IUPUI Assessment Institute, 2022

The Importance of DIF: An important tool for identifying potential bias in Performance Assessment

Identifying Biases in Performance Assessments: The Differential Item Functioning Approach

Marylee Demeter, Ed.M., M.A., *Senior Assessment Developer*
Heather Hayes, Ph.D. *Psychometrician*
Goran Trajkovski, PhD. *Senior Lead College Compliance Advisor*
Western Governors University

Learning Outcomes

- Participants will be able to define differential item functioning (DIF).
- Participants will be able to explain the difference between adverse and benign DIF.
- Participants will be able to identify the potential for DIF in a performance assessment based on item content.
- Participants will be able to give suggestions on how to improve the wording or phrasing of an item with adverse DIF.

What is Differential Item Functioning (DIF)?

- When an item functions differently for members of one group (relative to another) after controlling for differences in ability.
  - Difficulty of item (likelihood of endorsing item)
  - Discrimination of item scores (precision in measurement)
- Focal vs. Reference Groups
  - Focal: minority, female
  - Reference: Caucasian, male

Why study DIF?

- DIF *can* negatively impact validity (construct-irrelevant variance)*
  - Theory as to why:
    - The factor driving item scores should be total score/ability
    - If there is a secondary factor (item score is multidimensional), and groups differ on the second dimension (group based on latent factor scores)
    - A by-product of multicollinearity
  - Adverse vs. benign DIF (Roussos & Stout, 1996)
    - Nuisance dimension vs. auxiliary dimension

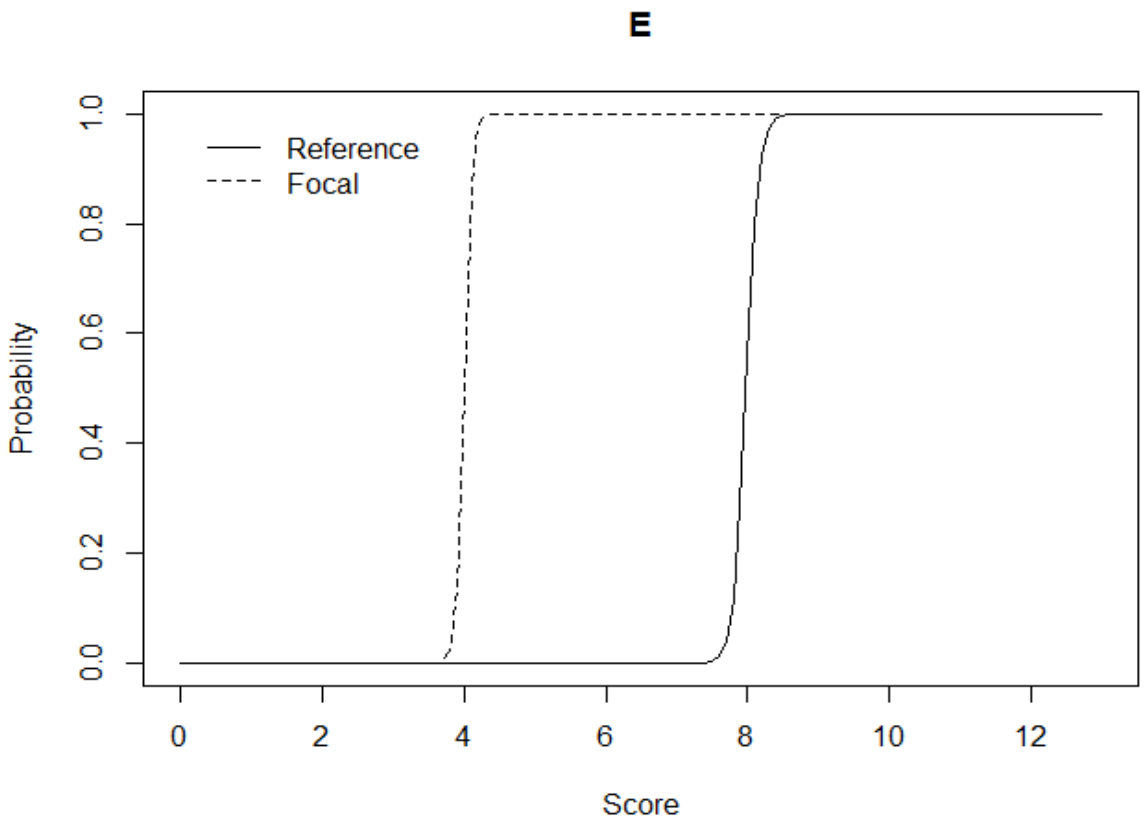
Forms of DIF

- Uniform DIF
  - An item is more difficult for one group than another when controlling for ability (the item is consistently harder for one group - for all ability levels)
- Non-Uniform DIF
  - Differences in item difficulty between groups varies along the ability spectrum (likelihood of item endorsement based on interaction between ability and group)
    - E.g., focal group favored at low ability levels,
    - E.g., reference group favored at high ability levels
- When an item has higher discrimination for one group than another when controlling for ability.

Statistical Methods

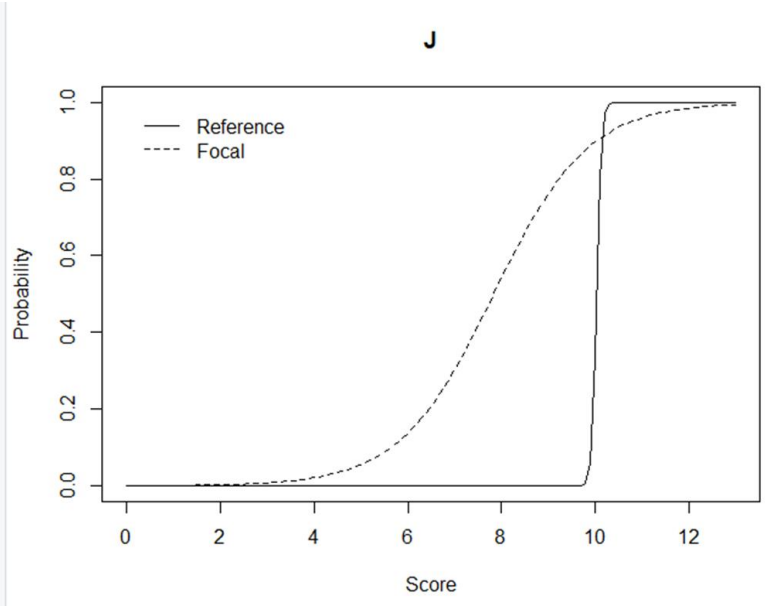
- (Generalized) Mantel-Haenszel:* - uniform only
  - Match respondents on ability level (total score)
  - Chi-square test for one item at a time, one ability level at a time: group (2+) x item score (0,1)
  - If >0 chi-square tests is statistically significant, then item has DIF
  - Direction of impact (focal vs. reference)= Delta MH stat
  - + chi-square = effect size (C+ or -)
- Logistic approach - both (uniform and non-uniform)
  - logistic regression p(item1=1) with 1) ability and 2) group as predictors; LRT for best model fit; best fit model(if 2), R^2 = effect size
- Lord's Chi-square (χ^2) – both (uniform and non-uniform)
  - IRT model parameters (fit IRT model to data first);
  - chi-square test – difference in matrix of parameter estimates for an item b/w focal and reference group)

Visual for Uniform DIF (Logistic)

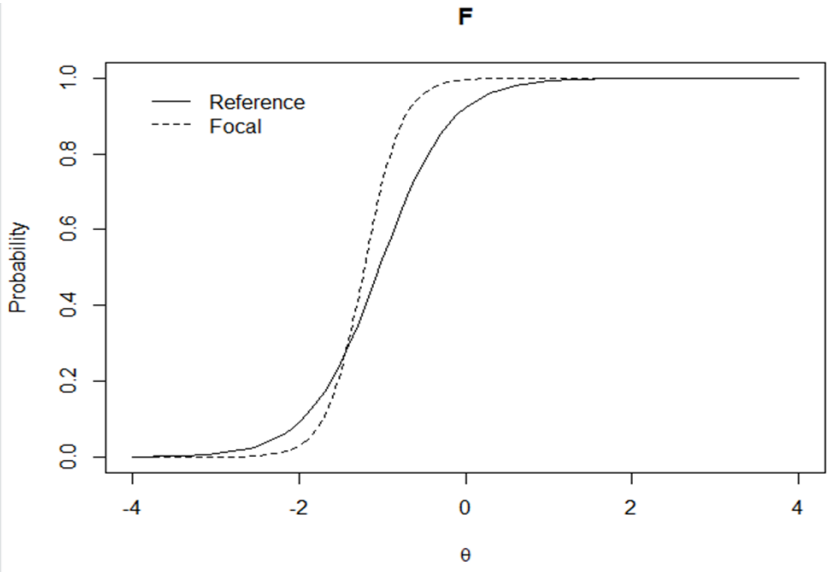


Visuals Non-uniform DIF

Logistic – regression curve



Lord's χ2 – item characteristic curve



Performance Assessments

- Test DIF for each aspect score (pass/fail) associated with a task prompt within the given task.
- All performance assessments involve data management, manipulation, and analyses applied to a given scenario.
- Sample of performance assessment scenarios (2018-present):
  - Census data per state by year (python programming) (n=247)
  - Census data per state by year (regression analysis) (n=350)
  - Increase/decrease in police funding (n=693)*
    - *Note: first course in program serious (remove drop-outs; only used students to attemped next course in series where the performance assessment scenario involves census data per state by year.*
  - Emergency bank loans, finance in a downsized economy (n=760)
- Groups considered:
  - Gender (female), Ethnicity (minority), Income (low income), and Rural/Urban (rural)

The Three Methods

- **Mantel-Haenszel (MH)** picked up more DIF than the other two methods.
- **Logistic** would not run if the proportions of group sample sizes are too large/offset.
  - E.g., easy items (only missed by 5 students)
- **Lord's χ^2** is most precise (if IRT model fit is good), so this method was ultimately chosen for determining presence of DIF.

PAs with no DIF (except in MH)

Scenarios:

1. Census data per state by year scenario (python programming)
2. Census data per state by year scenario (regression analysis)

Police Funding Scenario

- Student must run analyses of types of police activity, frequency, # officers on scene, duration (hours, days), etc. to determine whether additional funding should be awarded
- Pre – George Floyd (2017 – May 2020) n = 927
- Post – George Floyd (May 2020-present) n = 443

Police Funding Scenario: Results

Task Aspect	Gender DIF		Minority DIF		Income DIF		Rural DIF	
(11 total)	Pre (18%)	Post (0%)	Pre (27.3%)	Post (18.2%)	Pre (18%)	Post (9.1%)	Pre (0%)	Post (36.4%)
A) Clean data								
B) Explain prep				M (low θ) α -				
C) Data sheets								R α -
D) Summary stats		F						
E) Linear regression								R (hi θ) α -
F) Outlier impact								
G) Residual plot								
H) Explain funding			M (low θ)					R (hi θ) α -
I) Precautionary behavior	F (low θ) α +							
J) Citations/ref			M (uni)					
K) Org/structure	F (low θ) α +		M (uni)	M (uni)		Low (low θ); α +		R (hi θ) α -

Police Funding Scenario: Conclusions

- Aspect-level DIF increased from pre tp post George Floyd: 15% DIF → 24.2% DIF
- Detectable patterns
  - Direction is consistent but scattered about (aspect)
  - DIF consistently found for ethnicity (minority) and somewhat for gender (females)
  - DIF decreases in each group category except Community/Rural (which increases dramatically)
  - There is more community type (***rural***) DIF post-George Floyd
  - In crucial aspects not typically a problem (data sheets, linear regression, explain results) plus the more common organizational structure
  - (income DIF is minor - 9%)
- George Floyd shooting – this event did not seem to exacerbate item bias for any group other than rural students
- But, why? Scenario or ability/competency?
  - Minority, Female, "suddenly" Rural

Bank Finances/Emergency Loans Scenario

- Student must determine gaps in (bank's) data required to determine and award equitable emergency loans to small businesses and individuals
- 1/1/2018 – 12/31/2021
- N=350

Bank Funding/Emergency Loans Scenario : Results

Aspect	Gender (0%)	Minority (42.9%)	Income (42.9%)	Rural (0%)
A) Description status of data				
B) ID gaps		Min α +		
C) ID origin of gaps		Min α +		
D) Recommendation			L	
D1) Justification			L α +	
E) Citations/ref				
F) Org/structure		M at low θ α +	L	

Bank Finances/Emergency Loans Scenario: Conclusions

- 21.4% DIF overall
- Detectable patterns:
  - Ethnicity (minority) and Income (low) in almost half the aspects
  - Minorities are most disadvantaged at IDing gaps and locating the origin of the gaps, as well as overall organization/structure of submission
  - Low-income students are most disadvantaged at giving an acceptable recommendation and its explanation, as well as overall organization/structure of submission.
- But, why? Scenario or ability/competency?
  - Rural

Overall Conclusions

- How much of this DIF is adverse vs. benign?
  - None of what was presented today is benign, but there is another PA with DIF that *is* benign (scenario involving cars: 46.2% DIF)
- IT PAs – Construct: Competence in Data Analytics
  - DIF found only for PAs with a controversial scenario (e.g., what data you're analyzing shouldn't be affected by the topic *unless one group is more experienced with or interested in that topic in a way that gives them an advantage; or vice-versa*)

Deep Thoughts…..”What-ifs?”

- How do we deal with DIF?
  - Maybe, who cares? Just throw the item away just incase? Maybe…. Once we have a sea of items… (problem will be obsolete)
  - But maybe not if the item functions excellently in all other way; perhaps it’s mostly a validity issue seeking to be explored?
    - Dig deeper, research! We need to understand *why* adverse and benign DIF is happening (e.g., course details; student background; multicollinearity)
- Implications for our assessments
  - Sterile vs. diversity – with diversity, not only do we increase DIF but also added variance in preference/interest level in topic/scenario
    - Formative assessments, course material: No sterilization; diversify.
    - Summative:
      - Match student profile to topic category schema/profile (only in summative; in order to control for factor of personal interest, adverse DIF)
      - Incorporate category scheme/profile into AIG

References

- Maki, 2017
- Roussos & Stout, 1996

Get Involved!

- This presentation will be available on the Assessment Institute website where you can access links embedded in this poster presentation.

- Are you interested in reviewing excerpts from the performance assessments analyzed in this study?
- Do you have ideas about how to improve language in performance assessments to ensure equity and attainment?
- We want to hear from you!
  - Please use this embedded link to visit our online form and provide feedback on excerpts from our performance assessments!

Thank you!

- Marylee Demeter, Ed.M., M.A., Senior Assessment Developer; Western Governors University
- Goran Trajkovksi, Senior Lead College Compliance Advisor; Western Governors University

- Heather Hayes, Ph.D., Senior Lead Psychometrician; Western Governors University