Differential Item Functioning

Lecture #11

ICPSR Item Response Theory Workshop

Lecture Overview

- Detection of Differential Item Functioning (DIF)
  - Distinguish “Bias” from “DIF”
  - Test vs. Item-level bias

- Revisit some test score equating issues

DIFFERENTIAL ITEM FUNCTIONING

Outline

- Review relevant assumptions
- Concept of potentially biased test items (or “DIF,” as we’ll call it)
- IRT-based methods of detecting DIF

Some Notes

- We will focus on:
  - DIF with 0-1 test items
  - DIF with polytomous items is more complicated, though similar approaches are used
  - IRT methods only
  - DIF as a statistical issue
    - Interpretation of “why?” can be quite a bit trickier

Why study DIF?

- We're often interested in comparing cultural, ethnic, or gender groups
- Meaningful comparisons require that measurement equivalence holds
- Classical test theory methods confound "bias" with true mean differences
  - Not IRT
- In IRT terminology, item/test "bias" is referred to as DIF/DTF

Definition of DIF

- A test item is labeled with “DIF” when people with equal ability, but from different groups, have an unequal probability of item success

- A test item is labeled as “non-DIF” if examinees having the same ability have equal probability of getting the item correct, regardless of group membership

DIF versus DTF

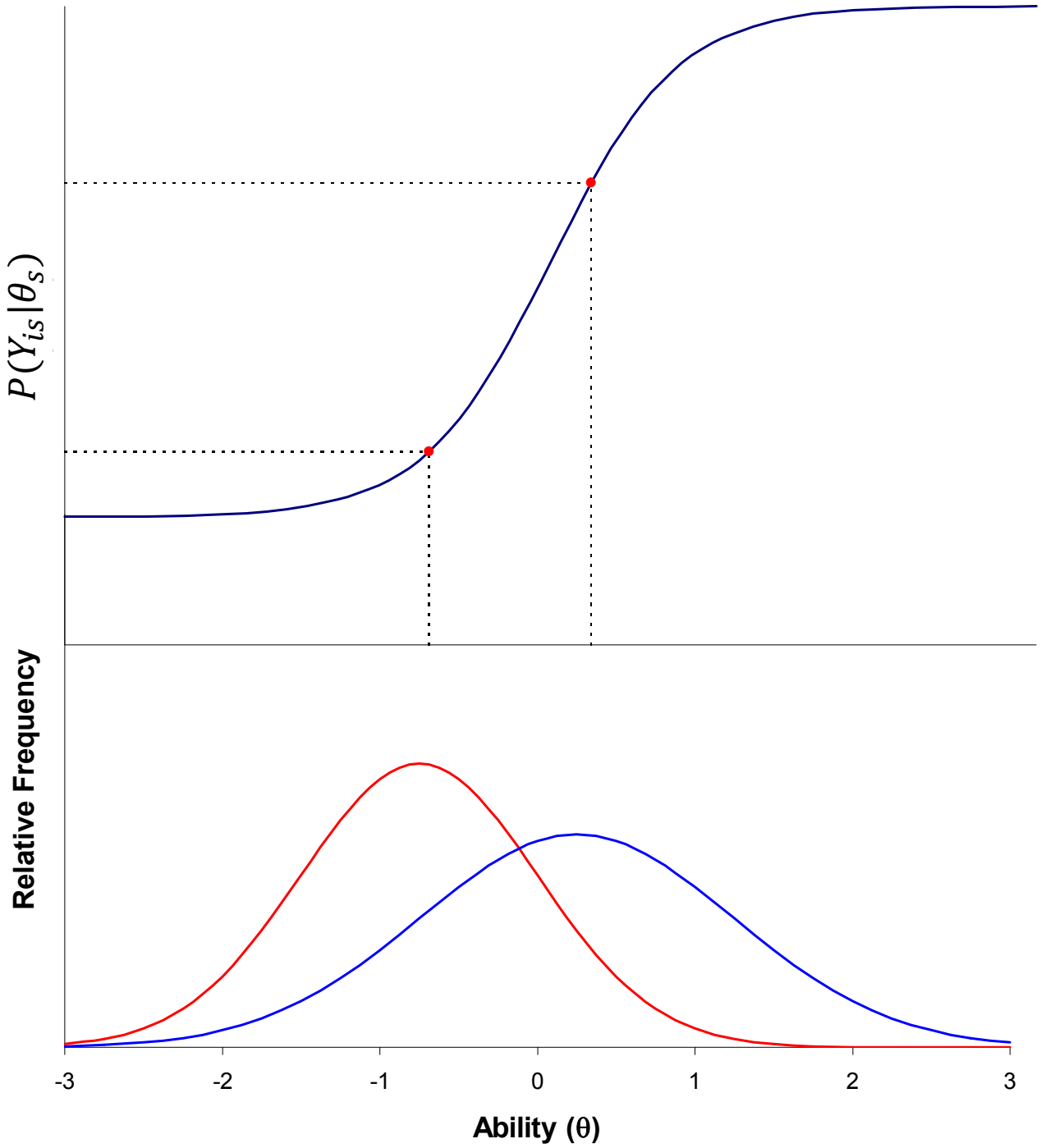
- DIF is basically found by examining differences in ICCs across groups

- Differential Test Functioning, or DTF, is the analogous procedure for determining differences in TCCs

- DTF is arguably more important because it speaks to impact
  - DIF in one item may be significant, but might not have too much practical impact on test results

Classical Approach to DIF

- Compare item p-values in the two groups
  - Criticism: p-value difference may be due to real and important group differences
  - Need to compare p-value difference for one item in relation to the differences for other items
  - This provides a baseline to interpret the difference
    - But item p-value differences will be confounded by discrimination differences


$P(Y_{is}|\theta_s)$
Relative Frequency
-3
-2
-1
0
1
2
3
Ability (θ)

Important Assumptions

- Unidimensionality
- Parameter Invariance
- Violations of these assumptions across identifiable groups basically provide us with a definition of DIF

Unidimensionality

- The test measures ONE construct
  - e.g., math proficiency, verbal ability, trait, etc.

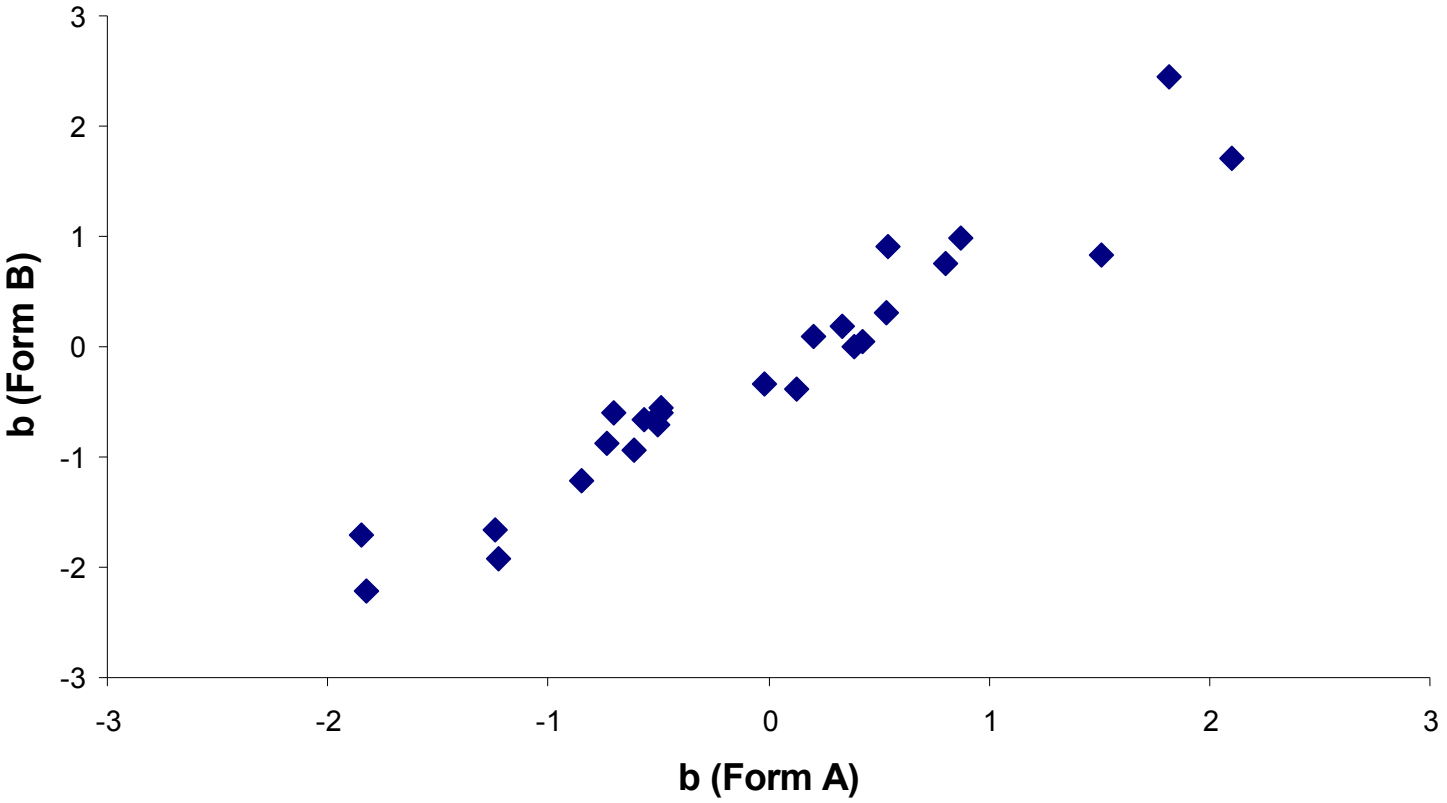
- Group membership (e.g., gender, ethnicity, high-low ability) should not differentially impact success on the test (or on an item)

- If group membership impacts or explains performance, then the assumptions of the model are not being met
  - DIF=Dimensionality (unintended)

Parameter Invariance

- Invariance of Item parameters (a,b,c)
  - Compare item statistics obtained in two or more groups (e.g., high and low performing groups; Ethnicity; Gender)
  - If the model fits, different samples should still produce close to the same item parameter estimates
  - Scatterplots of b-b, a-a, c-c based on one sample versus the other should be strongly linearly related
  - Relationship won't be perfect: sampling error does enter the picture
  - Those estimates far from the best-fit line represent a violation of invariance

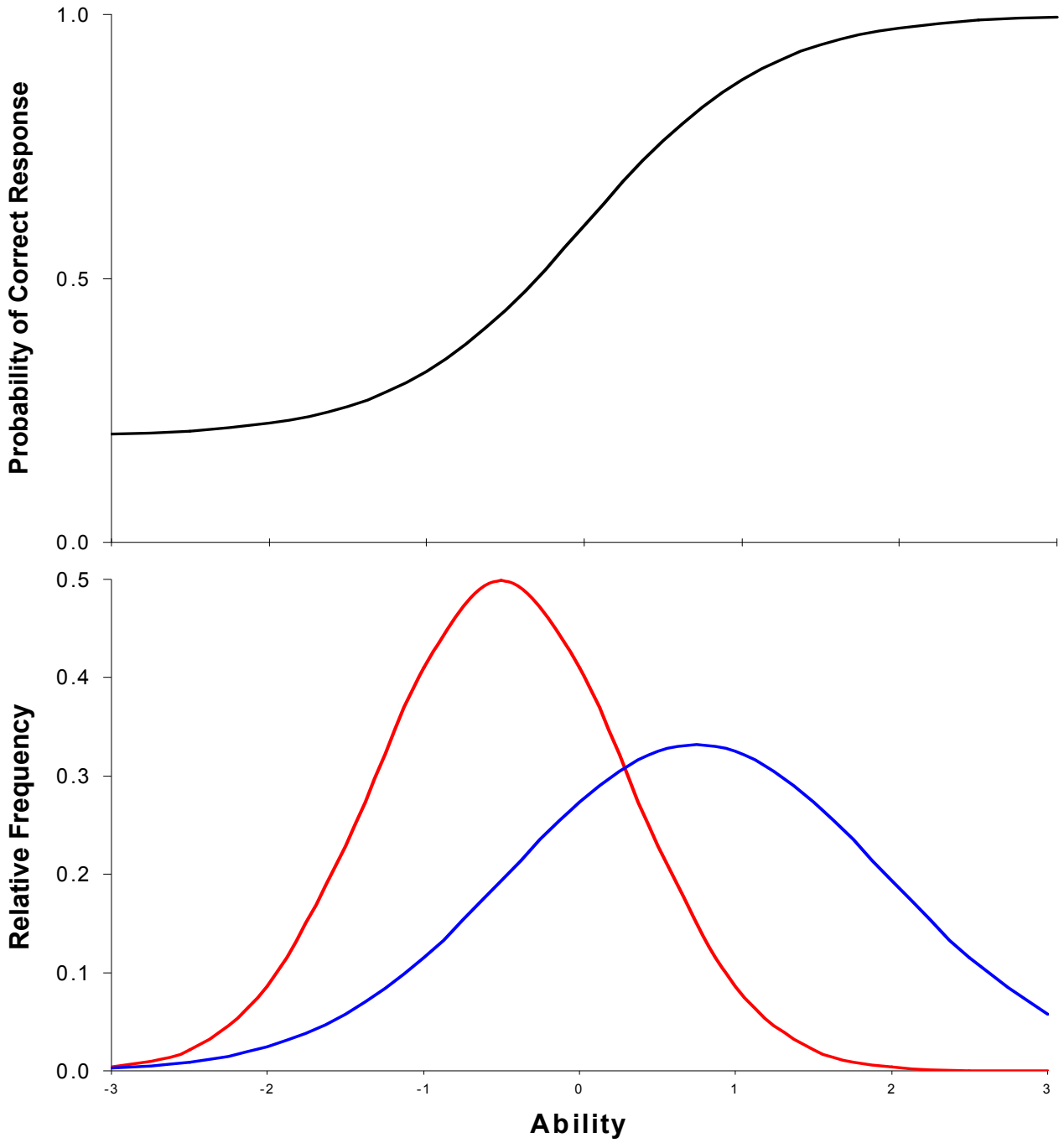
Parameter Invariance
b-plot



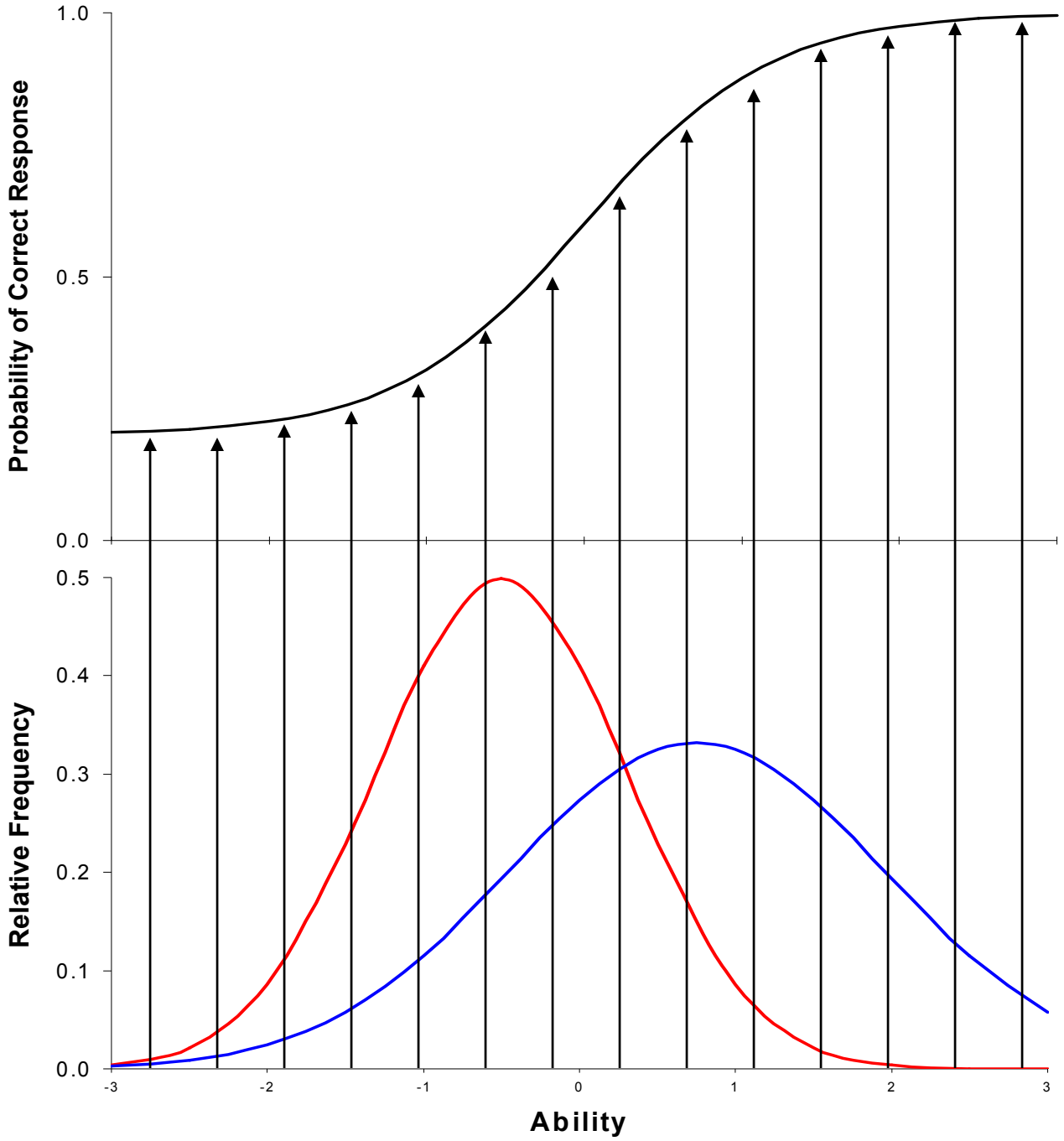
Check Item Invariance

- If…
  - The test is unidimensional -and-
  - The parameters are invariant across groups

- Then…
  - Even though there may be differences in the distributions of ability, similar (linearly related) results will be obtained for item parameter estimates



Two groups may have different distributions for the trait being measured, but the same model should fit



Frequencies may differ, but matched ability groups should have the same probability of success on the item

Check Item Invariance

- Differential Item Functioning (DIF)

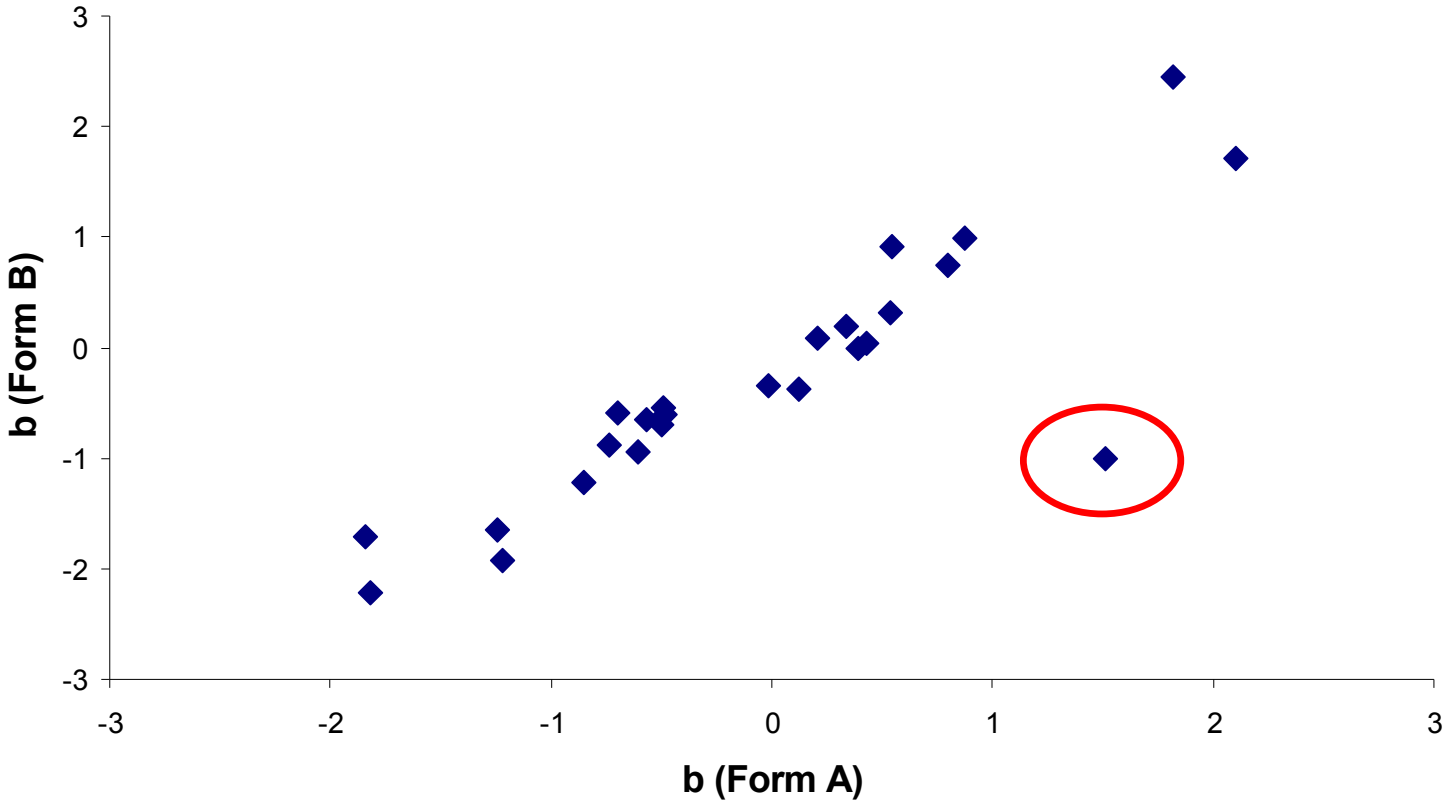
- This violates parameter invariance because items aren’t unidimensional

DIF in Practice

- "Reference" versus "Focal" groups
  - Males – Females
  - Majority – Minority
  - English – ESL groups

- One example of DIF is "item drift" due to over- or under-emphasis of content measured over time

May have to remove items: Parameter Drift



DIF in Practice

- "Differential Item Functioning" NOT "item bias" is the term used

- DIF is a statistical property, which states that matched-ability groups have differential probabilities of success on an item

- Bias is a substantive interpretation of any DIF that may be observed

DIF in Practice

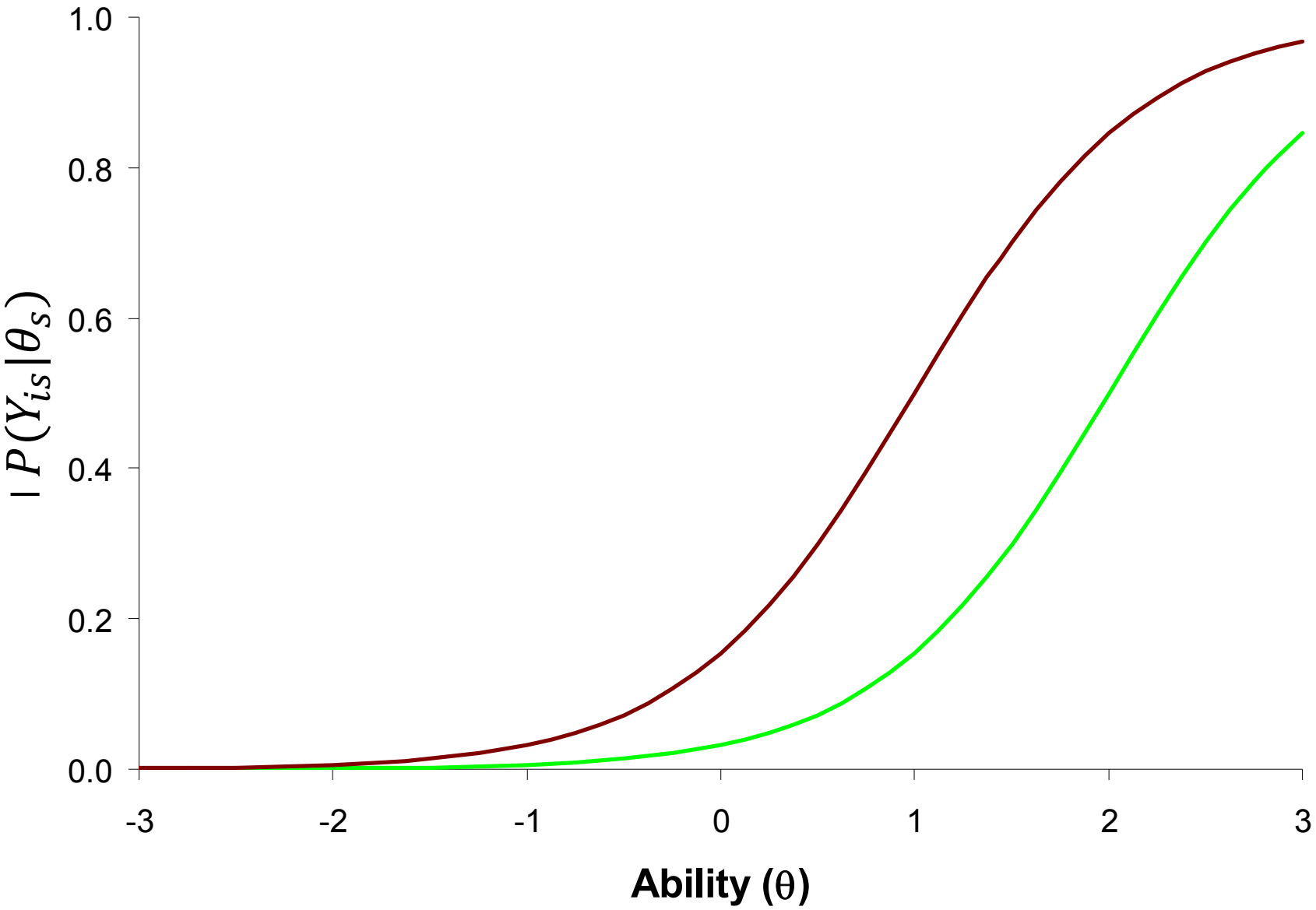
- DIF studies are absolutely essential in testing programs with high stakes

- Potential gender and/or ethnicity bias could negatively impact one or more groups in a construct irrelevant way
  - We're only interested in theta!

Identifying DIF

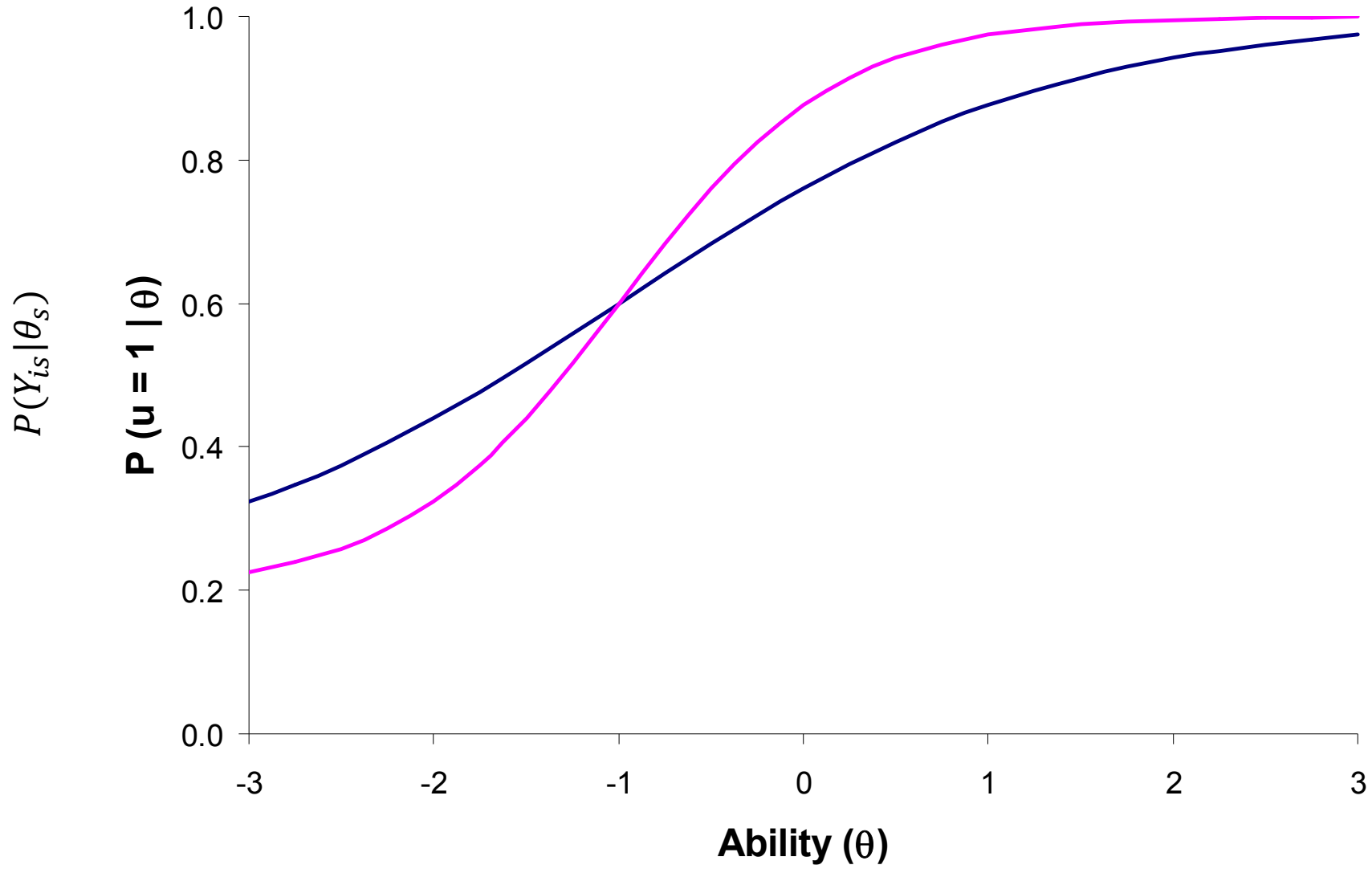
- Uniform DIF: item is systematically more difficult for members of one group, even after matching examinees on ability (theta)

- Cause: shift in b-parameter


1.0
0.8
0.6
0.4
0.2
0.0
$P(Y_{is}|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)

Identifying DIF

- Non-Uniform DIF: shift in item difficulty is not consistent across the ability continuum
- Increase/decrease in P for low-ability examinees is offset by the converse for high-ability examinees
- Cause: shift in a (and possibly b)


$P(Y_{is}|\theta_s)$
P (u = 1 | θ)
1.0
0.8
0.6
0.4
0.2
0.0
-3
-2
-1
0
1
2
3
Ability (θ)

Steps in a DIF Study

- Identify Reference and Focal groups of interest (usually two at a time)
- Design the DIF study to have samples which are large as possible
- Choose DIF statistics which are appropriate for the data
- Carry out the statistical analyses
- Interpret DIF stats and delete items or make item changes as necessary

Sample Size v. Effect Size

- CAUTION: Don't assume if DIF is found with statistical tests that there are definitely problems.
- With big samples, the power exists to detect even small conditional differences
- The flip side is that with small samples, sufficient power may not exist to identify truly problematic items

IRT DIF Analyses

- Combine data from both groups and obtain c-parameter estimates
- Fix the c-parameters and obtain a- and b-parameter estimates in each group from separate calibrations
- If multivariate DIF test is done, two completely separate calibrations may be conducted (not always)

IRT DIF Analyses

- Before the ICCs in the two groups can be compared, the must be placed onto the same scale (equated).

- Typically, item parameter estimates from the Focal group (minority) are placed onto the scale of the Reference group (majority)

Mean and Sigma Equating Method

- After separate calibrations, determine the linear transformation that matches the mean and SD of anchor item b-values across administrations:

$$x = \frac{\sigma_{b-FormA}}{\sigma_{b-FormB}} \qquad y = \mu_{b-FormA} - x\mu_{b-FormB}$$

- This transformation places the scale of item parameters from "Form B" onto the scale of item parameters from "Form A"

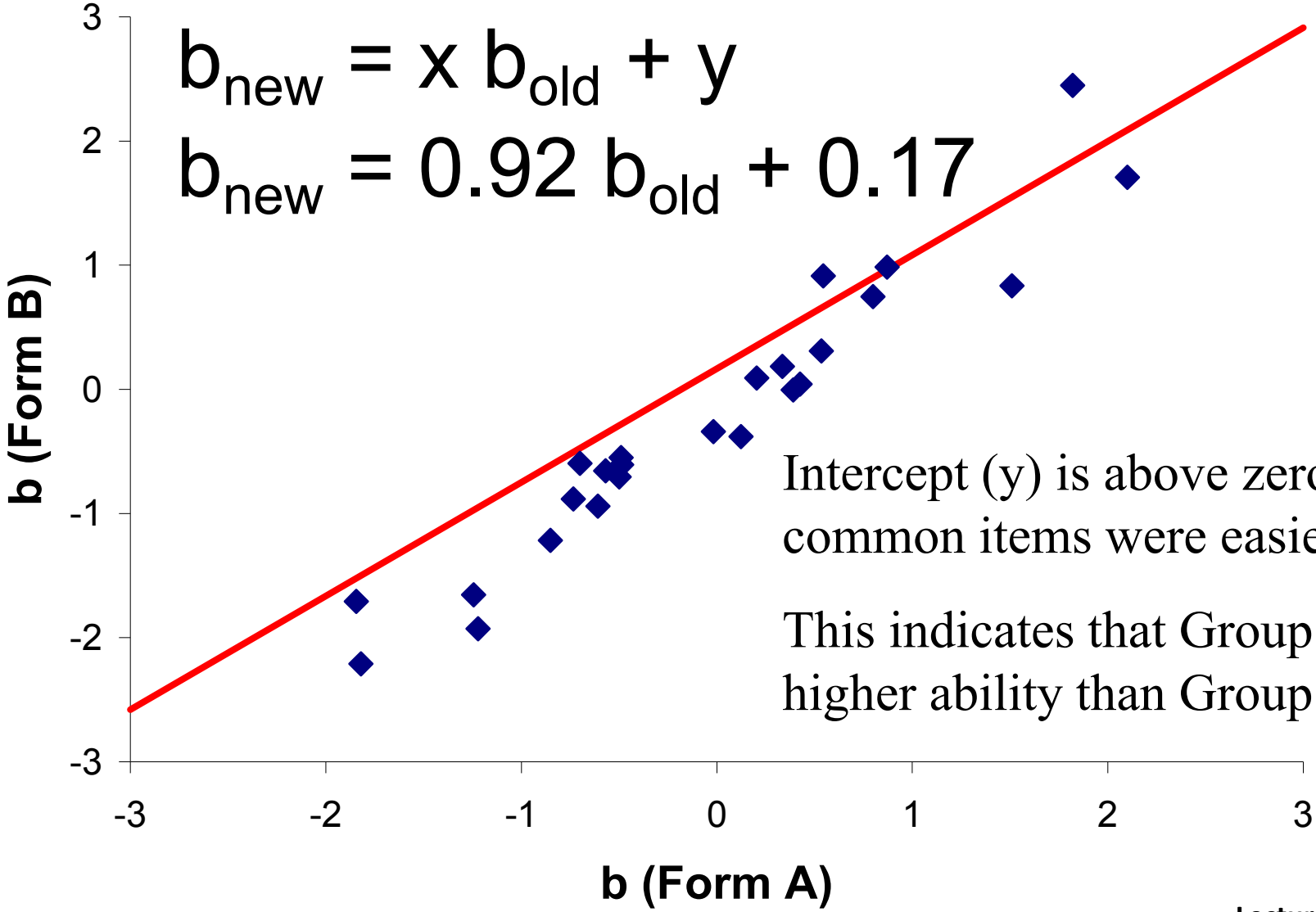
Mean & Sigma Example

- 25 Linking items
  - Form A: mb-FormA = -0.06, sb-FormA = 1.03
  - Form B: mb-FormB = -0.25, sb-FormB = 1.12
    - x = 1.03 / 1.12 = 0.92
    - y = -0.06 – (0.92 * -0.25) = 0.17
- $b_{new} = x\ b_{old} + y$
- $b_{new} = 0.92\ b_{old} + 0.17$

M & S Transformation of Form B linking b-parameters

$b_{new} = x\ b_{old} + y$

$b_{new} = 0.92\ b_{old} + 0.17$



Intercept (y) is above zero (i.e., common items were easier).

This indicates that Group B had higher ability than Group A.

if

then

$$\theta_{new} = x\theta + y$$

$$b_{new} = xb + y$$

$$a_{new} = \frac{a}{x}$$

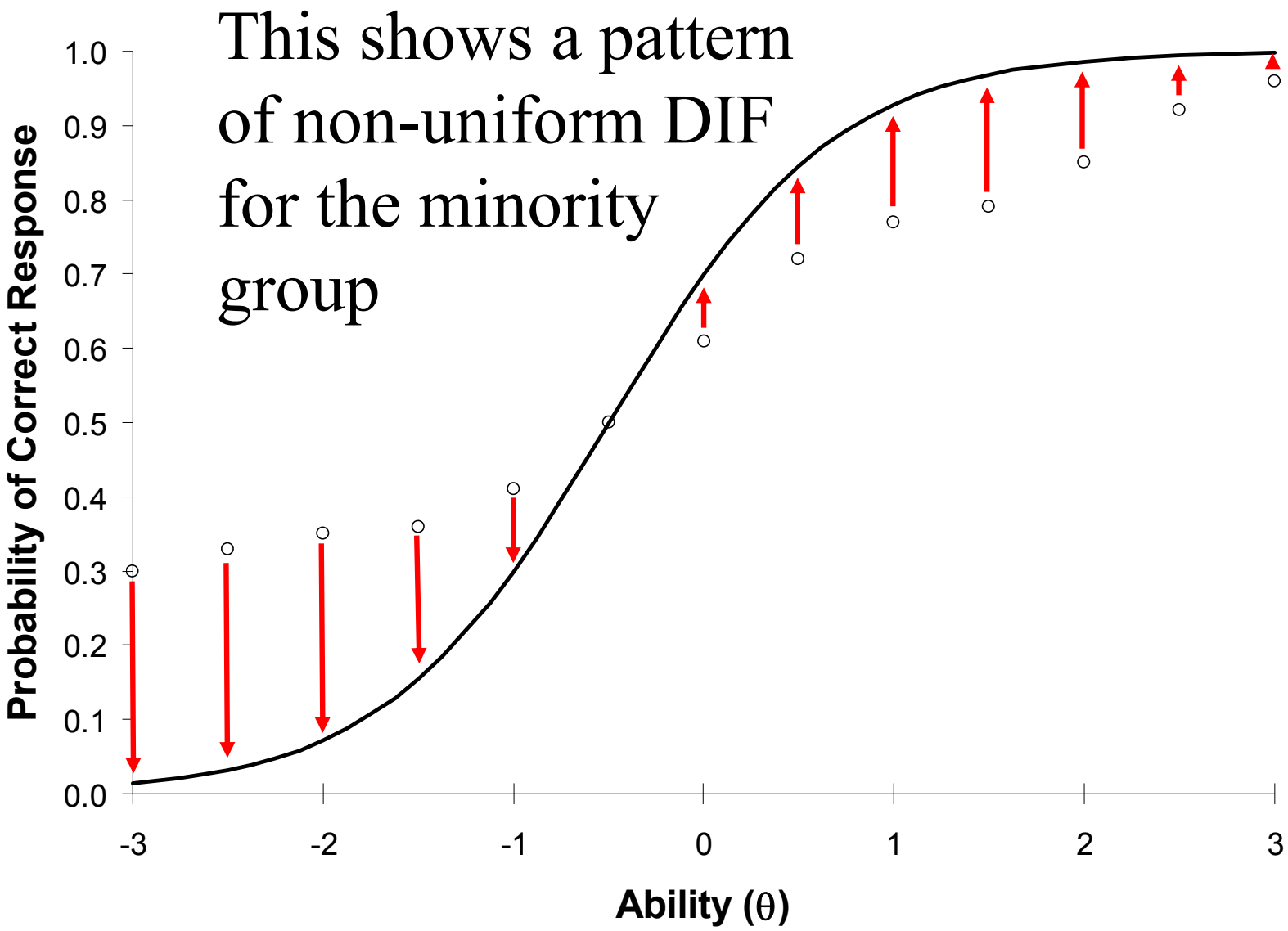
$$c_{new} = c$$

After item parameters are adjusted, the same transformation of b is done for all θ…now Group B will look more able (as they should).

Sample Size Issue

- We might also not have enough data to estimate two calibrations (e.g., the Focal N may be small)

- One solution is to estimate the ICCs with just the majority group, score the minority group with those parameters, and examine their “residuals” against the model-predicted ICC


This shows a pattern of non-uniform DIF for the minority group
Probability of Correct Response
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Ability (θ)

Important Note

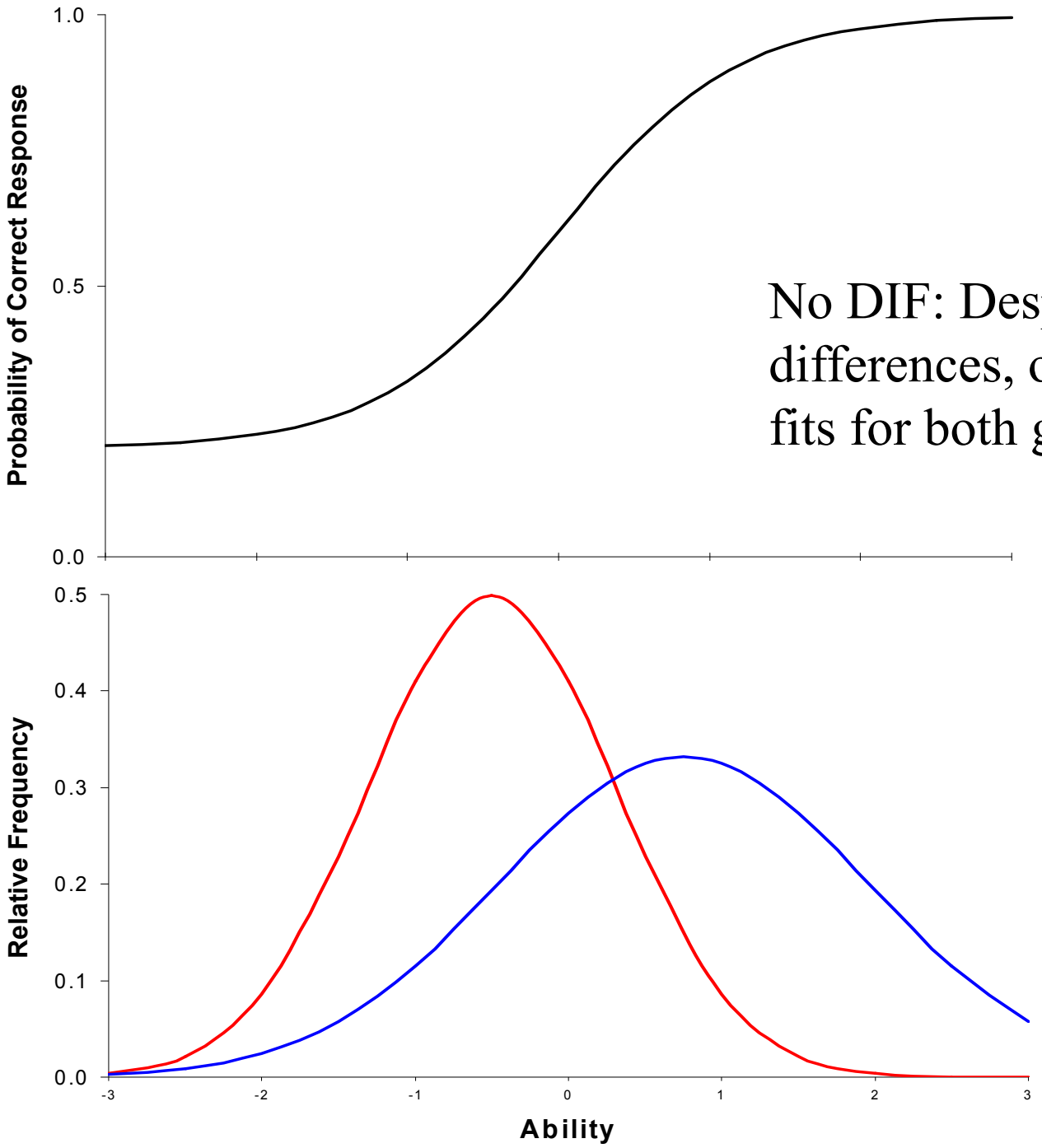
- Group differences do NOT (necessarily) mean DIF (or bias, for that matter)

- Two groups can have differences in their means, and if the model fits, the same ICCs will be estimated

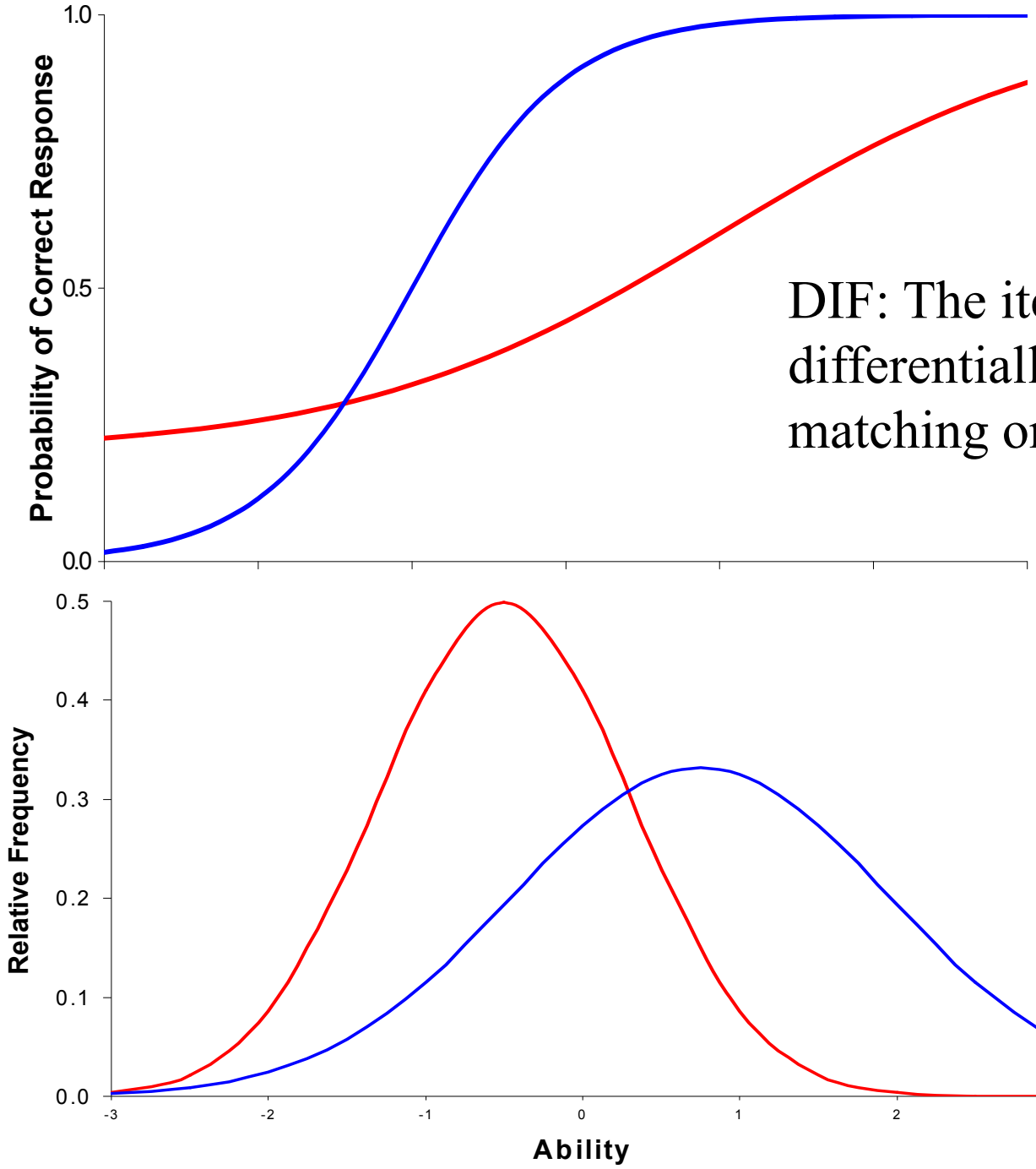
Important Note

- The point here is that if, after being matched on ability, different ICCs occur, then the item displays DIF

- If matched ability examinees have the same probability of success on the item, then there is no DIF, even if one group is smarter than the other



No DIF: Despite group differences, one ICC fits for both groups



DIF: The item performs differentially even after matching on ability

IRT DIF Analyses

- After equating parameters, compare the items across groups using one of the following methods we'll discuss
- Optional: delete items that have large DIF statistics, re-run analyses, and re-estimate equating constants
- Carefully review any flagged items for substantive review, interpretation

Two-stage Methods

- Conduct a DIF analysis
- Flag DIF items, temporarily remove them from criterion (raw score, θ)
- Re-evaluate DIF with new criterion

DIF Detection Methods

- Mantel-Haenszel: condition on raw score, statistical test of contingency tables
- Logistic Regression: condition on raw score, model group-response relationship
- IRT Methods: condition on ability (θ) Compare item parameters or ICCs

IRT DIF Detection

- Compare Item Parameter Estimates
  - Multivariate test (b, a, and, c)
  - t-tests on b-values

- Area Methods
  - Total Area (e.g., Raju, 1988, 1990)
  - Squared Differences
  - Weighted Areas and Differences

Parameter Comparisons

- Null Hypothesis H_0:
  - b1 = b2
  - a1 = a2
  - c1 = c2

- Alternative H_1:
  - b1 ≠ b2
  - a1 ≠ a2
  - c1 ≠ c2

Parameter Comparisons

- Chi Squared test of parameter differences, after parameters have been equated:

$$\chi^2 = (a_{diff} \;\; b_{diff} \;\; c_{diff})' \; \Sigma^{-1} \; (a_{diff} \;\; b_{diff} \;\; c_{diff})$$

- Df = p, the number of parameters;
  - Sometimes done with common c, which is left out (df = 2)

- Criticism:
  - Asymptotic properties known, but it's not clear how big a sample you need to do it

Parameter Comparisons

- t-test of difference between b-parameters, after parameters have been equated

$$t_{DIF} = \frac{b_{ref} - b_{foc}}{\sqrt{SE^2(b_{ref}) + SE^2(b_{foc})}}$$

- This is a very simple procedure which may be informative for identifying items which call for a closer look, but not too common to rely solely on this

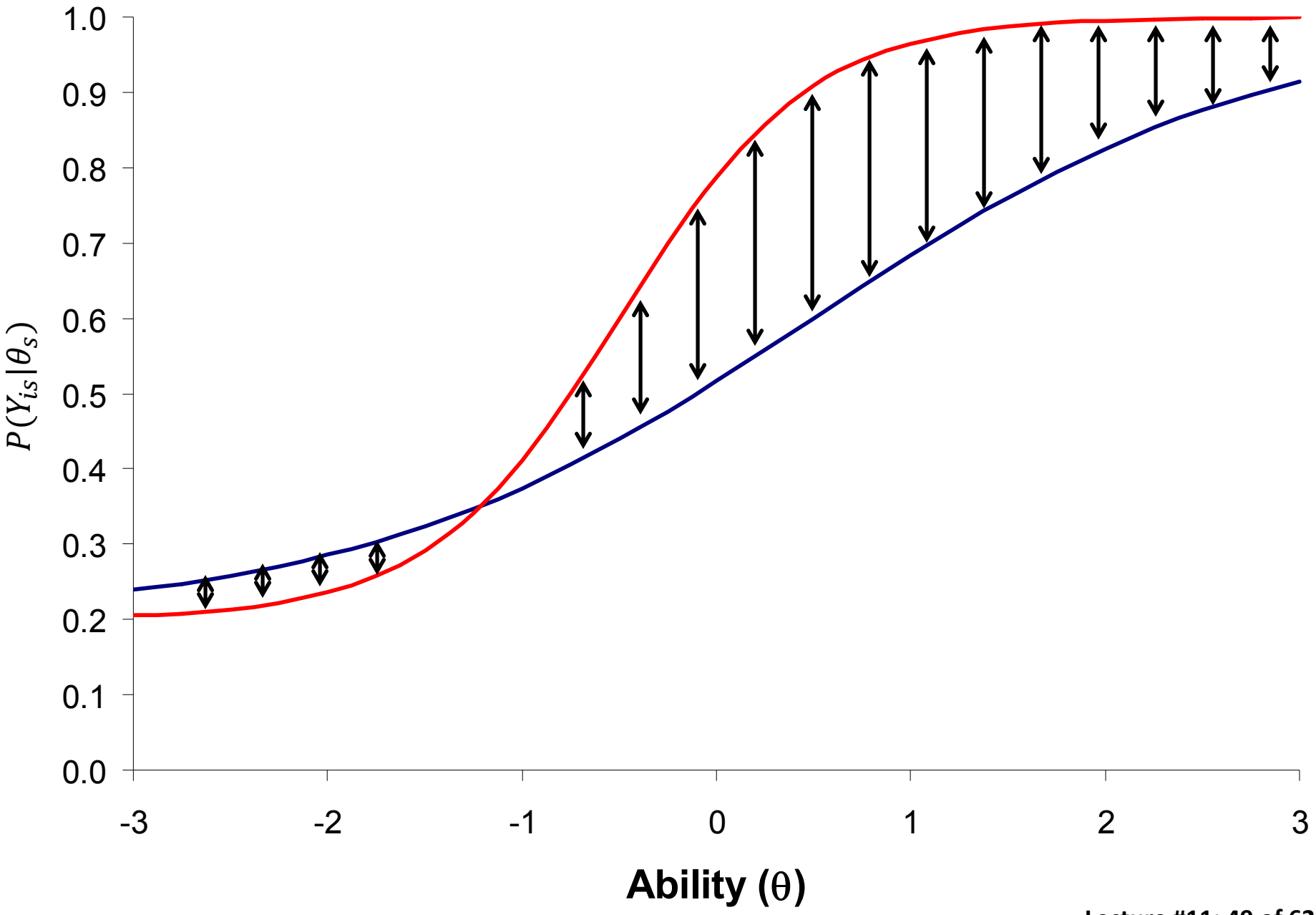
- Doesn't account for a- and c-parameters, which may vary even for fixed b-value

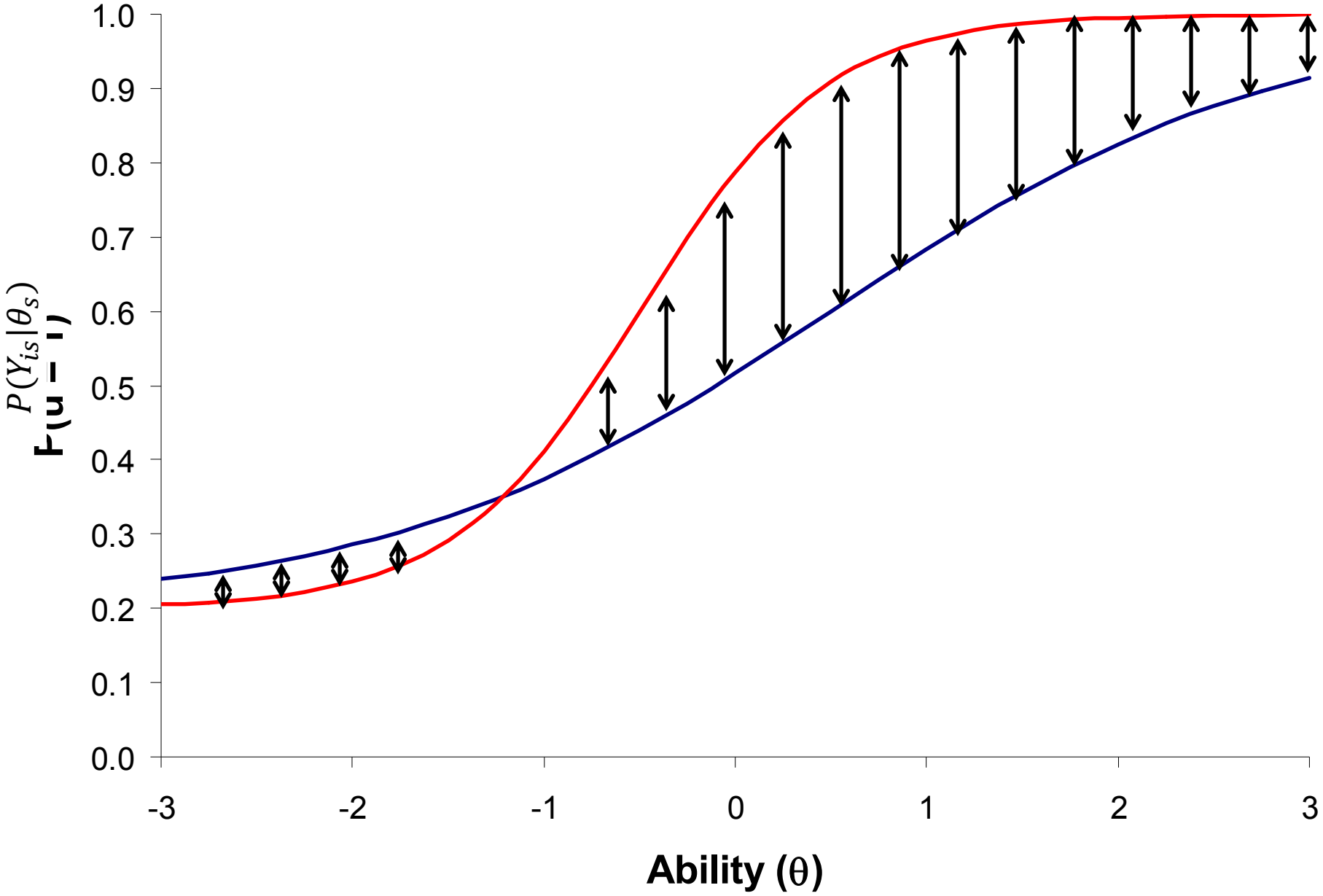
IRT Area Methods

- These methods are more common, as they compare an ICC from one group against an ICC from the other and look at how much area is between the two…
  - Doesn't depend on differences in parameters, but actual differences in conditional probability

- The basic approach to these methods is to evaluate the amount of "space" between the two ICCs…
  - The smaller, the better

- Recall that if the item parameters were invariant across groups, the two ICCs would be coincident


1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
$P(Y_{is}|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)

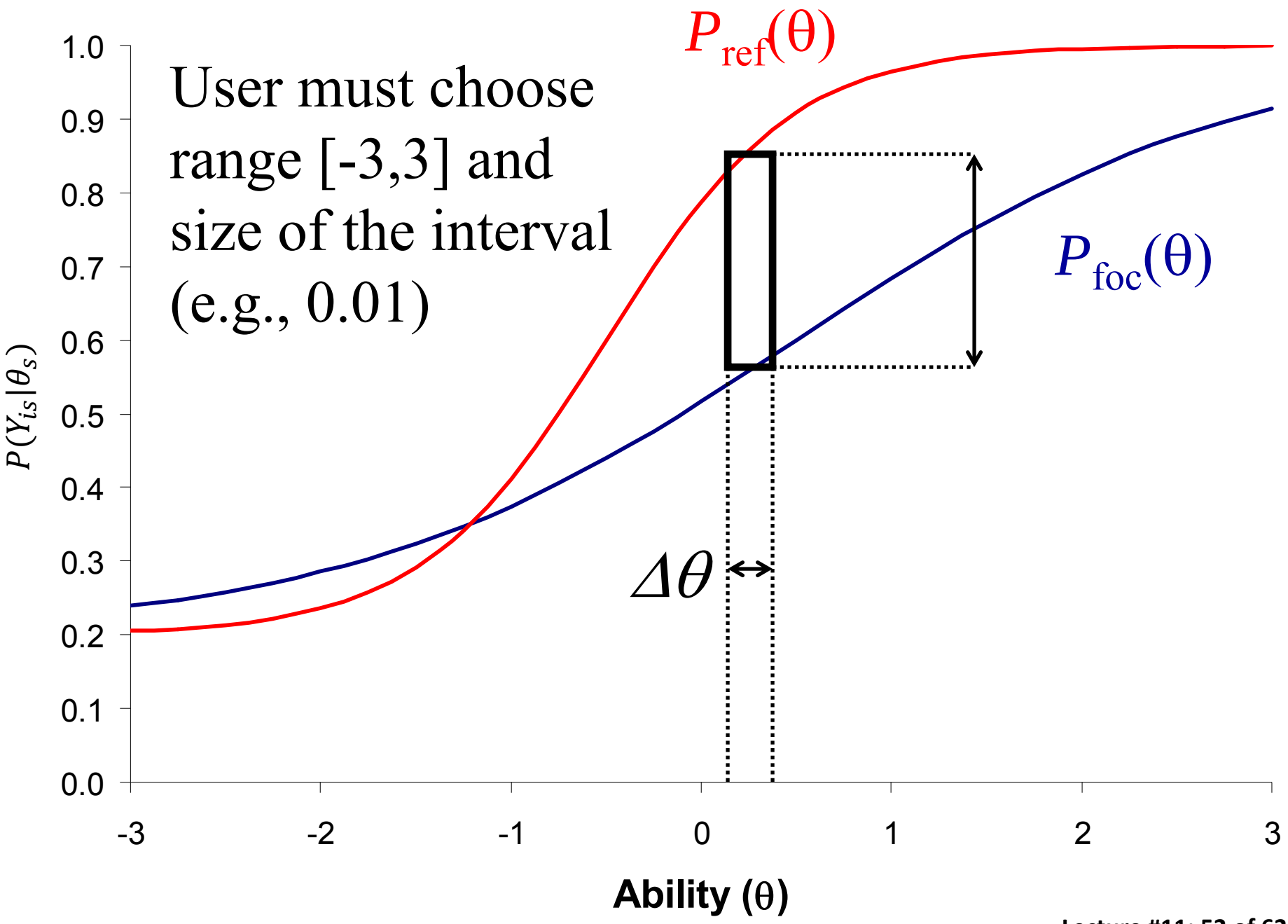

1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
$P(Y_{is}|\theta_s)$
-3
-2
-1
0
1
2
3
Ability (θ)

Calculating Area

- The area between the ICCs is defined as

$$Area = \sum_{k=1}^{m} \Delta\theta_k \mid P_{ref}(\theta) - P_{foc}(\theta) \mid$$

where $\Delta\theta_k$ is the width of a quadrature node


$P_{ref}(\theta)$
$P_{foc}(\theta)$
User must choose range [-3,3] and size of the interval (e.g., 0.01)
$\Delta\theta$
$P(Y_{is}|\theta_s)$
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Ability (θ)

DIF, DTF Interpretation

- DTF can be examined in the same way, only using a comparison of TCCs instead of ICCs

- How much is too much?
  - No significance test, so one common approach is to determine the amount of DIF present in simulated non-DIF data

Simulated Data for Comparison

- Combine groups and estimate item and ability parameters

- Use item and person parameter estimates to simulate item response data
  - Any DIF present will only be due to statistical artifact

Simulated Data for Comparison

- For each item, determine the amount of DIF present, which will only be a result of sampling error

- This amount of DIF becomes the cutoff point; any DIF greater and the item is flagged

Other Area Methods

- Absolute value of each difference is the most commonplace approach

- Less Common
  - Signed DIF – positives and negatives can cancel out (with non-uniform DIF)
  - Squared DIF – less intuitive to interpret
  - Weighted DIF – magnitude of differences are weighted by conditional sample sizes

Practical Concerns

- When all is said and done, after an item has been identified as DIF, the questions then becomes “why?”

- In the absence of interpretation, it is difficult to justify deleting an item

*Example DIF against Females

Decoy : Duck :: _____ : _____

(A) Net : Butterfly

(B) Web : Spider

(C) Lure : Fish

(D) Lasso : Rope

(E) Detour : Shortcut

*from Holland & Wainer, 1993

Practical Concerns

- Some “essential” items (either due to content or statistical properties) may be left on the test
  - Even with DIF

- Ameliorated by constructing forms balanced to include some other item or items which favor the focal group
  - Make TCCs match

CONCLUDING REMARKS

Concluding Remarks

- DIF is a heavily researched topic:
  - No shortage of articles comparing methods
    - Or creating new ones
- Methods presented today are general but very useful
- The nature and spirit of DIF is sometimes called "invariance testing"
  - More from a factor analytic perspective
  - Attempts the same goal with different sets of parameters (i.e., intercepts and slopes)

Up Next…

- Equating Example