Because it might not make a big DIF: Assessing differential test functioning

David B. Flora
R. Philip Chalmers
Alyssa Counsell

Department of Psychology, Quantitative Methods Area
Differential item functioning

- Differential item functioning (DIF) refers to any situation in which an item within a test measures an intended construct differently for one subgroup of a population than it does for another.

- In short, DIF is item bias.

- DIF is often assessed using item response theory (IRT).
Differential item functioning (DIF)

• DIF can be assessed for items fitted to any IRT model, including three-parameter logistic (3PL) and graded-response models (GRM).

• An item has DIF if two respondents from distinct groups who have equal levels of the latent trait ($\theta$) do not have the same probability of endorsing response $k$ for that item:

\[
P(u_i^1 = k \mid \theta) \neq P(u_i^2 = k \mid \theta)
\]

where the super-script indexes group membership.

• DIF is readily assessed using a multiple-group generalization of an item response model.

• E.g., for the two-parameter logistic (2PL) model:

\[
P_i(u_i^g = 1 \mid \theta) = \frac{1}{1 + \exp(-a_i^g (\theta - b_i^g))}
\]
\[ P_i(u_i^g = 1 \mid \theta) = \frac{1}{1 + \exp(-a_i^g (\theta - b_i^g))} \]

\[ a^1 = a^2 \]
\[ b^1 \neq b^2 \]

\[ a^1 \neq a^2 \]
\[ b^1 = b^2 \]

\[ a_i^1 \neq a_i^2 \]
\[ b_i^1 \neq b_i^2 \]
Our questions:

Of course DIF is a property of an individual item. To what extent does DIF lead to biased total scores, that is, **differential test functioning** (DTF)? Can we quantify DTF?
• If a subset of items has DIF effects which consistently favor one group over another, then the overall impact of DIF on total test scores may be substantial.

• But DIF impact on total scores may be small when

  -individual DIF effects are small, especially if the test is long and only a small number of items show DIF.

  -DIF effects in some items favor one group, but DIF in other items favors the other group.
Test scoring with IRT

- First, represent the observed item response as a function of $\theta$, given estimated item parameters $\psi_i$ (item score function):

$$S_i(\theta, \psi_i) = \sum_{k=1}^{K} (k - 1) P(u = k | \theta, \psi_i)$$
From these item-score functions, we can then represent the expected total score as a function of $\theta$, given the set of all item parameters, $\psi$:

$$T(\theta, \psi) = \sum_{i=1}^{I} S_i(\theta, \psi_i)$$

This total-score function can be generalized to the multiple-group context:

$$T(\theta, \psi^g) = \sum_{i=1}^{I} S_i^g(\theta, \psi_i^g)$$

Then we can visualize how individual DIF effects accumulate across items to create distinct total-score functions, or DTF, across groups:
Test characteristic curves by group

No DTF

DTF consistently favoring group 1

DTF favors group 1 at high $\theta$, DTF favors group 2 at low $\theta$.

DTF mostly favoring group 2
Quantifying DTF

\[ s_{DTF} = \int [T(\theta, \psi^1) - T(\theta, \psi^2)] g(\theta) d\theta \]

\[ u_{DTF} = \int |T(\theta, \psi^1) - T(\theta, \psi^2)| g(\theta) d\theta \]

\( g(\theta) \) is a weighting function that makes these measures averages rather than total areas.

• Plot depicts large \( u_{DTF} \) but small \( s_{DTF} \).
• If \( s_{DTF} = 1 \), then group 1 total scores are one point higher than group 2 scores on average (despite equal \( \bar{\theta} \)).
• Estimates of \( u_{DTF} \) and \( s_{DTF} \) are obtained with numerical quadrature.
• These estimates are subject to sampling error in \( \psi^g \)...
\[ \hat{\psi} \sim N(\psi, \Sigma(\hat{\psi})) \]

where \( \Sigma(\hat{\psi}) \) is the inverse of the observed information matrix.

Idea based on Thissen & Wainer (1990):

- Stochastically impute plausible values of the item parameters from the MVN distribution of \( \hat{\psi} \);

- calculate the \( sDTF \) and \( uDTF \) statistics in each imputed \( \hat{\psi} \);

- use collection of “imputed” \( \hat{\psi} \)s to form confidence envelopes for TCCs and confidence intervals for \( sDTF \) and \( uDTF \).
95% confidence envelopes for two TCCs
Comparison with a previous approach

Our $sDTF$ and $uDTF$ offer several advantages over $DTF$ developed by Raju et al. (1995):

• No need to rely on stand-in latent trait estimates, $\hat{\theta}$, which makes $DTF$ overly sensitive to sample characteristics and introduces additional uncertainty.

• Choice of reference vs. focal group remains arbitrary, but for $DTF$ it affects prediction of stand-in $\hat{\theta}$ estimates.

• $sDTF$ and $uDTF$ remain in the metric of expected test scores.
Simulation results

• When *no* DIF existed for any items, omnibus tests for *sDTF* and *uDTF* retained nominal or slightly conservative (for *sDTF*) Type I error rates.

• *sDTF* was more effective at detecting differential scoring when there was systematic DIF in intercepts, whereas *uDTF* was more effective when there was systematic DIF in slopes.

• When bidirectional DIF existed, these DIF effects cancelled each other out to produce small DTF statistics.

• Overall, the proposed DTF statistics demonstrated desirable properties that have not been obtained for previous DTF methods.
Example application

- Responses from the General Self-Efficacy Scale assessed for DIF and DTF across samples from Canada ($n = 277$) and Germany ($n = 219$).

- 10 items with 4-point ordinal response scale assessing degree to which one generally views one’s own actions as responsible for good outcomes → Graded response model

- DIF is present in 6 of the 10 items...how does that manifest at the total score level?
• Generating 1000 imputations for the DTF statistics resulted in significant $sDTF$ ($p = .002$), but the effect was small:

• $sDTF = -0.629$ (95% CI: -1.038 to -0.270)
• $uDTF = 0.663$ (95% CI: 0.297 to 1.079)

• Bias in the total scores was approximately 0.629 raw score points (or 1.57% of possible total score = 40) in favor of the German population.
At average to lower levels of $\theta$, Canadians tend to score lower on the GSE.
Shaded areas show 95% confidence envelopes.
• Ignoring measurement bias, the mean total score was slightly *higher* among Germans \((M = 30.54)\) than Canadians \((M = 29.94)\), but not significantly, *\(p = .19\).*

• But IRT analyses showed that Germans had a *lower* latent mean \((\bar{\theta} = -.11)\) than Canadians \((\bar{\theta} = 0)\).

• This discrepancy occurred because of DTF that favored the German group overall (i.e., observed German total scores were higher than they should be).

• But this DTF effect is quite small, despite that 6 of 10 items showed DIF.
Conclusion

• Differential *item* functioning does not necessarily produce badly biased *test* scores.

• $sDTF$ and $uDTF$ measure the extent to which DIF effects accrue across an entire test to produce biased test scores.

• Sampling variability of $sDTF$ and $uDTF$ can be effectively measured using a stochastic imputation procedure.

• Researchers can use our DTF measures to assess the extent to which measurement bias influences test scores *overall* as well as within specific ranges of $\theta$. 
References

