Introduction to Differential Item Functioning

American Board of Internal Medicine
Item Response Theory Course
Overview

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

• Revisit some test score equating issues.
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 examinees 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.
Important Assumptions

• Unidimensionality.

• Parameter Invariance.

• Violations of these assumptions across identifiable groups basically provide us with a definition of DIF.
Unidimensionality

• This assumption basically states that the test of interest 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.
Unidimensionality

• If group membership impacts or explains performance, then the assumptions of the model are not being met.

• Basically, DIF=Dimensionality
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 examinee samples should still produce close to the same item parameter estimates.
Parameter Invariance

- Invariance of item parameters (a, b, c)
  - Scatterplots of b-b, a-a, c-c based on one sample of examinees 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 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
  – Whites – Blacks
  – 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.
The diagram shows the probability function $P(u = 1 | \theta)$ as a function of ability $\theta$. The graph illustrates the relationship between the probability of a binary outcome and the underlying ability level, with different curves for different conditions or parameters.
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).
$P(u = 1 \mid \theta)$
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.
The problem of sample 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, they 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 & Sigma Equating

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

\[ x = \frac{\sigma_{b-ref}}{\sigma_{b-foc}} \quad y = \mu_{b-ref} - x \mu_{b-foc} \]
Mean & Sigma Equating

• This transformation places the scale of item parameters from the Focal Group onto the scale of item parameters from the Reference Group:

\[ x = \frac{\sigma_{b-ref}}{\sigma_{b-foc}} \quad y = \mu_{b-ref} - x \mu_{b-foc} \]
Mean & Sigma with DIF

• All items comprise the “anchor test” as both groups took the same form.
• Treat the entire test as a block of linking items, determine the equating constants, x and y, and transform.
• Another transformation procedure, such as TCC or robust M & S, could alternatively be utilized.
Mean & Sigma Example

• 25 items
  – Reference: \( \mu_b = -0.06, \sigma_b = 1.03 \)
  – Focal: \( \mu_b = -0.25, \sigma_b = 1.12 \)
• \( x = 1.03 / 1.12 = 0.92 \)
  \( y = -0.06 - (0.92 \times -0.25) = 0.17 \)
• \( b_{\text{new}} = x \, b_{\text{old}} + y \)
• \( b_{\text{new}} = 0.92 \, b_{\text{old}} + 0.17 \)
M & S Transformation of Focal Group b-parameters

\[ b_{\text{new}} = x \, b_{\text{old}} + y \]

\[ b_{\text{new}} = 0.92 \, b_{\text{old}} + 0.17 \]

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

This indicates that the Focal Group had higher ability than the Reference Group.
if \[ \theta_{new} = x\theta + y \]

then

\[ b_{new} = xb + y \]

\[ a_{new} = \frac{a}{x} \]

\[ c_{new} = c \]

After item parameters are adjusted, the same transformation of b is done for all \( \theta \)…now the Focal examinees will look more able (as they should).

These transformations preserve the probability:

\[
P(u = 1 \mid \theta_{new}) = P(u = 1 \mid \theta)
\]
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.
This shows a pattern of non-uniform DIF for the minority group.
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, \( \theta \)).
• 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 (\(\theta\))
  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

• $H_0$: $b_1 = b_2$
  $a_1 = a_2$
  $c_1 = c_2$

• $H_1$: $b_1 \neq b_2$
  $a_1 \neq a_2$
  $c_1 \neq c_2$
Parameter Comparisons

- $\chi^2$ test of parameter differences, after parameters have been equated:

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

- This info can be obtained from BILOG-MG.
- 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.
• 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.
t-test for b-parameter

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

- More useful in a Rasch situation, because this is the same as the multivariate test!
t-test for b-parameter

• This may be done automatically for you in BILOG-MG, using a DIF command (see Help menu).
• DIF is only assessed in terms of item difficulty (b-parameter), no consideration to non-uniform DIF possibilities.
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.
IRT Area Methods

• 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.
$P(u = 1)$ vs Ability ($\theta$)
Calculating Area

• The area between the ICCs is defined as:

\[ \text{Area} = \sum_{k=1}^{m} \Delta \theta_k \left| P_{\text{ref}} (\theta) - P_{\text{foc}} (\theta) \right| \]

where \( \Delta \theta_k \) is the width of a quadrature node.
User must choose range [-3,3] and size of the interval (e.g., 0.01)
Calculating Area

• Closed form solution (requires that a common c-parameter be fixed across calibrations):

\[
\text{Area} = (1-c) \left| \frac{2(a_{\text{ref}} - a_{\text{foc}})}{D a_{\text{ref}} a_{\text{foc}}} \ln \left[ 1 + e^{D a_{\text{ref}} a_{\text{foc}} (b_{\text{ref}} - b_{\text{foc}})} \right] - (b_{\text{ref}} - b_{\text{foc}}) \right|
\]

• Otherwise, the area between ICCs would be infinite!
IRT Area Methods

• In practice, a very useful way to examine DIF is to actually construct the ICCs for both groups and visually inspect any differences.

• Let’s go through an example of an actual DIF study to see how this would work…
DIF Example

• To demonstrate the graphical methods of DIF, we will now work through an example.

• Data from a single test were collected:
  – Simulated data (for example purposes).
  – 50 multiple choice items total.
    • All scored correct/incorrect.
  – 10,000 examinees took the test.
  – Two groups:
    • 5,000 examinees per group.

• The data for each group can be found in the files group1.dat and group2.dat.
Analysis Steps

1. Run a 2PL model in BILOG on Group 1.
   • Save the item and examinee parameters.
2. Run a 2PL model in BILOG on Group 2.
   • Again, save the item and examinee parameters.
3. Load the estimated item parameter values into Excel.
   • Find the parameters for all items for each exam.
Analysis Steps

4. Using the estimated b parameters, compute:

\[ x = \frac{\sigma_{b-\text{Group1}}}{\sigma_{b-\text{Group2}}} \]

\[ y = \mu_{b-\text{Group1}} - x \mu_{b-\text{Group2}} \]

5. Transform the a and b parameters from Group 2 onto the scale of the Group 1 calibration.

6. Reconstruct the ICCs for each group and visually inspect the slides.
   • Use the macro in Excel.
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.
- This is ameliorated by constructing forms balanced to include some other item or items which favor the focal group...make TCCs match.
Conclusion

• Good reference:
  Holland & Wainer (Eds.) (1993) *Differential Item Functioning*

• DIF is a *heavily* researched topic, no shortage of articles comparing methods (or creating new ones).

• Methods presented today are general but very useful.
Next...

- Practice with software
- Operational analyses.