

Twenty-six (26) showed evidence of intercept differences

Hunter, Schmidt, and Raschenberger (1984)

In selection, tests are found to be biased

Predictive bias is interpreted to be against majority group members (e.g., Anglo-Americans)

# Fairness in Cognitive Ability Testing: Predictive Bias Evidence

Cognitive ability tests overpredict the performance of minority group members (Bartlett, Bobko, Mosier, & Hannan, 1978; Jensen, 1980)

Little evidence of tests being statistically biased, in the predictive sense, against minorities or focal groups

In the vast majority of conditions, differential prediction has been found when there is a difference in sub-group means in favor of the majority group (Schmidt & Hunter, 1974).

In contrast to what most researchers imply about predictive bias and DIF being related and mutually supportive of each other, empirical evidence suggests otherwise (Bryant, 2004)

# Investigation of the potential for bias in the FCAT

- Little evidence exists in evaluating FCAT
  - No item level data for demographic groups (no DIF analysis)
  - No external criterion (no predictive bias analysis)
  - No support from the state in obtaining item level data for demographic groups

# Investigation of the potential for bias in the FCAT

## Evidence of test content

- Math (2005)
- Reading (2005)
- Obtained online from FCAT site
  - SSS Standards
  - Item level difficulty (aggregated)
  - Subject matter expert's ratings of content difficulty

# Investigation of the potential for bias in the FCAT

- Use two Standards from Fairness in Testing and Test Use
  - 7.7: “In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct.”

# Investigation of the potential for bias in the FCAT

- Use two Standards from Fairness in Testing and Test Use
  - 7.10: “When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests.”

# Investigation of the potential for bias in the FCAT

- Use two Standards from Fairness in Testing and Test Use
  - 7.10 (con't): "Where mean differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct under-representation or construct irrelevant variance. While initially the responsibility of the test developer, the test user bears the responsibility for uses with groups other than those specified by the developer."

# Investigation of the potential for bias in the FCAT

- Two FCAT Mathematic sections are evaluated:
  - 4 grade (2005)
  - 8 grade (2005)
- Two demographic groups are examined
  - Whites
  - Blacks

# Investigation of the potential for bias in the FCAT

- Is there a difference in the proportions attaining Level 3 or higher in Reading?
- Does adverse impact (i.e., pass rate ratio for demographic groups less than 80%) exist on the FCAT Reading section of the FCAT for 4<sup>th</sup> and 8<sup>th</sup> graders?

# Investigation of the potential for bias in the FCAT

- Proportions of students at Level 3 or higher differ for Whites and Blacks ( $p < .01$ )
  - Reading 4<sup>th</sup>, Whites .81
  - Reading 4<sup>th</sup>, Blacks .56 (Adverse Impact, 69%)
  - Reading 8<sup>th</sup>, Whites .56
  - Reading 8<sup>th</sup>, Blacks .24 (Adverse Impact, 43%)

# Investigation of the potential for bias in the FCAT

- Is there a difference between Blacks and Whites in the proportion attaining Level 3 or higher in Mathematics? (Standard 7.10)
- Does adverse impact exist on the Mathematics section of the FCAT for 4<sup>th</sup> and 8<sup>th</sup> graders?

# Investigation of the potential for bias in the FCAT

- Proportion of students at Level 3 or higher differ for Whites and Blacks ( $p < .01$ )
  - Math 4<sup>th</sup>, Whites .74
  - Math 4<sup>th</sup>, Blacks .44 (Adverse Impact, 59%)
  - Math 8<sup>th</sup>, Whites .71
  - Math 8<sup>th</sup>, Blacks .36 (Adverse Impact, 51%)

# Investigation of the potential for bias in the FCAT

- Is the reading level of questions appropriate for the Mathematics section of the FCAT?
- Standards would suggest that reading level should be below reading level assessed in the FCAT Reading section
- 4<sup>th</sup> Grade Math
  - Questions should be at the 3<sup>rd</sup> grade level or lower
- 8<sup>th</sup> Grade Math
  - Questions should be at the 7<sup>th</sup> grade level or lower

# Investigation of the potential for bias in the FCAT

- Is the reading level of questions appropriate for the Mathematics section of the FCAT?
- 4<sup>th</sup> Grade
  - Average Reading Level = 6.5 (Flesch-Kincaid Readability Test)
  - Number of questions = 40
  - Reading Level tested against grade level of 3
  - Found to be significantly higher than 3<sup>rd</sup> grade ( $p < .001$ )

# Investigation of the potential for bias in the FCAT

- Is the reading level of questions appropriate for the Mathematics section of the FCAT?
- 8<sup>th</sup> Grade
  - Average Reading Level = 7.6 (Flesch-Kincaid Readability Test)
  - Number of questions = 50
  - Tested against population grade level of 7.
  - Found to be significantly higher than 7<sup>th</sup> grade ( $p < .05$ )

# Investigation of the potential for bias in the FCAT

- Does the reading level of questions in the Mathematics section account for systematic differences in question performance after accounting for SSS content?
- 4<sup>th</sup> Grade (multiple regression analysis, N=40)
  - Criterion: Proportion of correct responses to question (item difficulty)
  - Predictors: SSS content rating, Flesch-Kincaid Reading level
  - Model accounts 24% of variance in item difficulty
  - Content ratings found to be a significant predictor of difficulty ( $p < .05$ )
  - Reading level was not significant but was related to content ratings ( $p < .05$ )

# Investigation of the potential for bias in the FCAT

- Does the reading level of questions in the FCAT Mathematics section account for systematic differences in question performance after accounting for content?
- 8<sup>th</sup> Grade (multiple regression analysis, N=50)
  - Criterion: Proportion of correct responses to question (item difficulty)
  - Predictors: SSS content rating, Flesch-Kincaid reading level
  - Model accounts 22% of variance in item difficulty
  - Content rating found to be a significant predictor of difficulty ( $p < .05$ )
  - Reading level found to be a significant predictor of difficulty ( $p < .05$ )

# Summary

- The FCAT Mathematic section may not be a fair assessment tool of mathematical achievement
- Scores from the FCAT Mathematics section will lead to biased interpretations and inferences due to the excessive reading demands

# Summary

- Implications
  - Schools
    - Loss of funding due to biased interpretation
    - State takeover
  - Teachers
    - Performance management in math due to student performance in reading
    - Job loss
  - Students
    - Retention in grade
    - Psychological consequences (e.g., self-esteem and self-efficacy)
- Alternatives
  - Low stakes assessments for development
  - Portfolio assessments and standardized rating systems
- Future research
  - Expose issues and propose research agenda to solve fairness concerns



# Questions