Part 3: DIF via Robust Scaling

Peter F. Halpin

## Overview of Workshop

* Part 1. Intro + factor analysis + MI
* Part 2. IRT + DIF
* Part 3. Robust scaling + DIF + DTF ${\color{green}\leftarrow}$

## Overview of Part 2

* IRT-based scaling and its relation to DIF
* Robust scaling
* Tests of DIF based on robust scaling
* Tests of DTF (impact) based on robust scaling
* Worked example

## Organization

* Website: [peterhalpin.github.io/RDIF-workshop/](https://peterhalpin.github.io/RDIF-workshop/)
* Slides: These slides in HTML format
* Notes: These slides DOCX format (translated, editable)
* Code: Just the code from these slides

# Scaling and its Relation to DIF

## Review: Shortcomings of DIF analysis

* “Anchor items”
  + To test if one item has DIF, we have to assume some other items do not have DIF
* Anchors are required to estimate impact, otherwise tests of DIF are confounded by impact
  + Estimating impact also called “scaling” – more on this today
* Logical circularity: If we could figure out which items were anchors, we could use that same approach on the rest of the items, too!!

## DIF and scaling

|  |
| --- |
|  |

* Goal of DIF: compare items over groups
* Requirement for DIF analysis: multigroup scaling / estimating impact

## DIF and scaling

|  |
| --- |
|  |

* Anchor item selection: Kopf et al. (2015)
* Recent review: Teresi et al. (2022)

## DIF and scaling

|  |
| --- |
|  |

* Item pairs: Bechger & Maris (2015); Yuan et al. (2021);
* Regularization: Belzak & Bauer (2020); Magis et al., (2015); Schauberger & Mair (2020)

## DIF and scaling

|  |
| --- |
|  |

* Scaling: He et al. (2015); He & Cui (2020); Stocking & Lord (1983)
* DIF: Halpin (2022); Wang et al. (2022)

## What is scaling?

* To allow scores from different test forms to be compared
  + Putting two tests “on the same scale”
* Mostly applicable in large scale educational testing
  + Multiple versions (forms) for test security
  + Different tests administered at different time points
  + TOEFL, SAT, ACT, GRE, …
* Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking. Springer

## Types of scaling

* Test scores: Observed scores vs IRT-based scores
* Models for IRT-based scores: Concurrent calibration (multi-group models) vs separate calibration (separate models)
* Research design: equivalent vs non-equivalent groups, different or overlapping test items

## Types of scaling

* Test scores: Observed scores vs **IRT-based scores**
* Models for IRT-based scores: Concurrent calibration (multi-group models) vs **separate calibration (separate models)**
* Research design: equivalent vs **non-equivalent groups**, different or **overlapping test items**

## Comparable items, non-equivalent groups (CINEG)

* Two (or more) non-equivalent groups of respondents
  + Fall vs spring SAT
* Partially overlapping items
  + Anchor items, appear on both test forms
  + Separate items, appear on only one test form
* CINEG scaling: assumes anchor items have the same item parameters in both groups

## CINEG is formally the same as MI / DIF

* “non-equivalent groups of respondents” = impact
* “assumes anchor items have the same item parameters in both groups” = invariance

. . .

* Robust scaling: considers that some anchors perform differently across groups = DIF

. . .

* Different application: In scaling, groups of respondents are defined by what test they took, not pre-existing social groups

## How it works: Scaling functions

* Assume:
  + in reference group
  + in comparison group
* In the scaling context, and are called *scaling parameters*
* If we know these parameters, we can put two test forms on the same scale
* Earlier, we talked about these same parameters in terms of *impact*

## How it works: Scaling functions

* Scaling functions are used to compute scaling parameters using item parameters
* e.g., “mean-mean” scaling for 2PL model
* There are many scaling functions, we focus on this kind of “moment-based” function
* See Appendix for more details

## Summary

* Scaling is about putting scores from different test forms on the same scale
* Our focus: IRT-based scaling, separate calibrations, CINEG design
* DIF and IRT-based scaling with CINEG design are formally similar
  + “Two sides of same problem” but different research applications
* Scaling functions provide a useful too for DIF analysis

# Robust Scaling

* Items with DIF translate into outliers in scaling

## DIF, scaling, and robust regression

|  |
| --- |
|  |

* CINEG scaling via linear regression in the presence of DIF. Points represent difficulty parameters from the 2PL model, estimated in two groups The red point is an item with DIF. The scaling parameters are written as and . DGP = data generating parameters; LAD = least absolute deviation; OLS = ordinary least squares.

## From regression to scaling

* “Out of the box” robust regression doesn’t work very well for this problem
* Regression model misses some peculiar aspects of the scaling problem
  + Without DIF, item parameters have an exact linear relationship
  + Heteroskedastic error in both variables ( depends on )
  + We have estimates of – but how to use them??

## Robust scaling: Overall approach

* Implicitly define scale parameter (impact) via M-estimating equation for a location parameter
* is a scaling function based on parameters of item
* is the variance of obtained via the delta method
* is a “redescending” loss function chosen to flag outliers (i.e., items with DIF)
  + Use Tukey’s bi-square for computations
* Details in Halpin (2022)

## How items with DIF are flagged

* Estimation via iteratively weighted least squares (IRLS)
* Intuitive idea: set for items with DIF
  + i.e., “flag” items with DIF while estimating

## Weight function (Tukey’s bisquare)

|  |
| --- |
|  |

* If item does not have DIF, we know
* Choose per item tuning parameter based on desired Type I error rate for testing DIF
* Result: if item is outside of 95% confidence interval for “no DIF”

## Robust scaling: In practice

* Step 1. Maximum likelihood estimation of a focal psychometric model
  + Estimate separately in both groups, or use configural model
* Step 2. Extract model parameter estimates and their standard errors
* Step 3. Robust scaling is implemented as a post-estimation step to
  + Provide an estimate of impact that is robust to DIF
  + Flag item parameters with DIF at the desired Type I Error rate
* Step 4. Follow-up chi-square (Wald) tests for item-level DIF
* …

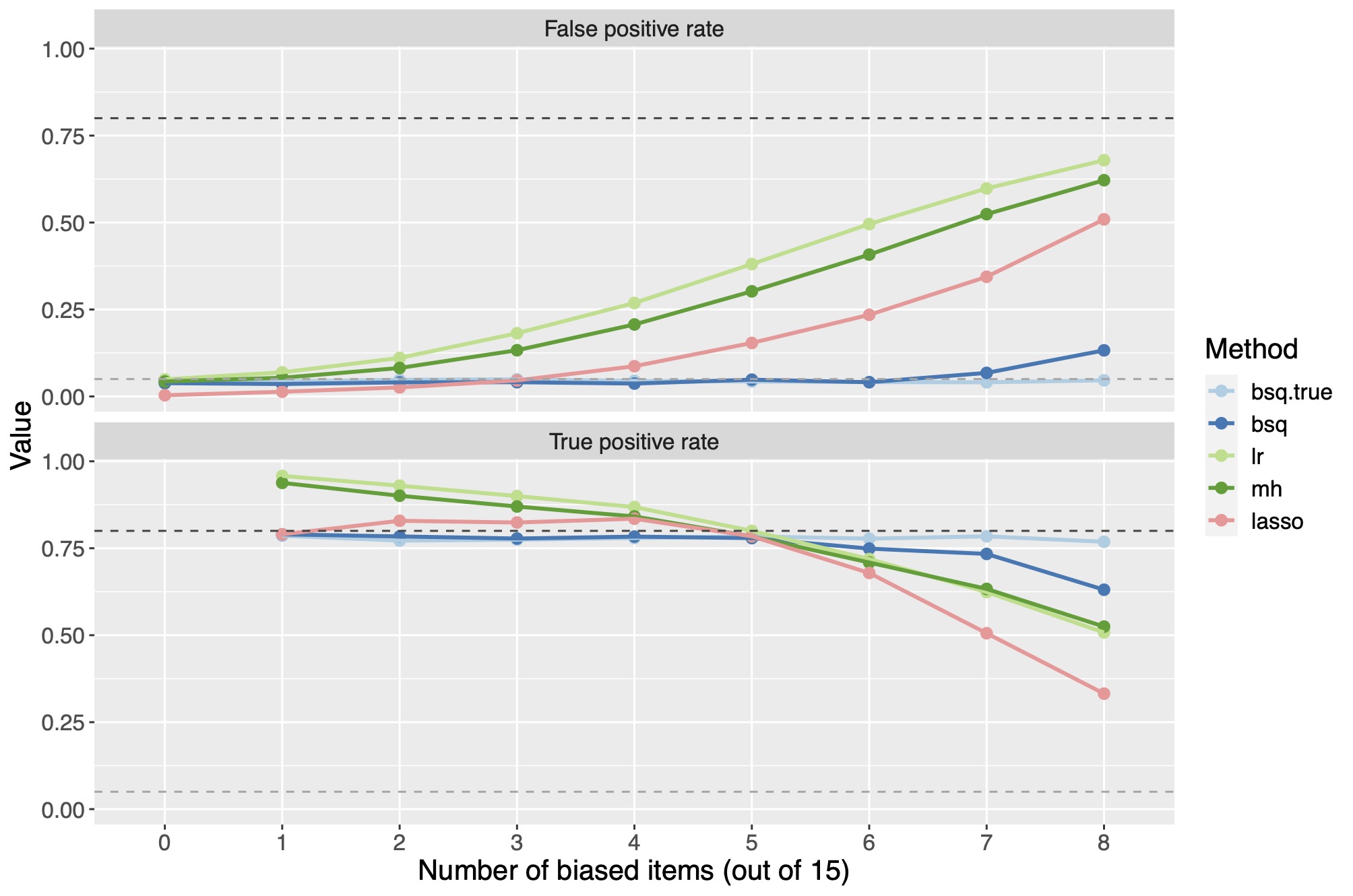
## Robust scaling: Additional details

* The approach does not require specification of anchor items
* Theoretical results guarantee that the procedure can tolerate up 50% of items with DIF
  + Traditional methods that use anchors fail at < 25%
* Data simulation results show that it performs well compared to other methods

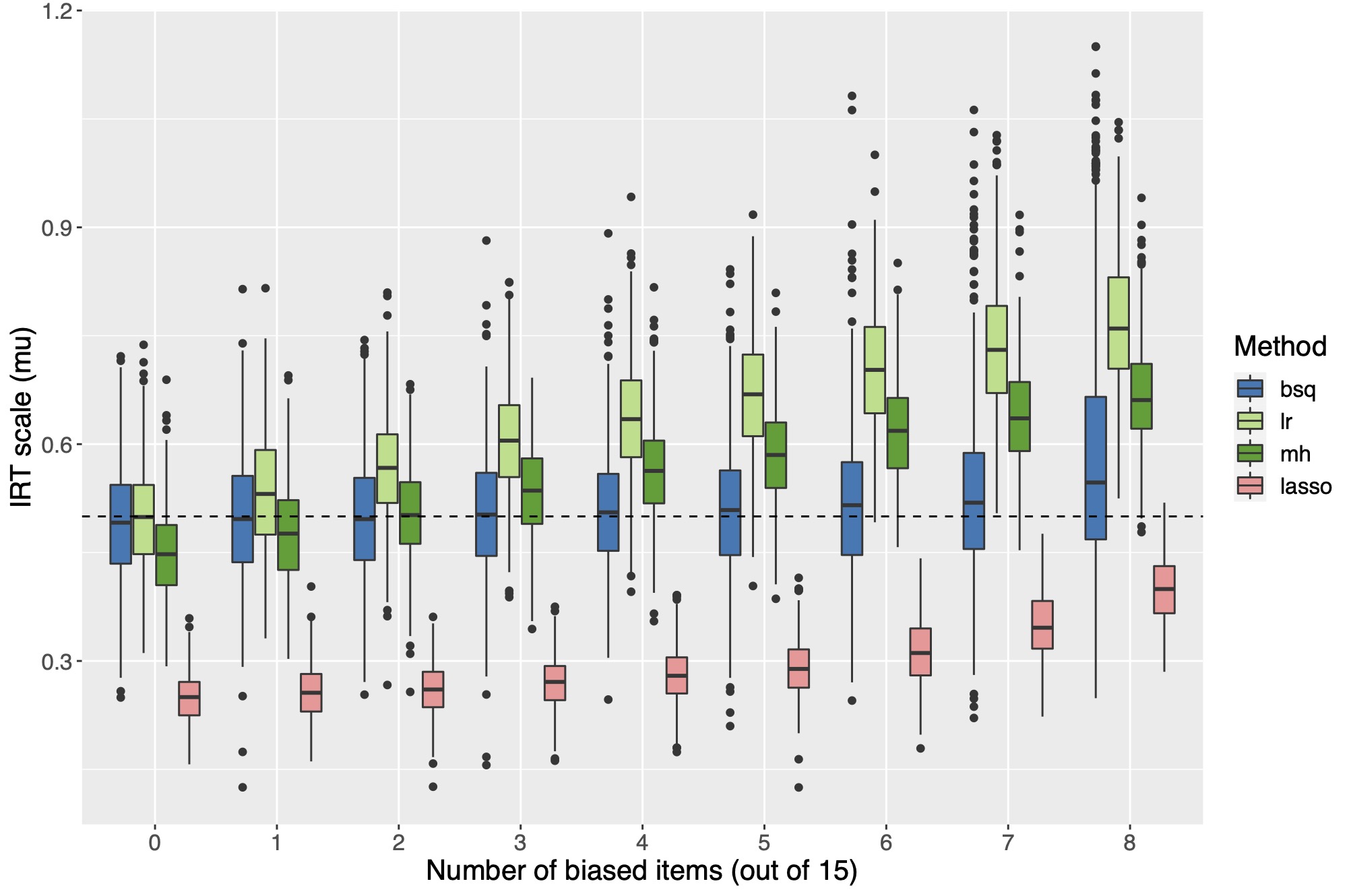
## A simulation study

* 2PL in two groups, DIF in item difficulty only
* DIF on item difficulties (intercepts) only ()
* Impact on mean only ()
* Focal factors
  + Number of items with DIF: 0 to , randomly selected
  + Method: LRT-DIF, Mantel-Haenszel, GPCM lasso, proposed M estimator
* Design factors
  + reps per number of biased items
  + person per group
  + items
  + ;
  + ;

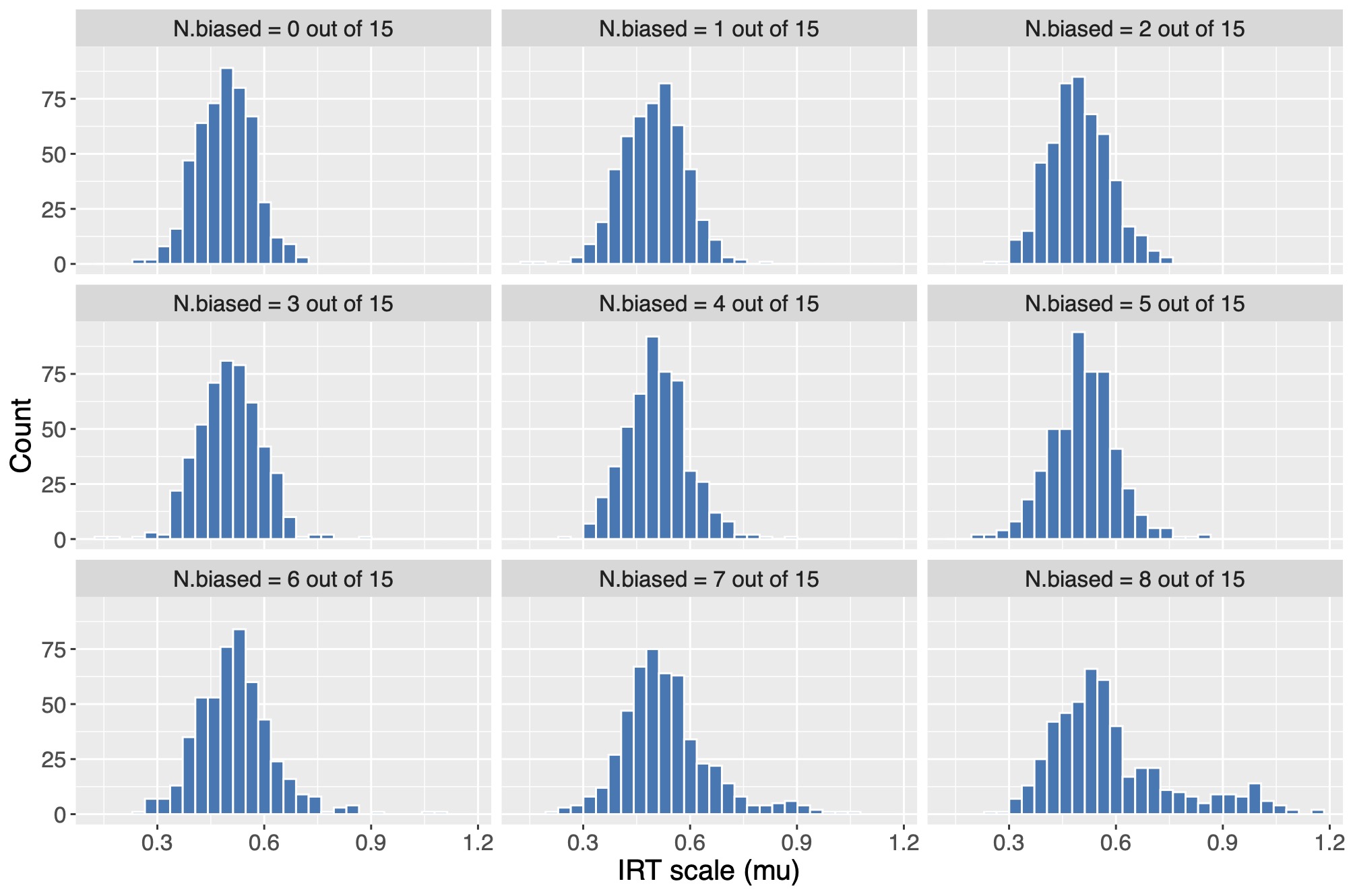
## Simulation Results



## Simulation Results



## Simulation Results



# Example

## Getting set up with R

* robustDIF package has been updated recently, so let’s re-install now

Loading required package: Matrix

# installer for github   
install.packages(remotes)  
  
# install robustDIF from github  
remotes::install\_github("peterhalpin/robustDIF")  
  
# load library   
library(robustDIF)

## Step 1. Estimate IRT model

* Can use configural model in mirt or list of two separate fits

Loading required package: stats4

Loading required package: lattice

library(mirt)  
  
# Set up data for mirt  
cint <- read.csv("cint\_data.csv")  
depression\_names <- c("cint1", "cint2", "cint4", "cint11",   
 "cint27", "cint28", "cint29", "cint30")  
depression\_items <- cint[, depression\_names]  
gender <- factor(cint$cfemale)  
  
# Estimate model (no invariance constraints)  
config.mod <- multipleGroup(depression\_items,   
 group = gender,   
 itemtype = "graded",   
 SE = T) # <- make sure to request SEs  
  
# Print parms (in slope-intercept format)  
coef(config.mod, IRTpars = F, simplify= T)

## Step 1. Estimate IRT model

coef(config.mod, IRTpars = F, simplify= T)

$`0`  
$items  
 a1 d1 d2 d3  
cint1 1.517 2.099 -0.442 -2.866  
cint2 1.323 1.100 -0.704 -2.362  
cint4 1.106 1.760 -0.025 -2.130  
cint11 0.931 1.723 0.249 -2.313  
cint27 1.647 0.799 -0.780 -2.607  
cint28 0.909 1.194 -0.312 -2.141  
cint29 1.074 -0.663 -1.908 -3.528  
cint30 1.101 0.766 -0.525 -2.278  
  
$means  
F1   
 0   
  
$cov  
 F1  
F1 1  
  
  
$`1`  
$items  
 a1 d1 d2 d3  
cint1 1.580 3.076 0.411 -2.226  
cint2 1.242 1.566 -0.142 -1.961  
cint4 1.122 2.480 0.431 -1.851  
cint11 1.420 2.371 0.712 -1.689  
cint27 1.748 1.190 -0.216 -2.205  
cint28 1.443 1.941 0.303 -1.860  
cint29 1.468 0.628 -0.627 -2.619  
cint30 1.271 0.702 -0.501 -2.192  
  
$means  
F1   
 0   
  
$cov  
 F1  
F1 1

## Step 2. Extract model parameters

* Extract model parameters from mirt

# Extract model parameters  
mirt.parms <- get\_model\_parms(config.mod)  
  
## Check output  
mirt.parms$est

$group.1  
 a1 d1 d2 d3  
item1 1.5167838 2.0990325 -0.44236088 -2.866174  
item2 1.3226290 1.0996419 -0.70402161 -2.361671  
item3 1.1060247 1.7599867 -0.02459558 -2.130449  
item4 0.9305090 1.7228653 0.24874502 -2.313067  
item5 1.6472558 0.7990658 -0.77984423 -2.606872  
item6 0.9094216 1.1941310 -0.31173439 -2.141181  
item7 1.0741429 -0.6629446 -1.90846725 -3.528486  
item8 1.1011885 0.7658870 -0.52489886 -2.277890  
  
$group.2  
 a1 d1 d2 d3  
item1 1.580371 3.0757984 0.4107194 -2.226117  
item2 1.242215 1.5663720 -0.1421765 -1.960686  
item3 1.122109 2.4799671 0.4308424 -1.850987  
item4 1.419502 2.3706138 0.7115116 -1.688975  
item5 1.748094 1.1899711 -0.2163709 -2.205330  
item6 1.442756 1.9410021 0.3027256 -1.859963  
item7 1.467982 0.6282358 -0.6270320 -2.619477  
item8 1.270718 0.7024522 -0.5005097 -2.192099

## Step 3. Run robust DIF analysis

* rdif is internal function for estimation

# "raw" output with weights  
rdif(mirt.parms, par = "intercept")

$est  
[1] 0.404222  
  
$weights  
 [1] 0.3839973 0.4828362 0.9999925 0.9835501 0.9494398 0.9304414 0.4264334  
 [8] 0.9999393 0.7666714 0.9440296 0.8000310 0.9806982 0.1649757 0.7915727  
[15] 0.6049233 0.6791094 0.9837524 0.4172171 0.0000000 0.0000000 0.6549483  
[22] 0.0000000 0.0000000 0.1326015  
  
$n.iter  
[1] 23  
  
$epsilon  
[1] 7.273842e-08

## Step 3. Run robust DIF analysis

* rdif\_z\_test and rdif\_chisq\_test for user-friendly output

# Test of individual item parameters (intercepts / thresholds)  
rdif\_z\_test(mirt.parms, par = "intercept")

z.test p.val  
cint1.d1 1.209 0.227  
cint1.d2 1.083 0.279  
cint1.d3 0.004 0.997  
cint2.d1 -0.178 0.859  
cint2.d2 0.314 0.754  
cint2.d3 -0.369 0.712  
cint4.d1 1.155 0.248  
cint4.d2 0.011 0.991  
cint4.d3 -0.691 0.489  
cint11.d1 0.330 0.741  
cint11.d2 -0.637 0.524  
cint11.d3 0.193 0.847  
cint27.d1 -1.510 0.131  
cint27.d2 -0.651 0.515  
cint27.d3 -0.924 0.356  
cint28.d1 0.822 0.411  
cint28.d2 0.177 0.859  
cint28.d3 -1.166 0.244  
cint29.d1 3.750 0.000  
cint29.d2 2.987 0.003  
cint29.d3 0.856 0.392  
cint30.d1 -3.189 0.001  
cint30.d2 -2.587 0.010  
cint30.d3 -1.563 0.118

## Step 3. Run robust DIF analysis

* rdif\_z\_test and rdif\_chisq\_test for user-friendly output

# Test of individual item parameters (slopes)  
rdif\_z\_test(mirt.parms, par = "slope")

z.test p.val  
cint1.a1 -0.023 0.982  
cint2.a1 -0.684 0.494  
cint4.a1 -0.180 0.857  
cint11.a1 2.238 0.025  
cint27.a1 0.102 0.919  
cint28.a1 2.471 0.013  
cint29.a1 1.558 0.119  
cint30.a1 0.603 0.547

## Step 3. Run robust DIF analysis

* rdif\_z\_test and rdif\_chisq\_test for user-friendly output

# Item-level tests   
rdif\_chisq\_test(mirt.parms)

chi.square df p.val  
cint1 2.269 4 0.686  
cint2 2.347 4 0.672  
cint4 2.798 4 0.592  
cint11 12.688 4 0.013  
cint27 3.360 4 0.499  
cint28 11.099 4 0.025  
cint29 30.338 4 0.000  
cint30 11.697 4 0.020

## Comparison: robust DIF and LR Test

* Note that in LR test of DIF, only cint29 and cint30 were found to have
* Will look at DIF on item slopes together during the workshop

## Comparison: robust DIF and LR Test

# DIF analysis for item slopes with mirt  
strong.invariance <- c("free\_mean", "free\_var", "slopes", "intercepts")  
strong.mod <- multipleGroup(depression\_items,  
 group = gender,  
 itemtype = "graded",  
 invariance = strong.invariance,  
 verbose = F)  
  
DIF(strong.mod,  
 which.par = c("a1"),  
 scheme = "drop")

groups converged AIC SABIC HQ BIC X2 df p  
cint1 0,1 TRUE 0.013 1.571 1.828 4.747 1.987 1 0.159  
cint2 0,1 TRUE -0.454 1.104 1.360 4.279 2.454 1 0.117  
cint4 0,1 TRUE -0.826 0.731 0.988 3.907 2.826 1 0.093  
cint11 0,1 TRUE -2.737 -1.179 -0.923 1.997 4.737 1 0.03  
cint27 0,1 TRUE 0.318 1.876 2.132 5.051 1.682 1 0.195  
cint28 0,1 TRUE -0.339 1.218 1.475 4.394 2.339 1 0.126  
cint29 0,1 TRUE -9.060 -7.503 -7.246 -4.327 11.06 1 0.001  
cint30 0,1 TRUE 1.246 2.803 3.060 5.979 0.754 1 0.385

## Summary of example

* Similar conclusions as LR test, but not exactly the same
  + Test of item thresholds found same items as LR test
  + Chi-square test found 2 additional items with DIF (due to on slopes)
* Next steps
  + Can follow up with partial invariance model as before to examine item-level effects
  + May consider revising or omit items …
* Other concerns
  + Different DIF methods lead to different conclusions (in general)
  + Does any of this affect conclusions about impact??

# DTF

* Differential test functioning: does DIF affect conclusions about impact?

## Recapping where we are

* DIF analysis is about items
  + Useful for test development
* DIF analysis does not provide a direct way of inferring whether DIF affects conclusions about impact
  + Often what we care about in research!

## Using robust DIF to address impact

* So far we have focused on using a robust estimate of scaling parameters (impact) as a way to test for DIF individual items
* We can also compare the robust estimate to a “naive” estimate that would arise if we ignored DIF
  + e.g., maximum likelihood estimate (MLE) of impact, based on the same scaling procedure
* If the two estimates give the “same” result, then DIF does not affect conclusions about impact
  + i.e., if we ignored DIF, we would arrive at the same conclusion about how groups differ

## Logic of test

* The logic of this test is the similar to the Hausman specification test
* Under the null hypothesis, both the robust estimate and the MLE are consistent (unbiased) estimates of the “true” impact
  + The MLE is more efficient, but this is not very important for us
* Under the alternative hypothesis, the both may be inconsistent, but the robust estimate will be less biased
  + Assuming < 50% of items exhibit DIF
* Consequently, the difference between the estimates can be used to test for the affect of DIF on impact

## Simulation studies

|  |
| --- |
|  |

## Simulation studies

|  |
| --- |
|  |

## Relation to MI

* In MI, we test whether (a subset of) item parameters are equal over groups
* May reject MI with little affect on impact
  + DIF on a single item may be negligible when averaged over many items
  + DIF in opposite directions can cancel out over items
* In testing DTF
  + We don’t test any item parameters (although may down weight!)
  + We just test whether two estimates of impact are equal
  + Note that items there may be still items with DIF even if impact is not affected!

## Implications for practice

* If all we want is to compare groups test scores:
  + Using robust approach, we can test for DTF without *before* having to do item-by-item DIF analysis
  + If there is no DTF, can proceed with group comparisons without doing DIF analysis
  + If there is DTF, can follow up with item-by-item analyses, test revisions, etc, before making comparisons

# Back to the example

delta\_test(mirt.parms)

rdif.est ml.est delta se.delta z.test p.val   
0.4042220 0.3955726 0.0086494 0.0396633 0.2180706 0.8273741

* Conclusion: The naive and robust gender mean-differences on depression do not differ

# Summary

* In test development, we almost always want to know about DIF at the item level
* In research settings, sometimes we just care about whether comparisons between groups are biased or not
* Using robust scaling, we can make inferences about DTF *before* doing an item-by-item DIF analysis
  + Trick: compare two estimates of impact, naive and robust
* Unlike tests of MI, we are not testing whether all items are DIF-free
  + We are just testing whether DIF affects conclusions about impact
* If we conclude there is no DTF, there may or not be DIF
  + If we really want to know about the individual items, need to do the DIF analysis!

# Wrapping up

## What we have covered today

* IRT-based scaling and its relation to DIF
  + More info on scaling in appendix
* Robust scaling
  + See Halpin 2022 for technical details
* Tests of DIF based on robust scaling
  + rdif\_z\_test and rdif\_chisq\_test
* Tests of DTF (impact) based on robust scaling
  + delta\_test
* Worked example

## Caveats and future directions

* robustDIF is in early stages of development
  + Just added support for categorical data last week!
  + Working on multiple groups this winter
  + Sure to be many bugs and issues!
* Please feel free to contact me with questions about the software or ideas for new developments!
  + peter.halpin@unc.edu

## References

Bechger, T. M., & Maris, G. (2015). A Statistical Test for Differential Item Pair Functioning. Psychometrika, 80, 317–340. https://doi.org/10.1007/s11336- 014- 9408- y

Belzak, W. C. M., & Bauer, D. J. (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. Psychological Methods, 25(6), 673–690. https://doi.org/10.1037/met0000253 He, Y., &

He, Y., & Cui, Z. (2020). Evaluating Robust Scale Transformation Methods With Multiple Outlying Common Items Under IRT True Score Equating. Applied Psychological Measurement, 44(4), 296–310. https://doi.org/10.1177/0146621619886050

He, Y., Cui, Z., & Osterlind, S. J. (2015). New Robust Scale Transformation Methods in the Presence of Outlying Common Items. Applied Psychological Measurement, 39(8), 613–626.https://doi.org/10.1177/0146621615587003

Huber, P. J., & Ronchetti, E. (2009). Robust statistics (2nd ed). Wiley.

Magis, D., Tuerlinckx, F., & De Boeck, P. (2015). Detection of Differential Item Functioning Using the Lasso Approach. Journal of Educational and Behavioral Statistics, 40(2), 111–135. https://doi.org/10.3102/1076998614559747

## References

Schauberger, G., & Mair, P. (2020). A regularization approach for the detection of differential item functioning in generalized partial credit models. Behavior Research Methods, 52(1), 279–294. https://doi.org/10.3758/s13428-019-01224-2

Stocking, M. L., & Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. Applied Psychological Measurement, 7(2), 201–210. https://doi.org/10.1177/014662168300700208

Teresi, J. A., Wang, C., Kleinman, M., Jones, R. N., & Weiss, D. J. (2021). Differential Item Functioning Analyses of the Patient-Reported Outcomes Measurement Information System (PROMIS) Measures: Methods, Challenges, Advances, and Future Directions. Psychometrika, 86(3), 674–711. https://doi.org/10.1007/s11336-021-09775-0

Wang, W., Liu, Y., & Liu, H. (2022). Testing Differential Item Functioning Without Predefined Anchor Items Using Robust Regression. Journal of Educational and Behavioral Statistics, 47(6), 666–692. https://doi.org/10.3102/10769986221109208

Yuan, K.-H., Liu, H., & Han, Y. (2021). Differential Item Functioning Analysis Without A Priori Information on Anchor Items: QQ Plots and Graphical Test. Psychometrika, 86(2), 345–377. https://doi.org/10.1007/s11336-021-09746-5

# Appendix

## More about scaling

* Specify 2PL IRT model in “slope-intercept” form
* Group 1
* Group 2
* Scaling involves solving for and
* We know this relationship holds for some choice of and because IRT models are identified only up to a linear transformation of .

## More about scaling

* Substituting in the scaling equations and doing the algebra gives:
* Taking the mean over items, gives the usual “naive” scaling (in slope-intercept form)
* In the CINEG design, we let the item parameters in the reference group stand-in for the “unscaled” item parameters

## Example with lavaan

library(lavaan)  
  
# Model (same as above)  
mod1 <- ' depression =~ cint1 + cint2 + cint4 + cint11 +   
 cint27 + cint28 + cint29 + cint30'  
# Fit model  
fit.config <- cfa(mod1,   
 data = cint,   
 std.lv = T,   
 ordered = T,   
 group = "cfemale") # <--- new   
   
# extract parms  
lavaan.parms <- get\_model\_parms(fit.config)  
  
# RDIF procedures (groups are reversed)  
delta\_test(lavaan.parms)  
rdif\_z\_test(lavaan.parms, par = "intercept")  
rdif\_z\_test(lavaan.parms, par = "slope")  
rdif\_chisq\_test(lavaan.parms)