
The Use of Statistical Analysis in Assessing Bias

Presented to The 2010 Annual Conference for Personnel Testing
Council of Northern California (Berkeley, California)

3/10/2010

Dale Glaser, Ph.D.
Principal, Glaser Consulting
Lecturer/Adjunct Faculty: SDSU/USD/AIU

Program Objectives

- Defining Bias
- Differentiating Bias (e.g., vs. Fairness)
- Statistical procedures and Interpretation
 - Regression Analysis
 - Differential Item Functioning

Recent Examples:

“Justices to Hear White Firefighters’ Bias Claims”

- NEW HAVEN — Frank Ricci has been a firefighter here for 11 years, and he would do just about anything to advance to lieutenant.
- The last time the city offered a promotional exam, he said in a sworn statement, he gave up a second job and studied up to 13 hours a day. Mr. Ricci, who is dyslexic, paid an acquaintance more than \$1,000 to read textbooks onto audiotapes. He made flashcards, took practice tests, worked with a study group and participated in mock interviews.
- Mr. Ricci did well, he said, coming in sixth among the 77 candidates who took the exam. But the city threw out the test, because none of the 19 African-American firefighters who took it qualified for promotion. That decision prompted Mr. Ricci and 17 other white firefighters, including one Hispanic, to sue the city, alleging racial discrimination.
- The suit brought by Mr. Ricci and his colleagues says that the city’s rationale for throwing out the test is illegitimate and that they were denied a chance for promotion on account of the color of their skin. Karen Lee Torre, a lawyer for the firefighters, declined to be interviewed and said she had instructed her clients not to speak to reporters.

And on 6/30/09.....”High Court Backs White Firefighters in Bias Case”

WASHINGTON — The U.S. Supreme Court ruled yesterday that white firefighters in Connecticut were subjected to race discrimination when the city of New Haven threw out a promotional examination on which they had done well and black firefighters poorly.

“The city rejected the test results solely because the higher scoring candidates were white,” Justice Anthony Kennedy wrote for the majority, adding that the possibility of a lawsuit from minority firefighters was not a lawful justification for the city's action.

“Fear of litigation alone cannot justify an employer's reliance on race to the detriment of individuals who passed the examinations and qualified for promotions,” Kennedy wrote.

Recent Examples:

“How do you spot raw legal talent? Take this Test”

- “University of CA, Berkeley psychology professor Sheldon Zedeck, PhD and retired Berkeley law professor Marjorie M. Schultz have designed a test that may predict legal savvy and success better than the LSAT.....which is often criticized for only predicting law school success and keeping African-Americans who tend to score low out of law school
- Polling law professors, legal clients, etc. they came up with 26 “effectiveness factors” such as applicant’s ability to problem-solve, write, listen, mentor, negotiate and advocate for one’s client [used situational judgment test.....i.e., hypothetical situations as predictors of how test-takers fared on the 26 traits]
- In line with past studies, LSAT scores predicted who fared well in law school....but the Zedeck and Schultz test indicated which lawyers were the most effective

- Monitor on Psychology, June 2009, p. 12

Recent Examples:

Supreme Court Fixture Faced Bias as Early Female Lawyer

- “Born in the era before women could vote, Patricia Dwinnell Butler shocked her mother when she insisted on going to law school. After earning a law degree in 1931, she had trouble getting hired by any law firms, which told her they couldn’t hire a woman. But years later, she would become a fixture at the U.S Supreme Court”
- SD Union, 6/4/09

Variety of Definitions of Bias

(1) UNIVERSITY POLICE DEFINITION OF BIAS MOTIVATED INCIDENT

"Any offense or unlawful act based on a victim's race, sex, age, gender, disability, ethnic origin, religion, or sexual orientation is a bias motivated incident." (University Police, Field Operations Manual, 1992)

(2) What is artifact (single-point) calibration? (engineering)/What is Bias?

The operational definition of bias is that it is the difference between values that would be assigned to an artifact by the client laboratory and the laboratory maintaining the reference standards. Values, in this sense, are understood to be the long-term averages that would be achieved in both laboratories.

Variety of Definitions of Bias

**(3) RACISM, GENDER-BIAS,
and Other Forms of BIGOTRY
In the Writings of
ALEISTER CROWLEY**

Prejudice means to pre-judge someone or something before coming to know it. Bias means to look at someone or something in a slanted, skewed, or unfair way.

(4) Bias A statistical sampling or testing error caused by systematically favoring some outcomes over others.

Bias.....

Bias is manifest in texts when authors present particular values as if they were universal. For example, bias can be conveyed in the media through the selection of stories, sequence, and slant in newscasts; the placement or omission of stories in newspapers; who is interviewed and left out in radio or television talk shows and news programs; the advertisements on webpages, television, magazines, radio shows targeted at specific audiences; the lyrics of commercial jingles and popular music, and the images displayed with them in broadcast commercials and music videos; the goals, procedures, and the rules of video games.

Excerpt from: *Crossing the Information Highway: The Web of Meanings and Bias in Global Media*

author: [Ladislav Semali](#), December 2002,
Readingonline.org

Test Bias.....

- “For psychometricians, **bias** is a factor inherent in a test that systematically prevents accurate, impartial measurement”
 - (Cohen & Swerdlik, 2010 p. 199)

Test Bias.....

- “Test bias is a fundamentally important issue in testing, as pervasive an systematic errors can lead to erroneous inferences regarding the interpretation and use of test scores, compromising validity of the instrument”
 - (Meade & Fetzner, 2009 p. 738)

Bias and Fairness

- Bias and fairness represent fundamentally different concepts (in the context of testing)
 - A biased test may lead to decisions that are regarded widely as fair
 - Whereas an unbiased test may lead to decisions that are condemned universally as unfair
 - Murphy and Davidshofer, 2005, p. 317

Bias vs. Fairness

■ Bias

- Refers to test scores
- Based on statistical characteristics
- Is defined empirically
- Can be scientifically determined

■ Fairness

- Refers to actions taken or decisions made
- Is a value judgment regarding outcomes
- Is defined in philosophical or political terms
- Cannot be scientifically determined

Murphy & Davidshofer (2005)

A LEGAL DEFINITION OF FAIRNESS (<http://www.psy-lab.org/fairness.htm>)

Legally, a selection procedure (e.g., interview, intelligence test, etc.) is fair if it is a valid measure of a necessary job requirement. When the selection procedure leads to a higher ratio of hires for one protected class (i.e., any group sharing race, sex, religion, color, national origin) over another protected class, there is **adverse impact**.

• Adverse impact as defined by the **4/5's rule** occurs if the selection ratio for any subgroup of people is less than 4/5's of the selection ratio for the largest group. The following is an example that looks at race.

$$\frac{20 \text{ whites hired}}{100 \text{ white applicants}} = \text{selection ratio of } .20$$

$$4/5 (80\%) \text{ of } .20 = .80 \times .20 = .16$$

So if selection ratio for any other race is $< .16$ there is adverse impact. If 50 blacks apply, at least 8 should be hired to avoid adverse impact.

$$\frac{8}{50} = .16$$

If there is adverse impact, validation study may be necessary to prove the test is job-related. If so, adverse impact is legal

Note: This example could use any 2 races, religions, colors, national origins, and of course the two sexes.

Complexity in Defining Bias

- Defining bias is not a simple task (cf., Hunter & Schmidt, 1976)
- Plausible alternative definitions are often contradictory
- Perhaps the most basic question asks to whom we should be fair---the individual or the group to which the individual belongs
 - Nunnally and Bernstein, 1994, p. 359

Ethical Concerns and Defining Bias:

Three Ethical Positions

- (1) **Unqualified Individualism**—use tests to select the most qualified individuals that can be found....indifferent to race or gender of applicants
- (2) **Quotas**—explicitly recognize race and gender differences (e.g., if 20% African Americans in the population, then select 20% African Americans for company)
- (3) **Qualified Individualism**—compromise between unqualified individualism and quotas.....subscribes to notion of selecting the best qualified personnel...but also takes into consideration race, gender, religion, etc....
 - Kaplan & Saccuzzo, 2009, p. 532

Some Key Questions Regarding Bias

- ❑ Does the test measure different things for different groups?

- ❑ Is there bias in prediction? (overestimate or underestimate)
 - Due to:
 - ❑ Differences in test scores
 - ❑ Differences in criterion performance
 - ❑ Differences in variability of test scores/performance
 - ❑ Difference in test validity

Murphy & Davidshofer (2005)

Sources/Remedies of Bias

❑ Sources

- ❑ Real mean differences in the attributes
- ❑ Differences are function of the test
- ❑ Content of test is familiar to some groups, but not others
- ❑ Method of presentation (e.g., written vs. video, etc.)
- ❑ Interaction with administrator

❑ Remedies

- ❑ Change content of test (Golden Rule)
- ❑ Employ multiple methods of assessment
- ❑ Change the testing method

Murphy & Davidshofer (2005)

In the Context of Ascertaining Bias, Why Use Statistics?

- Contingent on how the data was collected (i.e., random sampling, etc) statistical analysis can be one avenue by which to “objectively” assess test bias
 - However, critical to attend to:
 - Sampling methodology: random selection?
 - Measurement: quality of measures (e.g., reliability, validity, etc.)
 - Type of analysis (does the technique correspond to the issue/question at hand?)

Two Analytic Strategies

■ Internal Methods

- Comparisons of relationship between the latent trait and item responses (e.g., measurement invariance via confirmatory factor analysis (CFA), differential item functioning (DIF) via item response theory (IRT))

■ External Methods

- Makes use of a criterion measure that is external to the test (Cleary method testing slopes/intercept subgroup differences via regression analysis)

Methods for Analyzing Bias: Regression Analysis

- Regression analysis is a statistical technique often used for predicting an outcome/criterion (Cohen et al, 2003)
 - How well does SAT predict first year GPA at college?
 - To what extent does LSAT predict success in law school?.....or furthermore, as a practicing attorney?
 - Do personality inventories predict success on the job?
 -all of the above may be used for criterion/predictive validity

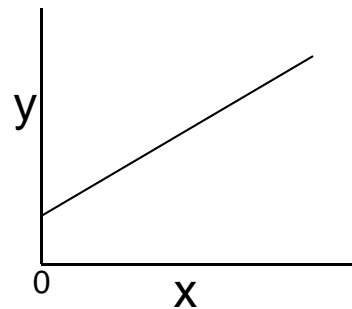
Methods for Analyzing Bias: Regression Analysis

- What is key is operationally defining...
 - **The predictor (x)**.....what exactly constitutes 'personality'?
 - **The criterion (y)**.....how would you measure success...as an attorney? Chef? CEO? Accounting clerk? Dental hygienist?
 - ...often referred to as the criterion problem...not always clear how you will measure/operationalize the outcome.....

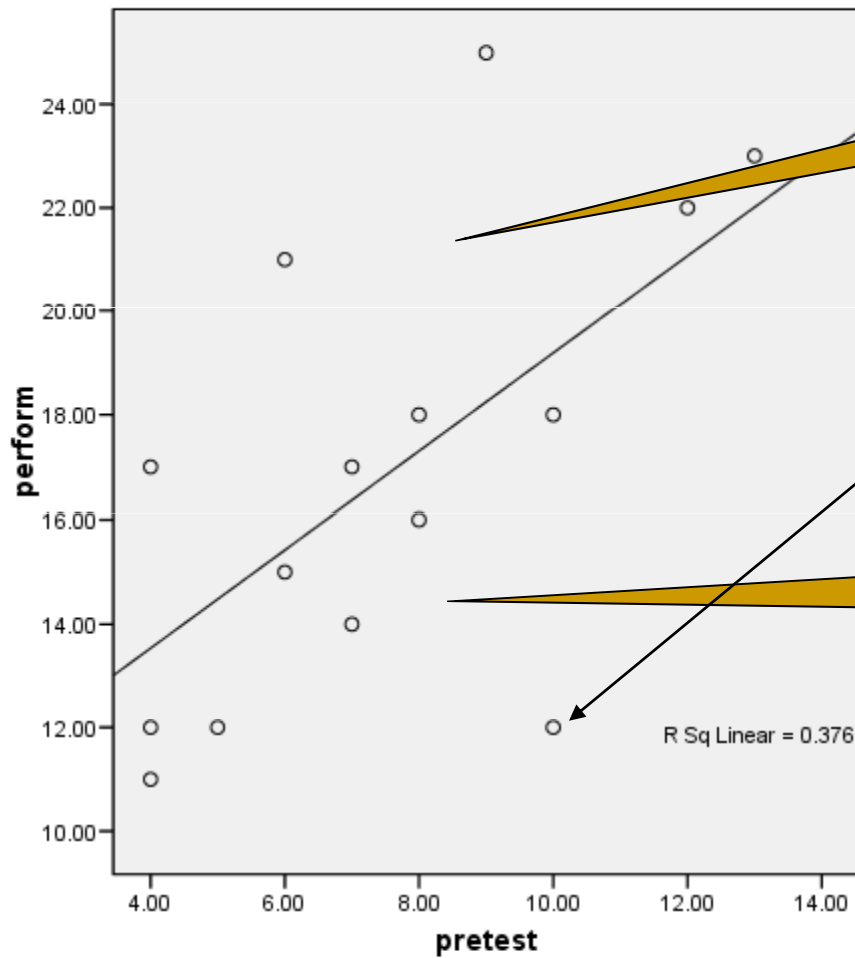
Methods for Analyzing Bias: Regression Analysis

■ The regression equation:

- The slope.....change in y (e.g., task performance) based on metric change in x (pretest...selection tool) [slope = $\Delta y / \Delta x$]
- The intercept (when $x = 0$, where does it cross the y axis)
 - There can be statistical biases in (1) the slope; (2) the intercept or (3) both



Graphic: the fit line



Above the line: under predicted

For the full (and very small) sample ($n = 15$) we see a relatively nice fit to the line.... though notice outliers

Below the line: over predicted

...and the equation..

Descriptive Statistics

	Mean	Std. Deviation	N
perform	16.8667	4.37308	15
pretest	7.5333	2.85023	15

No hard and fast rule but per Cohen (1988), small/medium/large R square = .01/.09/.25 for correlation squared and .02/.13/.26 for multiple regression R squared....
....so this means 37.6% of the variance in performance can be attributed to the pretest

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.614 ^a	.376	.328	3.58354	.376	7.849	1	13	.015

a. Predictors: (Constant), pretest

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.775	2.695		3.627	.003
	pretest	.941	.336	.614	2.802	.015

a. Dependent Variable: perform

Slope = .941.....as pretest goes from 5 to 6 performance goes up by .941.
Intercept = 9.775.....when pretest = 0 predicted performance = 9.775

About significance.....(statistical vs. practical)

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	9.775	2.695		3.627	.003
	pretest	.941	.336	.614	2.802	.015

a. Dependent Variable: perform

So, if we use the standard significance level of .05 (which means we are OK with wrongly rejecting the null hypothesis 5% of the time when the null is true (i.e., saying there is a difference when there isn't)), we see we have significance here ($p = .015$)

.....however.....must temper enthusiasm for obtaining significance with interpretation of magnitude/effect size...

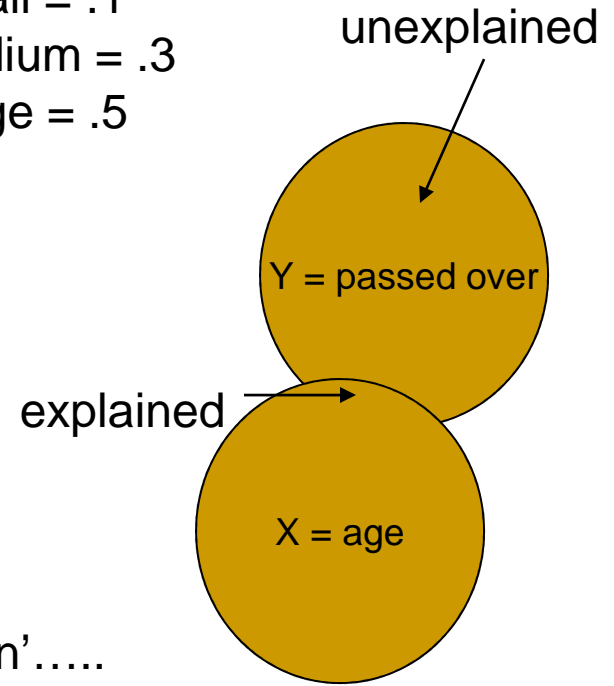
Just reporting p -values in isolation does not shed insight into strength of predictor/manipulation...hence, why Cohen (1994) and others for years have been exhorting reporting of effect sizes and confidence intervals (CI) besides p -values

Significant but....???

r
 Small = .1
 Medium = .3
 Large = .5

		age Age of Respondent	work4 Being Passed Over for Promotion
age Age of Respondent	Pearson Correlation	1	.111**
	Sig. (2-tailed)		.001
	N	1514	970
work4 Being Passed Over for Promotion	Pearson Correlation	.111**	1
	Sig. (2-tailed)	.001	
	N	970	971

** . Correlation is significant at the 0.01 level (2-tailed).

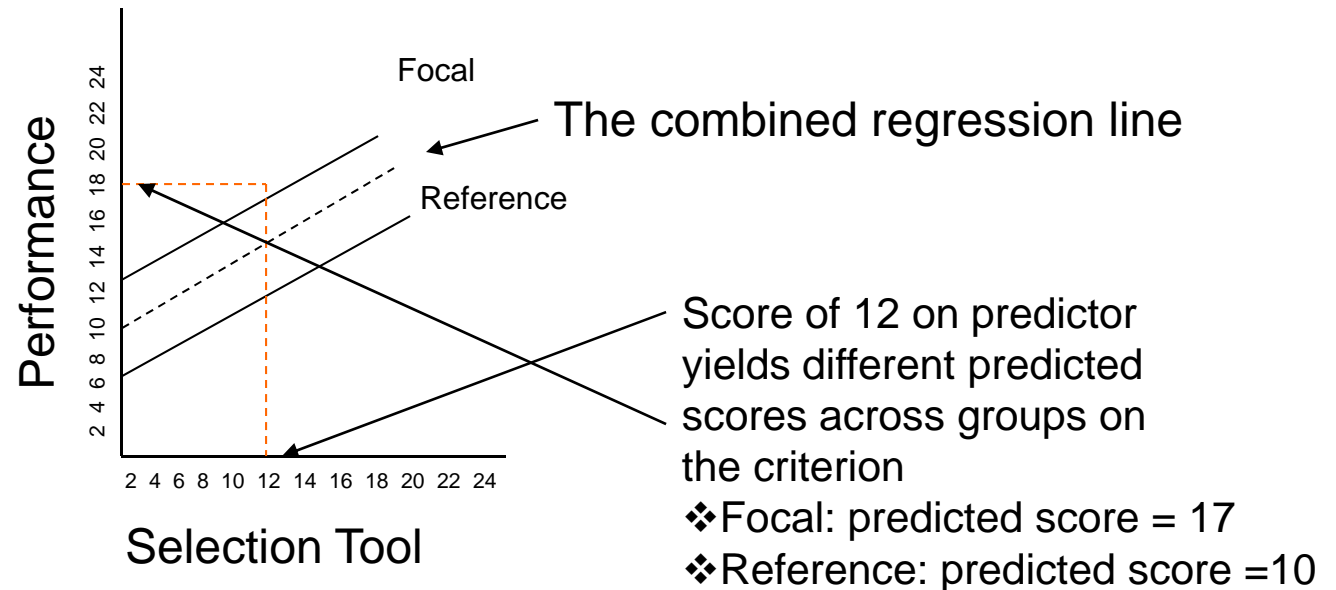


Here we have a “positive” and “significant ($p = .001$) relationship of age and ‘being passed over for promotion’.....
 But with $r = .111$, we have accounted for 1.23% of the variance in the outcome...meaning there is 98.77% error variance which constitutes a tremendous amount of unexplained variance!!!....and one of the contributors to significance?.....sample size: $n = 970$
 With large samples it becomes increasingly easier to reject the null....and to equate “statistical significance” with “practical significance”...i.e., equating “significance” with “importance”..so even if you find statistical evidence for bias via significance testing, you have to temper your conclusions with interpretation of effect size/magnitude.....

Testing for Bias

- Someone alleges that a pretest is biased against minority group (e.g., Latino)
- One option is to conduct a separate regression analysis for each group (focal vs. reference group) and see if the parameter estimates (slope, intercept, correlation) seem to vary across groups

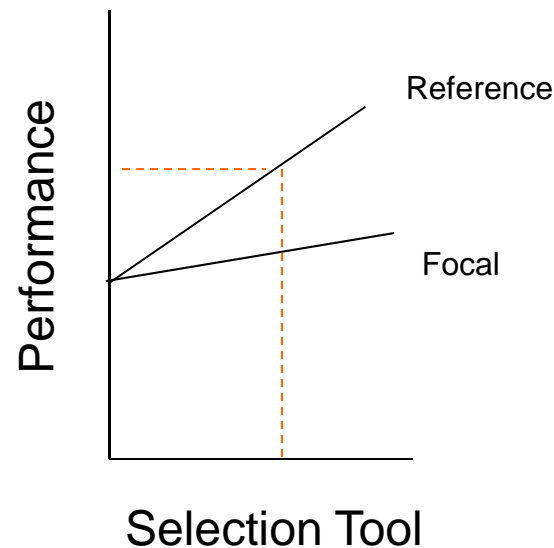
PATTERNS TO CONSIDER: INTERCEPT BIAS



- Intercept bias denotes unfairness to the focal group
- The test systematically underpredicts or overpredicts for the groups
- In the above example we see that the predictor systematically underpredicts performance for the focal (minority) group.....and systematically overpredicts performance for the reference (majority) group
.....hence the “use of a single regression line produces discrimination in favor of [the reference group] and against [the focal group]”

(Kaplan & Saccuzzo, 2009, p .522).

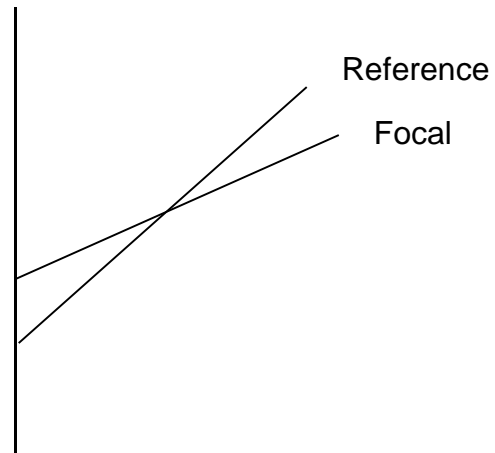
PATTERNS TO CONSIDER: SLOPE BIAS



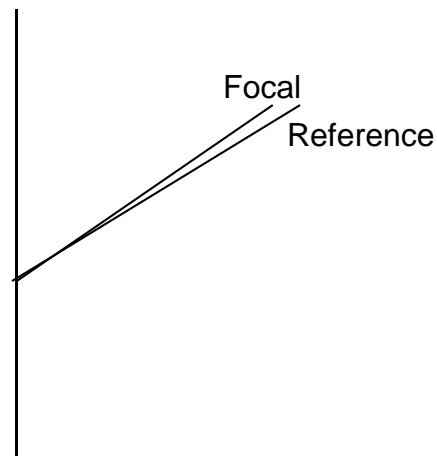
- Slope bias means that the tool has differential validity
-meaning that the correlation of the selection tool and performance will vary across groups....
- The test systematically underpredicts for the focal group
- Inappropriate to use a common regression line
- “this is the most clear-cut example of test bias”

(Kaplan & Saccuzzo, 2009 p. 523)

.....and other patterns to consider....



Slope bias &
Intercept bias



No Slope bias
No Intercept bias

Subgroup Analysis

Descriptive Statistics

ethnic		Mean	Std. Deviation	N
.00 white	perform	17.5714	4.35343	7
	pretest	8.0000	3.31662	7
1.00 Latino	perform	16.2500	4.59036	8
	pretest	7.1250	2.53194	8

Notice slightly higher mean on pretest for whites..

But is that due to 'bias' or true differences?

Correlations

ethnic			perform	pretest
.00 white	Pearson Correlation	perform	1.000	.796
		pretest	.796	1.000
	Sig. (1-tailed)	perform	.	.016
		pretest	.016	.
N	perform	7	7	
	pretest	7	7	
1.00 Latino	Pearson Correlation	perform	1.000	.415
		pretest	.415	1.000
	Sig. (1-tailed)	perform	.	.153
		pretest	.153	.
N	perform	8	8	
	pretest	8	8	

Uh oh, fairly substantive difference in correlations....

r = .796 for Whites

r = .415 for Latinos

...and the equation...

Model Summary

ethnic	Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
						R Square Change	F Change	df1	df2	Sig. F Change
.00 white	1	.796 ^a	.634	.561	2.88368	.634	8.675	1	5	.032
1.00 Latino	1	.415 ^a	.172	.034	4.51141	.172	1.247	1	6	.307

a. Predictors: (Constant), pretest

Coefficients^a

ethnic	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
.00 white	1	(Constant)	9.208	3.042		3.027	.029
		pretest	1.045	.355	.796	2.945	.032
1.00 Latino	1	(Constant)	10.891	5.057		2.154	.075
		pretest	.752	.673	.415	1.117	.307

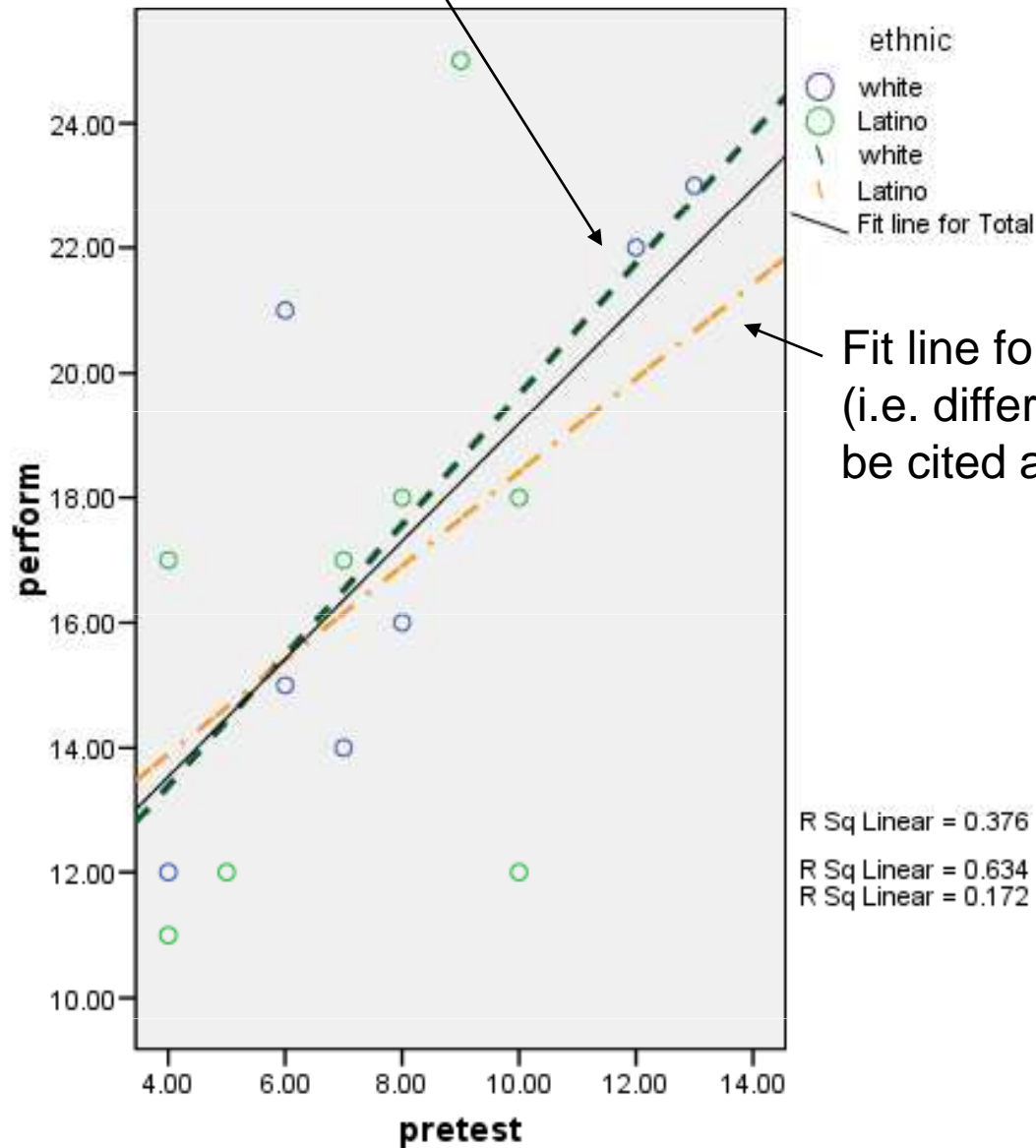
a. Dependent Variable: perform

$Y' = 9.208 + 1.045$ (pretest) for Whites ($p = .032$significant)

$Y' = 10.89 + .752$ (pretest) for Latinos ($p = .075$..not significant)

↑ ↑
intercept slope

Fit line for whites.. Any perceptible differences?



Fit line for latinos..this type of slope bias (i.e. difference in validity coefficient (r_{xy})) may be cited as evidence of **differential validity**

And this is not a trivial difference
In variation explained....

63.4% for Whites

17.2% for Latinos

Possible reasons for differences in intercepts.....

- Bias in the test
- Bias in the criterion
- Reliability of the test
- Omitted variables (Sackett, Laczko, & Lippe, 2003)

Another Option: Moderated Multiple Regression (MMR)

- Within a regression framework, can see if the relationship of the pretest and the outcome (i.e., performance) is contingent/varies across the levels of ethnicity (i.e., the categorical variable of interest)
- The above is called testing the interaction of Pretest x Ethnicity (Criterion = Performance)
- Two words that capture the spirit of an interaction: “it depends”!!!
- This is often called Moderated Multiple Regression (MMR) (Aguinis, 2004)
- And here you can test the interaction for statistical significance

Moderated Multiple Regression

Descriptive Statistics

	Mean	Std. Deviation	N
perform	16.8667	4.37308	15
pretest	7.5333	2.85023	15
ethnic	.5333	.51640	15
eth_pre Interaction of Ethnicity x Pretest	3.8000	4.09180	15

The interaction term..

Pretest* Ethn Int Term

4.00	.00	.00
7.00	.00	.00
6.00	.00	.00
8.00	.00	.00
6.00	.00	.00
12.00	.00	.00
13.00	.00	.00
10.00	1.00	10.00
4.00	1.00	4.00
5.00	1.00	5.00
7.00	1.00	7.00
8.00	1.00	8.00
4.00	1.00	4.00
10.00	1.00	10.00
9.00	1.00	9.00

Correlations

		perform	pretest	ethnic	eth_pre Interaction of Ethnicity x Pretest
Pearson Correlation	perform	1.000	.614	-.156	-.006
	pretest	.614	1.000	-.159	.132
	ethnic	-.156	-.159	1.000	.899
	eth_pre Interaction of Ethnicity x Pretest	-.006	.132	.899	1.000
Sig. (1-tailed)	perform	.	.007	.289	.492
	pretest	.007	.	.286	.319
	ethnic	.289	.286	.	.000
	eth_pre Interaction of Ethnicity x Pretest	.492	.319	.000	.
N	perform	15	15	15	15
	pretest	15	15	15	15
	ethnic	15	15	15	15
	eth_pre Interaction of Ethnicity x Pretest	15	15	15	15

* Review recommendation to center continuous level predictors 38

Moderated Multiple Regression

For MMR enter predictors at two steps:

- 1) the predictors (ethnicity and pretest)
- 2) the interaction term (ethnicity x pretest)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.616 ^a	.380	.277	3.71925	.380	3.677	2	12	.057
2	.623 ^b	.389	.222	3.85764	.009	.154	1	11	.702

a. Predictors: (Constant), ethnic, pretest

b. Predictors: (Constant), ethnic, pretest, eth_pre Interaction of Ethnicity x Pretest

And does the model show an appreciable improvement in fit when incorporating the interaction term.....even though this is a very small sample size, we see that the R square goes from 38% (individual predictors only) to 38.9% when adding the interaction; so with $p = .702$ for the R square change, we do not have evidence (in the context of significance testing) for bias.....however, we would not generally run such a model with $n = 15!!$

Moderated Multiple Regression

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	10.158	3.156		3.218	.007			
	pretest	.927	.353	.604	2.624	.022	.614	.604	.596
	ethnic	-.511	1.950	-.060	-.262	.798	-.156	-.075	-.060
2	(Constant)	9.208	4.069		2.263	.045			
	pretest	1.045	.475	.681	2.202	.050	.614	.553	.519
	ethnic	1.684	5.937	.199	.284	.782	-.156	.085	.067
	eth_pre Interaction of Ethnicity x Pretest	-.293	.746	-.274	-.393	.702	-.006	-.118	-.093

a. Dependent Variable: perform

Significance for Pretest

Nonsignificance for Ethnicity (the moderator) and the interaction term

*Discuss if $p = .053$, and terms such as “marginally significant”

Differences between slopes

❖ Can also extract very similar results as MMR by taking difference between slopes as shown below...

Coefficients^a

ethnic	Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
			B	Std. Error	Beta		
.00 white	1	(Constant)	9.208	3.042		3.027	.029
		pretest	1.045	.355	.796	2.945	.032
1.00 Latino	1	(Constant)	10.891	5.057		2.154	.075
		pretest	.752	.673	.415	1.117	.307

a. Dependent Variable: perform

95% CI = .293 +/- [2.16 x .7609]
 = [-1.35, 1.94]..since '0' included,
 conclude no significant difference in
 slopes..plus, beyond *p*-value can
 examine margin of error

Slope for Whites Slope for Latinos

$$t = (1.045 - .752) / (\sqrt{.355^2 + .673^2}) = .293 / .7609 = .385$$

And using *t*-Probability Density Function with df = 13

Compute *t* density=pdf.T(.385, 13).

.....we get a two-tailed *p*-value of .723.....mainly due to rounding error

this is the same as MMR (t = -.393, p = .702) !!! So, again, not significant differences in slopes (but there is 'practical' significance.....)

Multiple Testing...a caveat!!

- Give me the opportunity to do enough poking around a database, I can almost guarantee you I will find something significant!!
- For example, assuming you set your level of significance at the usual level of .05 (meaning you are OK with Type I error rate of 5/100 occasions of testing saying there are differences when there are not)
-and say you have 4 groups you want to potentially compare for bias.....well, you have $4(4-1)/2 = 6$ combinations of pairwise comparisons (e.g., Latinos vs. Caucasians, Caucasians vs. Filipinos, etc)

Multiple Testing...a caveat!!

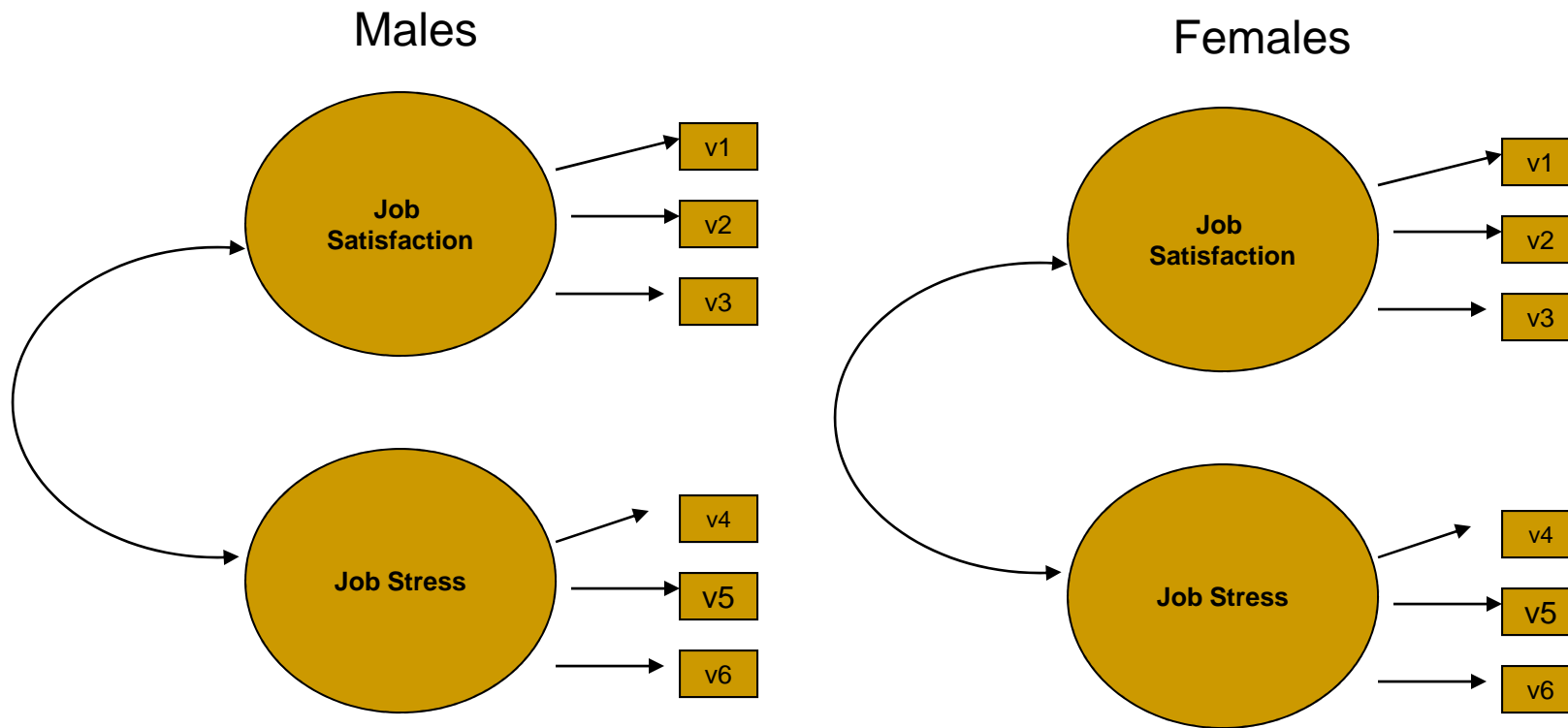
- Well, given $1 - (1 - \alpha)^k$ where α is your nominal level of significance (in this case, $\alpha = .05$) and $k = \#$ of pairwise comparisons ($k = 6$), then your Type I error rate goes from your 'allowable' .05 to .265!!
- So this means that 26.5 times out of 100 you will wrongly reject the null when the null is true (i.e., 26.5 times out of 100 you will say there is a difference when there isn't)
- Which means be very, very cautious about analysts that seem to be doing a lot of poking and prodding without paying heed to inflated Type I error rate.....you may arrive at some very wrong conclusions (in medical parlance, too many false positives)
 - e.g., in your search for 'bias' you compare majority/minority groups on 20 tests.....you're bound to find significance somewhere!!!

For recent 'state of the union' comments about measurement issues including Type I/II errors see Wainer (National Board of Medical Examiners), 2010

Construct Validity

- Is there the possibility that your test/items has different meaning(s) to different groups?
 - The onus is on the instrument developer/user to verify that the item/tests mean the same thing to different groups...e.g., assuming the following items measure the construct of 'organizational stress':
 - I am overwhelmed with the amount of work my job requires
 - I find my job to be too fast paced
 - I never have enough time to finish my tasks
 - I frequently feel uptight at work
- Maybe those four items (or any combination thereof) may measure dissimilar constructs across different groupings....

...tests of factorial invariance



Via confirmatory factor analytic procedures (Byrne, 1994), can test if there is “invariance” of the given instrument/inventory..... i.e., does the test exhibit similar structure/meaning across groups?

Differential Item Functioning

- Differential item functioning (DIF) occurs when people from different groups (commonly gender or ethnicity) with the same latent trait (ability/skill) have a different probability of giving a certain response on a questionnaire or test
- DIF analysis provides an indication of unexpected behavior by item on a test. An item does not display DIF if people from different groups have a different probability to give a certain response; it displays DIF if people from different groups of same underlying true ability have a different probability to give a certain response
- More precisely, an item displays DIF when the difficulty level (b), the discrimination (a) or the lower asymptotes (c) - estimated by item response theory (IRT)- of an item differs across groups

http://en.wikipedia.org/wiki/Differential_item_functioning

Differential Item Functioning

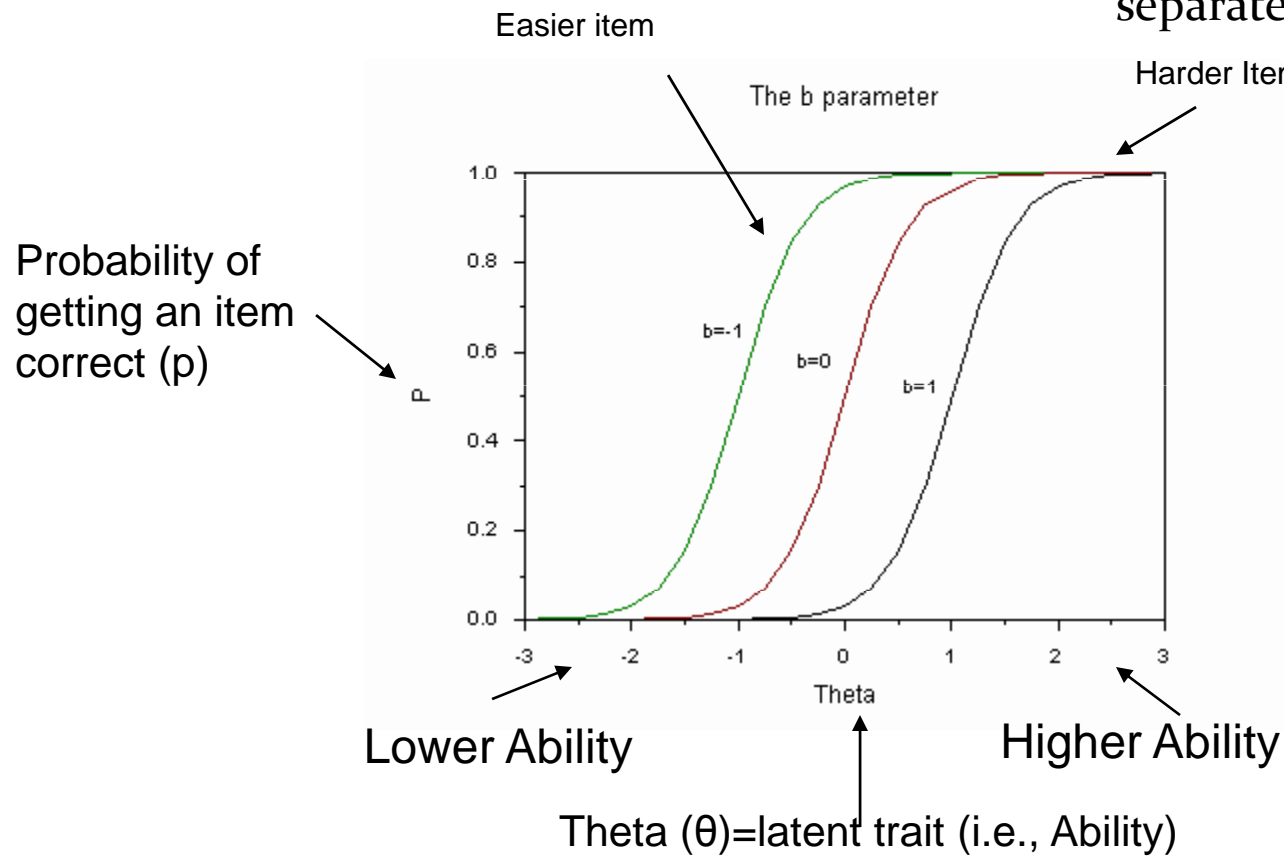
- Was created by Educational Testing Service (ETS).....(GRE, SAT, LSAT)
- DIF analysis attempts to identify items that are specifically biased against any ethnic, racial or gender group (Kaplan & Saccuzzo, 2009)

Item Response Theory (Applications)

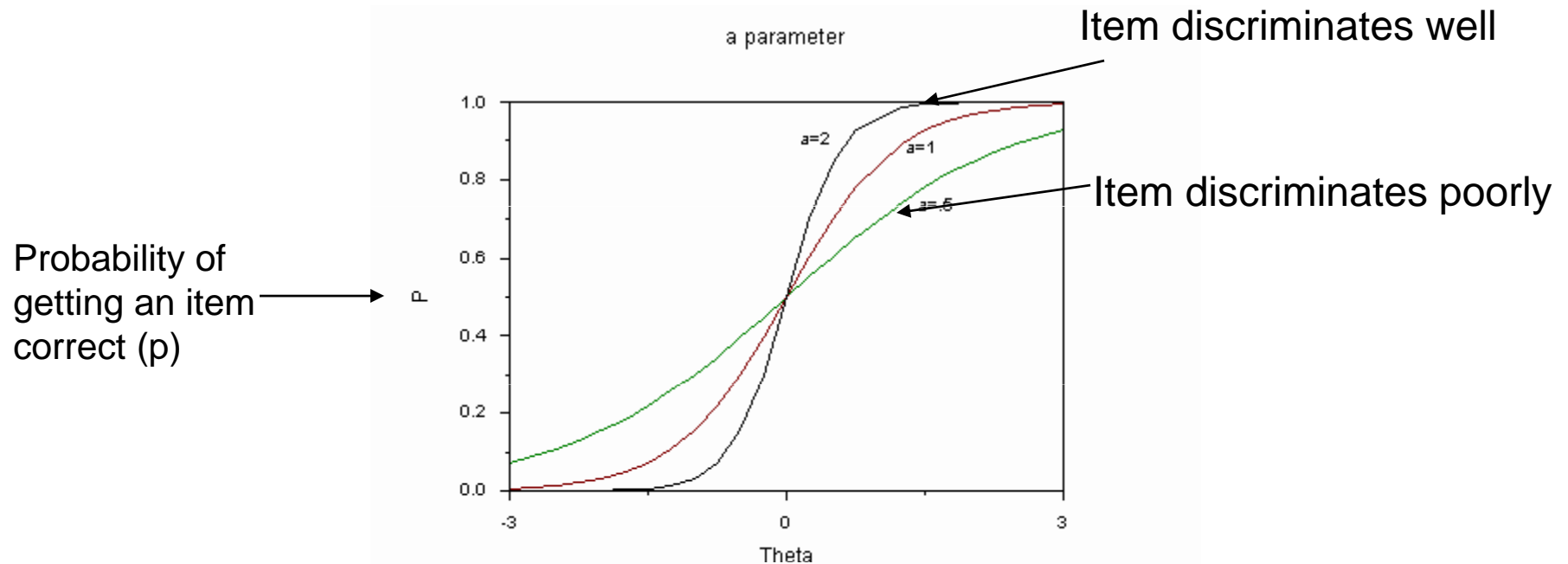
- Item bias analysis--IRT provides a test of item equivalence across groups.....can test whether an item is behaving differently for blacks and whites or for males and females, for example. The same logic can be applied to translations of attitude scales into different languages....can test whether the item means the same thing in English and French, for example.
- Equating--Sometimes we have scores on one test and we would like to know what the equivalent score would be on another test (e.g., versions or forms of the SAT). IRT provides a theoretical justification for equating scores from one test to another. [classic text by Kolen and Brennan, 2004]
- Tailored Testing--IRT provides an estimate of the true score that is not based on the number of correct items. This frees us to give different people different test items but still place people on the same scale. One particularly exciting feature of tailored testing is the capability to give people test items that are matched (close) to them. A tailored testing program for the SAT will give more difficult items to brighter test takers. This also has implications for test security -- different people get different tests.
 - <http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm>

Item Characteristic Curve (difficulty parameter)

For now, assume these are three separate items.....



Item Response Theory (discrimination parameter)



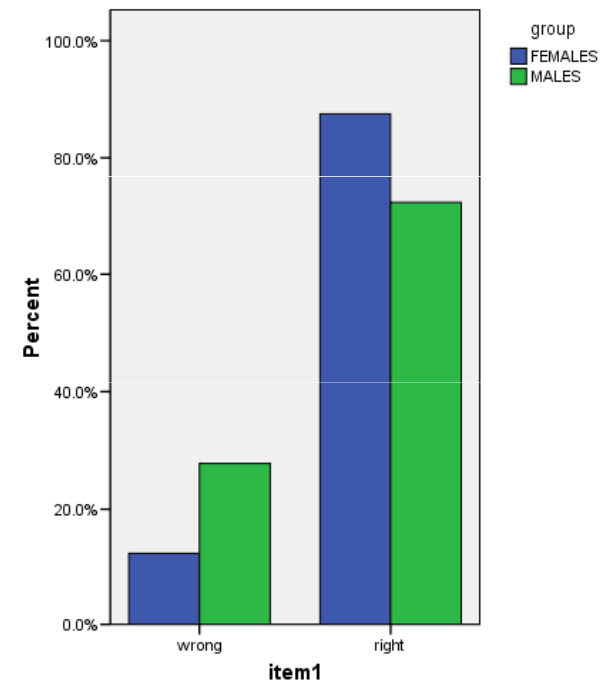
Discriminability parameter = slope
.....the steeper the s-curve (i.e., logistic curve) the better the item discriminates

DIF via Item Response Theory (IRT)....

item1

group				Frequency	Percent	Valid Percent	Cumulative Percent
1 FEMALE	Valid	0	3	12.5	12.5	12.5	
		1	21	87.5	87.5	100.0	
		Total	24	100.0	100.0		
2 MALES	Valid	0	5	27.8	27.8	27.8	
		1	13	72.2	72.2	100.0	
		Total	18	100.0	100.0		

Notice difference for item 1.....
✓87.5% of Females get item right.....
✓72.2% of Males get item right



IRT via BILOG.....

bias_dif

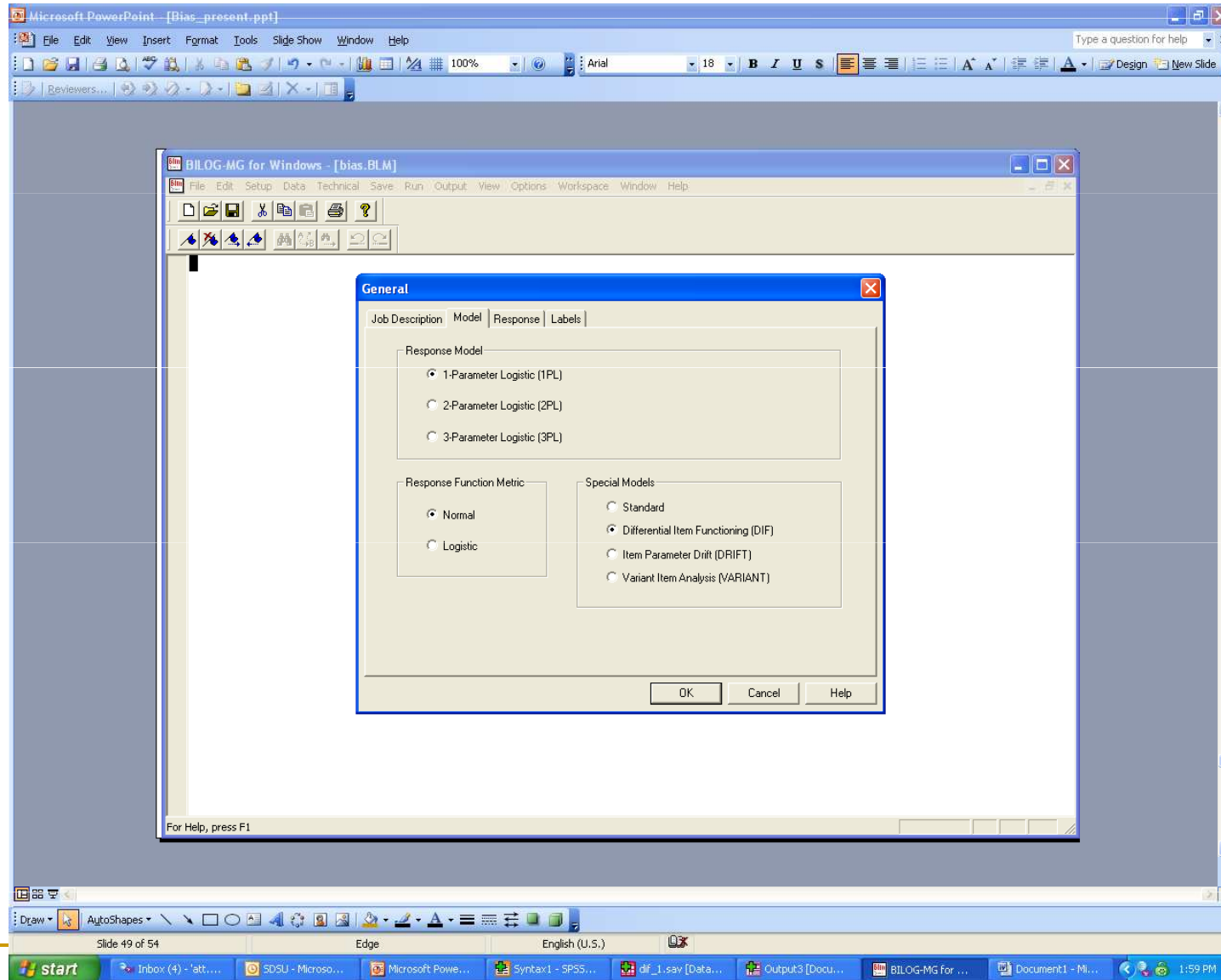
```
>GLOBAL DFName = 'C:\BIAS.dat',
      NPArm = 1;
>LENGTH NITems = (3);
>INPUT NTOtal = 3,
      NALt = 2,
      NIDchar = 2,
      NGRoup = 2,
      DIF;
>ITEMS ;
>TEST1 TName = 'TEST0001',
      INumber = (1(1)3);
>GROUP1 GName = 'GROUP001',
      LENgth = 3,
      INumbers = (1(1)3);
>GROUP2 GName = 'GROUP002',
      LENgth = 3,
      INumbers = (1(1)3);
(2A1, 11, 3A1)
>CALIB GROup-plots;
>SCORE FIT,
      MOMents;
```

females

males

Remember SPSS
mainframe?!!

IRT via BILOG.....



IRT via BILOG.....

ITEM STATISTICS FOR GROUP: 1 **Females**

ITEM	NAME	#TRIED	#RIGHT	PCT	ITEM*TEST CORRELATION		
					LOGIT/1.7	PEARSON	BISERIAL
1	ITEM0001	24.0	21.0	0.875	-1.14	0.225	0.361
2	ITEM0002	24.0	15.0	0.625	-0.30	0.685	0.874
3	ITEM0003	24.0	14.0	0.583	-0.20	0.655	0.827

ITEM STATISTICS FOR GROUP: 2 **Males**

ITEM	NAME	#TRIED	#RIGHT	PCT	ITEM*TEST CORRELATION		
					LOGIT/1.7	PEARSON	BISERIAL
1	ITEM0001	18.0	13.0	0.722	-0.56	0.207	0.276
2	ITEM0002	18.0	16.0	0.889	-1.22	0.366	0.607
3	ITEM0003	18.0	16.0	0.889	-1.22	0.366	0.607

IRT via BILOG.....Starting with one parameter logistic model (1PL = equal slopes)

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING

GROUP 1 GROUP001; ITEM PARAMETERS AFTER CYCLE 14

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
ITEM0001	2.189 0.628*	1.528 0.500*	-1.432 0.411*	0.837 0.274*	0.000 0.000*
ITEM0002	0.698 0.586*	1.528 0.500*	-0.457 0.383*	0.837 0.274*	0.000 0.000*
ITEM0003	0.462 0.544*	1.528 0.500*	-0.303 0.356*	0.837 0.274*	0.000 0.000*

* STANDARD ERROR

LARGEST CHANGE = 0.017674

GROUP 2 GROUP002; ITEM PARAMETERS AFTER CYCLE 14

ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE
ITEM0001	0.883 0.382*	1.528 0.500*	-0.578 0.250*	0.837 0.274*	0.000 0.000*
ITEM0002	1.871 0.315*	1.528 0.500*	-1.224 0.206*	0.837 0.274*	0.000 0.000*
ITEM0003	1.868 0.639*	1.528 0.500*	-1.222 0.418*	0.837 0.274*	0.000 0.000*

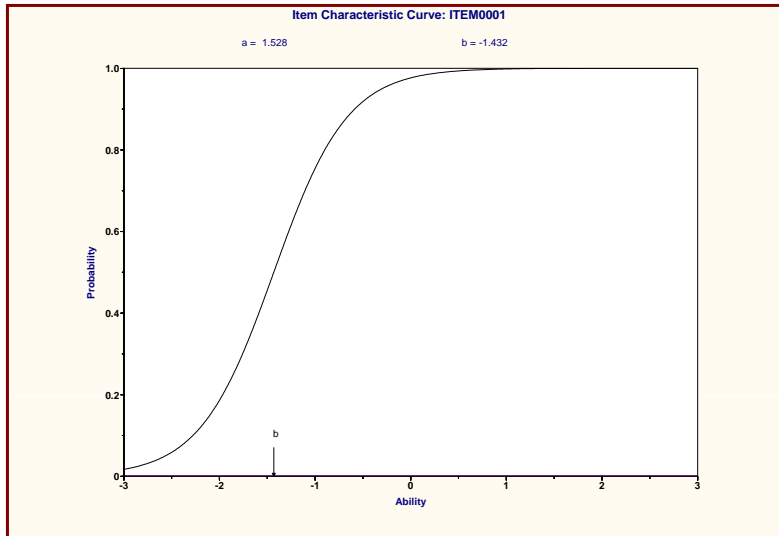
* STANDARD ERROR

LARGEST CHANGE = 0.017674

Lower Threshold means easier item for Females ...and notice next two items easier for males

NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE FOR TESTS WITH LESS THAN 20 ITEMS

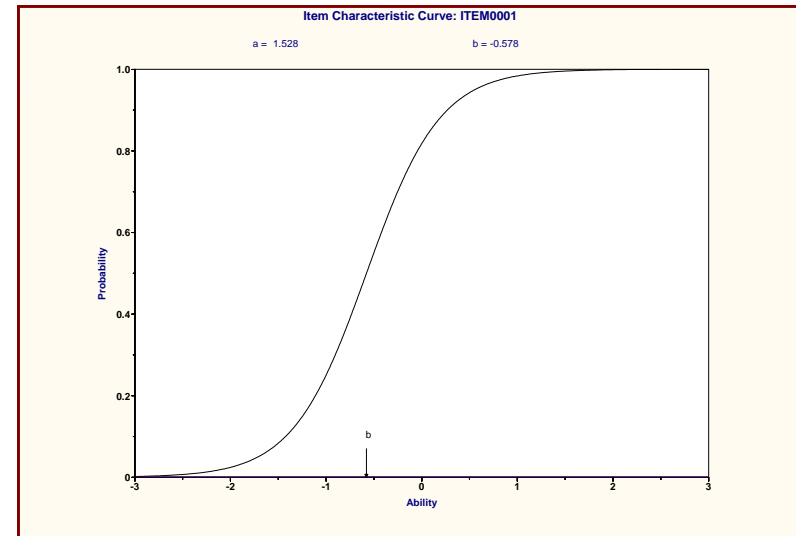
IRT via BILOG for 1PL.....



Selecting item 1.....

← Females

Males



Do they look different?

Note: discriminability (slope) for 1PL is constrained to equality, but it is an easier item for Males (difficulty free to vary)

IRT via BILOG.....2PL (now slopes to differ)

MODEL FOR GROUP DIFFERENTIAL ITEM FUNCTIONING

GROUP	1	GROUP001; ITEM PARAMETERS AFTER CYCLE				9
ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE	
ITEM0001	1.669	0.976	-1.710	0.699	0.000	
	0.461*	0.316*	0.569*	0.226*	0.000*	
ITEM0002	0.687	1.833	-0.375	0.878	0.000	
	0.621*	0.831*	0.342*	0.398*	0.000*	
ITEM0003	0.389	1.776	-0.219	0.871	0.000	
	0.563*	0.850*	0.330*	0.417*	0.000*	

* STANDARD ERROR

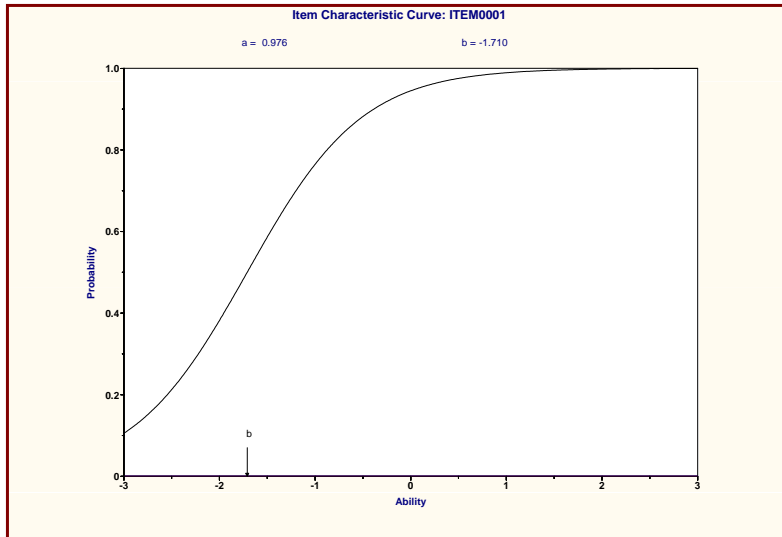
LARGEST CHANGE = 0.014627

GROUP	2	GROUP002; ITEM PARAMETERS AFTER CYCLE				9
ITEM	INTERCEPT	SLOPE	THRESHOLD	LOADING	ASYMPTOTE	
ITEM0001	0.658	0.976	-0.674	0.699	0.000	
	0.361*	0.316*	0.405*	0.226*	0.000*	
ITEM0002	1.904	1.833	-1.039	0.878	0.000	
	0.356*	0.831*	0.493*	0.398*	0.000*	
ITEM0003	1.865	1.776	-1.050	0.871	0.000	
	0.718*	0.850*	0.528*	0.417*	0.000*	

* STANDARD ERROR

LARGEST CHANGE = 0.014627

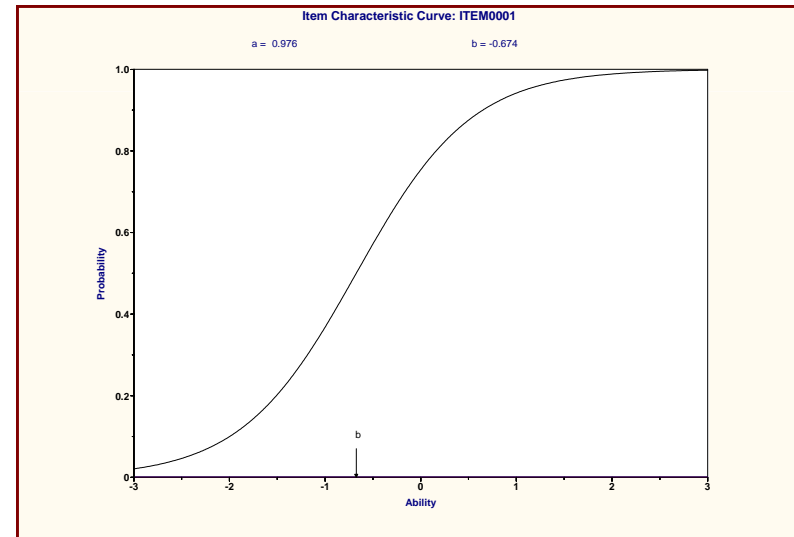
NOTE: ITEM FIT CHI-SQUARES AND THEIR SUMS MAY BE UNRELIABLE
FOR TESTS WITH LESS THAN 20 ITEMS



Selecting item 1.....

← females

Males



Do they look different now that slopes free to vary?

And 2PL in Mplus (one group).....

[note: couldn't obtain identification with three items, but able to do so once increased to k = 5 items]

- Mplus VERSION 5.21
- MUTHEN & MUTHEN
- 01/25/2010 3:36 PM

- INPUT INSTRUCTIONS

- TITLE: two-parameter logistic item response theory (IRT) model
- DATA: FILE IS 'C:\bias_mplus.dat';
- VARIABLE: NAMES ARE ID grp item1 item2 item3 item4 item5;
- CATEGORICAL ARE item1 item2 item3 item4 item5;
- Usevariables are item1 item2 item3 item4 item5;
- ANALYSIS: ESTIMATOR = MLR; ****uses numerical integration****
- MODEL: f BY item1-item5*;
- f@1;
- OUTPUT: TECH1 TECH8;
- PLOT: TYPE = PLOT3;

2PL in Mplus

SUMMARY OF CATEGORICAL DATA PROPORTIONS

ITEM1

Category 1 0.190

Category 2 0.810

ITEM2

Category 1 0.262

Category 2 0.738

ITEM3

Category 1 0.286

Category 2 0.714

ITEM4

Category 1 0.452

Category 2 0.548

ITEM5

Category 1 0.357

Category 2 0.643

2PL in Mplus

MODEL RESULTS

		Two-Tailed			
		Estimate	S.E.	Est./S.E.	P-Value
F	BY				
	ITEM1	-0.047	0.669	-0.071	0.944
	ITEM2	8.792	5.605	1.569	0.117
	ITEM3	3.896	2.166	1.798	0.072
	ITEM4	2.757	1.545	1.784	0.074
	ITEM5	0.279	0.428	0.651	0.515

slopes



Thresholds					
	ITEM1\$1	-1.448	0.393	-3.680	0.000
	ITEM2\$1	-5.705	1.587	-3.595	0.000
	ITEM3\$1	-2.418	1.713	-1.412	0.158
	ITEM4\$1	-0.387	0.903	-0.429	0.668
	ITEM5\$1	-0.597	0.336	-1.776	0.076

Thresholds (difficulty)



2PL in Mplus

IRT PARAMETERIZATION IN TWO-PARAMETER LOGISTIC METRIC
WHERE THE LOGIT IS $1.7 \cdot \text{DISCRIMINATION} \cdot (\text{THETA} - \text{DIFFICULTY})$

Item Discriminations

F	BY				
	ITEM1	-0.028	0.394	-0.071	0.944
	ITEM2	5.172	3.297	1.569	0.117
	ITEM3	2.292	1.274	1.798	0.072
	ITEM4	1.622	0.909	1.784	0.074
	ITEM5	0.164	0.252	0.651	0.515

Slopes (discrimination)



Item Difficulties

	ITEM1\$1	30.684	434.554	0.071	0.944
	ITEM2\$1	-0.649	0.446	-1.453	0.146
	ITEM3\$1	-0.621	0.384	-1.614	0.106
	ITEM4\$1	-0.141	0.322	-0.436	0.663
	ITEM5\$1	-2.141	3.279	-0.653	0.514

Thresholds (difficulty)



Variances

F	1.000	0.000	0.000	1.000
---	-------	-------	-------	-------

2PL in Mplus

Information Criteria

Number of Free Parameters	10
Akaike (AIC)	236.249
Bayesian (BIC)	253.626
Sample-Size Adjusted BIC	222.311
$(n^* = (n + 2) / 24)$	

Chi-Square Test of Model Fit for the Binary and Ordered Categorical (Ordinal) Outcomes

Pearson Chi-Square

Value	40.046
Degrees of Freedom	21
P-Value	0.0073

Likelihood Ratio Chi-Square

Value	30.340
Degrees of Freedom	21
P-Value	0.0854

Though this analysis is for entire group, you could conduct DIF using multigroup SEM option (and set certain parameters, such as slopes and intercepts to invariance) and see if model fit improves via information theoretic indices (lower is better) and chi-square tests.....(Woods, 2009)

.....last thoughts

- So, statistical analysis won't be the last word on furnishing unequivocal evidence of bias, for as we have seen there are no universally agreed-upon definitions of bias in a testing context
- However, at least it provides an empirically-based platform that goes beyond value judgment

References

- Aguinis, H. (2004). *Regression analysis for categorical moderators*. NY: Guilford.
- Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks, CA: Sage.
- Cohen, J., Cohen, P. West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum.....
- Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cohen (1988). *Statistical power analysis for the behavioral sciences*. (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, R. J., & Swerdlik, M. E. (2010). *Psychological testing and assessment*. (7th Ed). NY: McGraw Hill.
- Hunter, J. E., & Schmidt, F. L. (1976). Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychological Bulletin*, 83, 1053-1071.
- Kaplan, R. M., & Saccuzzo, D. P. (2009). *Psychological testing*. (7th Ed). Belmont, CA: Wadsworth.
- Kolen, M., J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. (2nd Ed.). NY: Springer.
- Meade, A. W., & Fetzner, M. (2009). Test bias, differential prediction, and a revised approach for determining the suitability of a predictor in a selection context. *Organizational Research Methods*, 12, 738-761.
- Murphy, K. R., & Davidshofer, C. O. (2005). *Psychological Testing: Principles and Applications* (6th Edition). Upper Saddle River, NJ: Prentice Hall.
- Nunnally, J. C., & Bernstein, I. H. *Psychometric theory*. (1994). (3rd Ed.). NY: McGraw-Hill.
- Sackett, P. R., Laczko, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors: The omitted variables problem, *Journal of Applied Psychology*, 88, 1046-1056).
- Wainer, H. (2010). 14 conversations about three things. *Journal of Educational and Behavioral Statistics*, 35, 5-25.
- Woods, H. (2009). Evaluation of MIMIC-Model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, 44, 1-27.