Parte 3: FDI mediante Escalado Robusto

Peter F. Halpin

## Descripción General del Taller

* Parte 1. Introducción + análisis factorial + IM
* Parte 2. TRI + FDI
* Parte 3. Escalamiento robusto + FDI + FDP ${\color{green}\leftarrow}$

## Descripción General de la Parte 2

* Escalamiento basado en TRI y su relación con FDI
* Escalamiento robusto
* Pruebas de FDI basadas en escalamiento robusto
* Pruebas de FDP (impacto) basadas en escalamiento robusto
* Ejemplo práctico

## Organización

* Sitio web: [peterhalpin.github.io/RDIF-workshop/](https://peterhalpin.github.io/RDIF-workshop/)
* Diapositivas: Estas diapositivas en formato HTML
* Notas: Estas diapositivas en formato DOCX (traducidas, editables)
* Código: Solo el código de estas diapositivas

# Escalamiento basado en TRI y su relación con FDI

## Revisión: Limitaciones del análisis de FDI

* "Ítems ancla"
  + Para probar si un ítem tiene FDI, tenemos que asumir que algunos otros ítems no tienen FDI
* Las anclas son necesarias para estimar impacto, de otra forma las pruebas de FDI se pueden confundir con impacto
  + Estimar impacto también se llama "escalamiento" – más sobre esto hoy
* Circularidad lógica: ¡Si pudiéramos descubrir cuáles ítems son anclas, podríamos usar ese mismo enfoque en el resto de los ítems, también!!

## FDI y escalamiento

|  |
| --- |

* Objetivo del FDI: comparar ítems entre grupos
* Requisito para el análisis de FDI: escalamiento multigrupo / estimación de impacto

## FDI y escalado

|  |
| --- |

* Selección de ítem ancla: Kopf et al. (2015)
* Revisión reciente: Teresi et al. (2022)

## FDI y escalamiento

|  |
| --- |

* Pares de ítems: Bechger & Maris (2015); Yuan et al. (2021);
* Regularización: Belzak & Bauer (2020); Magis et al., (2015); Schauberger & Mair (2020)

## FDI y escalamiento

|  |
| --- |

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

## ¿Qué es el escalamiento?

* Permite comparar las puntuaciones de diferentes formas de prueba
  + Poner dos pruebas "en la misma escala"
* Principalmente aplicable en pruebas educativas a gran escala
  + Varias versiones (formas) para la seguridad de la prueba
  + Diferentes pruebas administradas en diferentes momentos del tiempo
  + TOEFL, SAT, ACT, GRE, …
* Kolen, M. J., & Brennan, R. L. (2014). Test Equating, Scaling, and Linking. Springer

## Tipos de escalamiento

* Puntuaciones de prueba: Puntuaciones observadas vs puntuaciones basadas en TRI
* Modelos para puntuaciones basadas en TRI: Calibración concurrente (modelos multigrupo) vs calibración separada (modelos separados)
* Diseño de investigación: grupos equivalentes vs grupos no equivalentes, ítems de prueba diferentes o superpuestos

## Tipos de escalamiento

* Puntuaciones de prueba: Puntuaciones observadas vs **puntuaciones basadas en TRI**
* Modelos para puntuaciones basadas en TRI: Calibración concurrente (modelos multigrupo) vs **calibración separada (modelos separados)**
* Diseño de investigación: grupos equivalentes vs “grupos no equivalentes”, ítems de prueba **diferentes o** que se traslapan

## Ítems comparables, grupos no equivalentes (CINEG por su nombre en inglés)

* Dos (o más) grupos de encuestados no equivalentes
  + SAT de otoño vs primavera
* Ítems parcialmente traslapados
  + Ítems ancla, aparecen en ambas formas de la prueba
  + Ítems separados, aparecen solo en una forma de la prueba
* Escalado CINEG: asume que los ítems ancla tienen los mismos parámetros de ítem en ambos grupos

## CINEG es formalmente lo mismo que IM / FDI

* "grupos de encuestados no equivalentes" = impacto
* "asume que los ítems ancla tienen los mismos parámetros de ítem en ambos grupos" = invariancia

. . .

* Escalamiento robusto: considera que algunas anclas se desempeñan de manera diferente en grupos = FDI

. . .

* Diferente aplicación: En el escalamiento, los grupos de encuestados se definen por qué prueba tomaron, no por grupos sociales preexistentes

## Cómo funciona: Funciones de escalamiento

* Suponemos:
  + en el grupo de referencia
  + en el grupo de comparación
* En el contexto del escalamiento, y se llaman *parámetros de escalamiento*
* Si conocemos estos parámetros, podemos poner dos formas de prueba en la misma escala
* Anteriormente, hablamos sobre estos mismos parámetros en términos de *impacto.*

## Cómo funciona: Funciones de escalamiento

* funciones de escalamiento se utilizan para calcular los parámetros de escalamiento usando los parámetros de los ítems
* por ejemplo, escalado "media-media" para el modelo 2PL
* Hay muchas funciones de escalamiento, nos centramos en este tipo de función "basada en momentos"

## Resumen

* El escalamiento implica poner puntuaciones de diferentes formas de prueba en la misma escala
* Nuestro enfoque: escalado basado en TRI, calibraciones separadas, diseño CINEG
* El FDI y el escalado basado en TRI con diseño CINEG son formalmente similares
  + "Dos caras del mismo problema" pero con diferentes aplicaciones de investigación
* Las funciones de escalamiento proporcionan una herramienta útil para el análisis de FDI

# Escalamiento robusto

* Ítems con FDI se traducen en valores atípicos en el escalamiento

## FDI, escalado y regresión robusta

|  |
| --- |

* Escalamiento CINEG mediante regresión lineal en presencia de FDI. Los puntos representan parámetros de dificultad del modelo 2PL, estimados en dos grupos. El punto rojo es un ítem con FDI. Los parámetros de escalamiento se escriben como y . PGD = parámetros generadores de datos; DAM = desviación absoluta mínima; MCO = mínimos cuadrados ordinarios.

## De la regresión al escalamiento

- La regresión robusta "lista para usar" no funciona muy bien para este problema

* El modelo de regresión no capta algunos aspectos peculiares del problema de escalamiento
  + Sin FDI, los parámetros de los ítems tienen una relación lineal exacta
  + Error heterocedástico en ambas variables ( depende de )
  + Tenemos estimaciones de – pero, ¿cómo usarlas?

## Escalamiento robusto: Enfoque general

* Definir implícitamente el parámetro de escala (impacto) μ a través de una ecuación de
* es una función de escalamiento basada en parámetros del ítem
* es la varianza de obtenida mediante el método delta
* es una función de pérdida "redescendente" elegida para señalar valores atípicos (es decir, ítems con FDI)
* Bi-cuadrada de Tukey

. . .

* Detalles en Halpin (2022)

## Cómo se señalan los ítems con FDI

* Estimación mediante mínimos cuadrados ponderados iterativamente (IRLS por su nombre en inglés)
* Idea intuitiva: establecer para ítems con FDI
  + es decir, señalar ítems con FDI mientras se estima

## Función de ponderación (bi-cuadrada de Tukey)

|  |
| --- |

* Si el ítem no tiene FDI, sabemos que
* Escoger por ítem el parámetro de ajuste basado en la tasa de error de Tipo I deseada para probar FDI
* Resultado: si el ítem está fuera del intervalo de confianza del 95% para "sin DIF"

## Escalamiento robusto: En la práctica

* Paso 1. Estimación de máxima verosimilitud de un modelo psicométrico focal por separado en ambos grupos
* Paso 2. Extracción de las estimaciones de los parámetros del modelo (MLEs) y sus errores estándar
* Paso 3. El escalamiento robusto es un paso posterior a la estimación que utiliza el enfoque descrito anteriormente para:
  + Proporcionar una estimación de impacto que es robusta al FDI
  + Señalar ítems con FDI a la tasa de error de Tipo I deseada
* Paso 4. Pruebas de chi-cuadrado (Wald) de seguimiento para FDI a nivel de ítem

## Escalad]miento robusto: Detalles adicionales

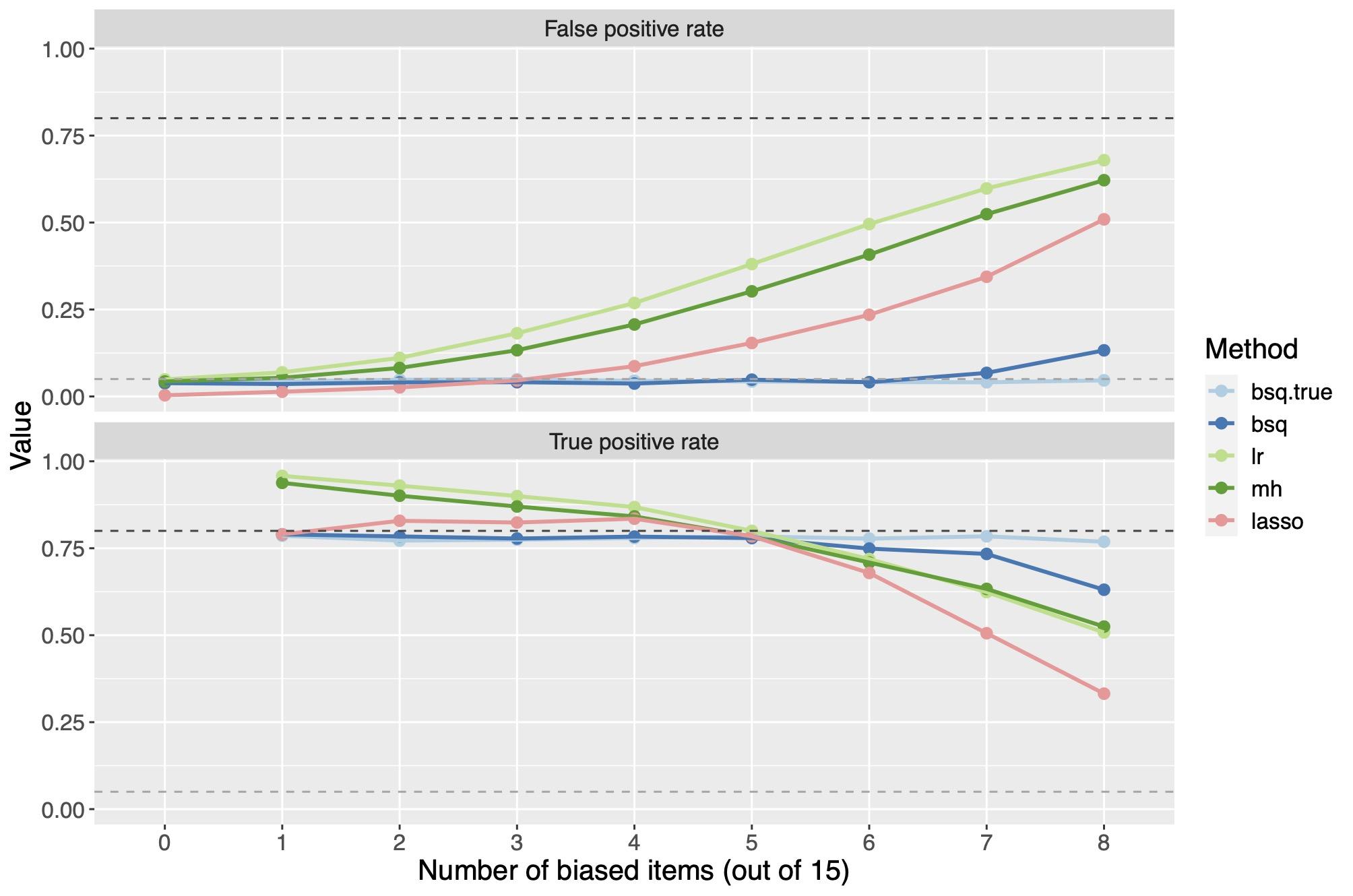
* El enfoque no requiere especificación de ítems ancla
* Los resultados teóricos garantizan que el procedimiento puede tolerar hasta el 50% de ítems con FDI
  + Los métodos tradicionales que usan anclas fallan con < 25%
* resultados de simulación de datos muestran que se desempeña bien en comparación con otros métodos

## Un estudio de simulación

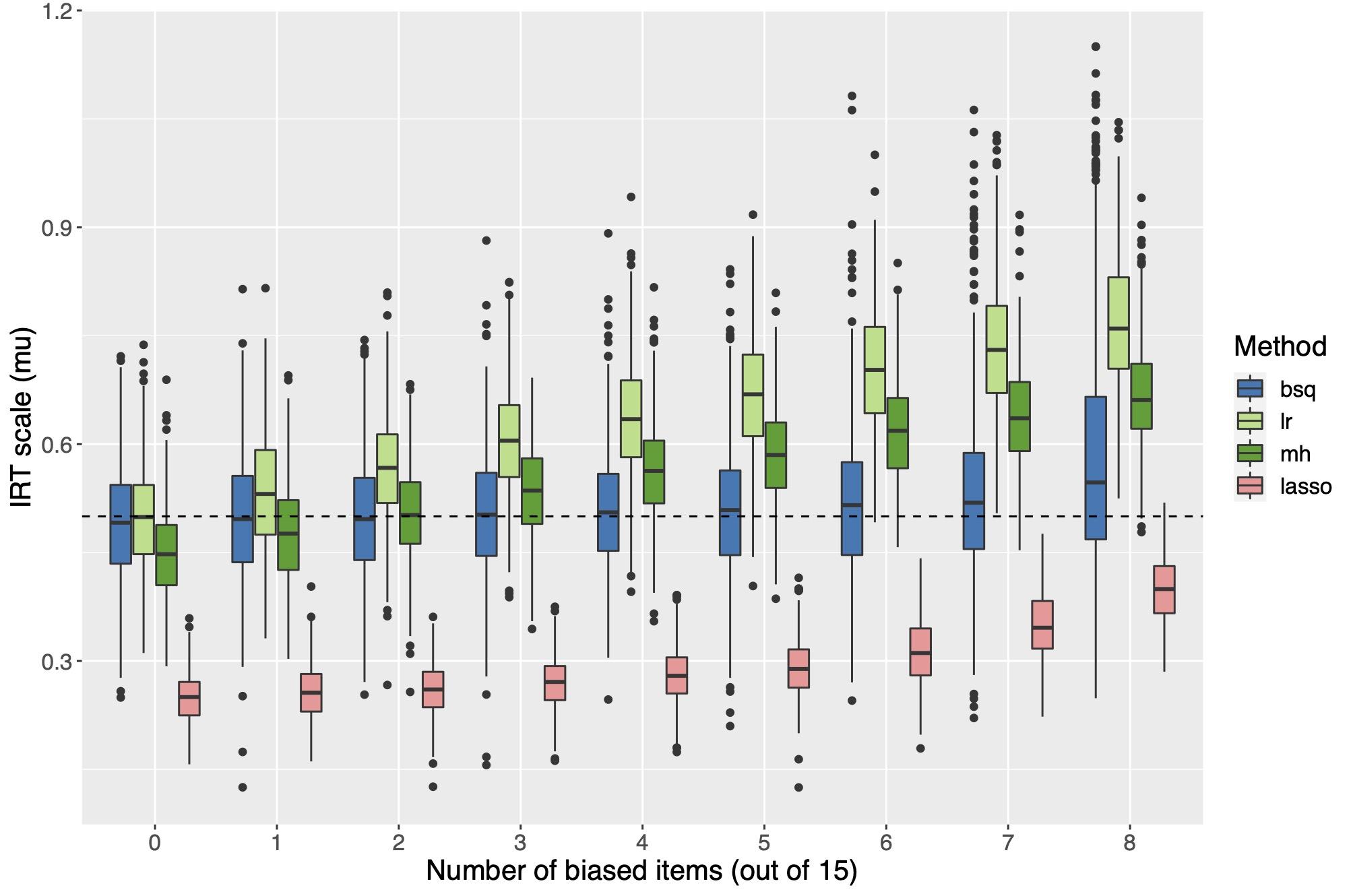
2PL en dos grupos, FDI en dificultad del ítem solo

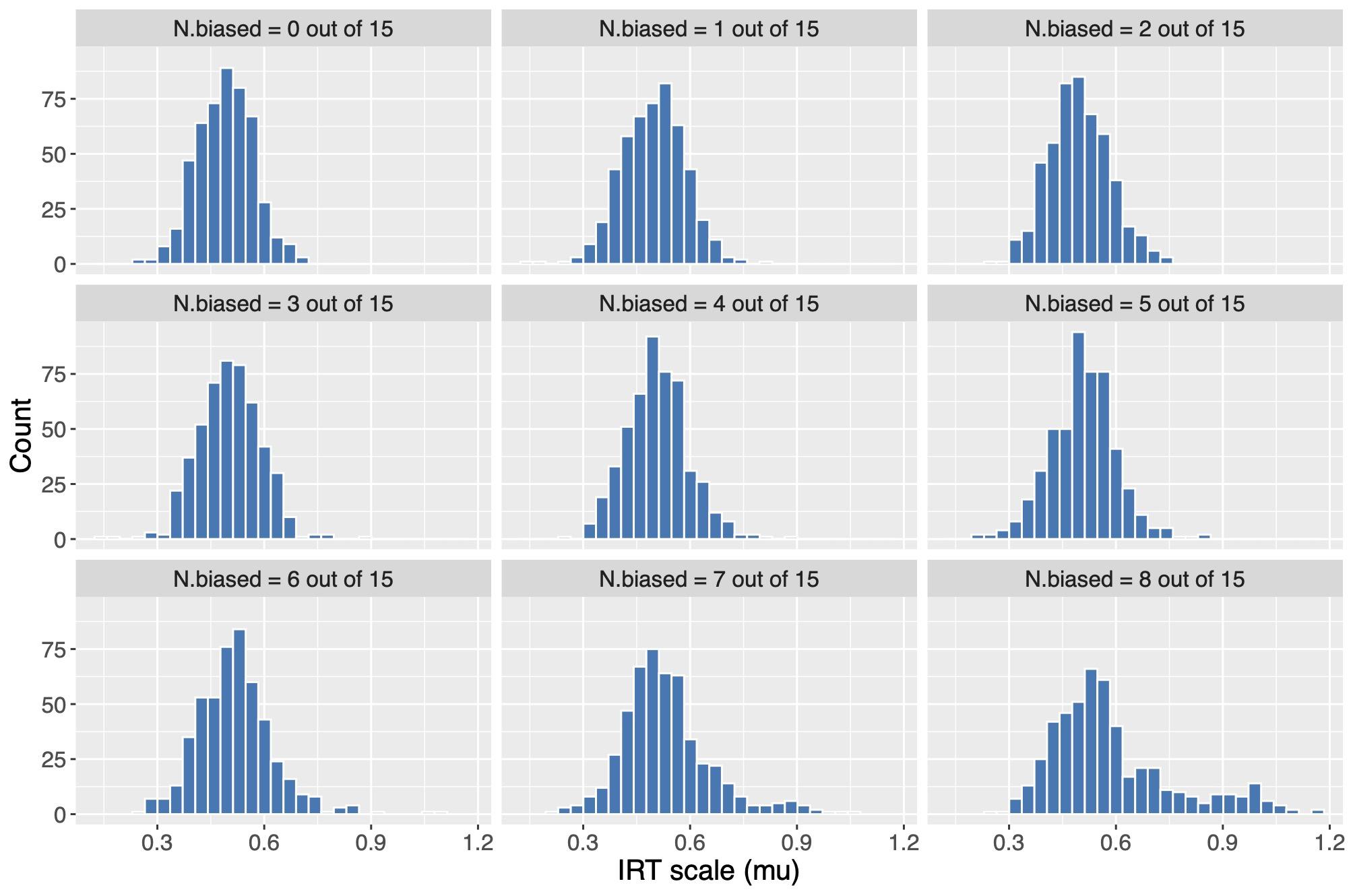
* FDI solo en las dificultades de los ítems (interceptos) ()
* Impacto solo en la media ()
* Factores focales
  + Número de ítems con DIF: 0 a , seleccionados aleatoriamente
  + Método: LRT-DIF, Mantel-Haenszel, GPCM lasso, estimador M propuesto
* Factores de diseño
  + repeticiones por número de ítems sesgados
  + personas por grupo
  + ítems
  + ;
  + ;

## Resultados de la simulación



## Resultados de la simulación



**Resultados de la simulación** 

# Ejemplo

## Configuración con R

• El paquete robustDIF ha sido actualizado recientemente, así que vamos a reinstalarlo ahora

# instalador para github  
install.packages(remotes)  
  
# instalar robustDIF desde gitihub  
remotes::install\_github("peterhalpin/robustDIF")  
  
# cargar biblioteca  
library(robustDIF)

## 

## Paso 1. Estimar el modelo IRT

• Puede utilizar el modelo configurar en mirt o la lista de dos ajustes separados

• lavaan también es compatible, vea el Apéndice

Loading required package: stats4

Loading required package: lattice

library(mirt)  
  
# Configurar datos para 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)  
  
# Estimatación del modelo (sin restricciones de invarianza)  
config.mod <- multipleGroup(depression\_items,  
 group = gender,  
 itemtype = "graded",  
 SE = T) # <- make sure to request SEs

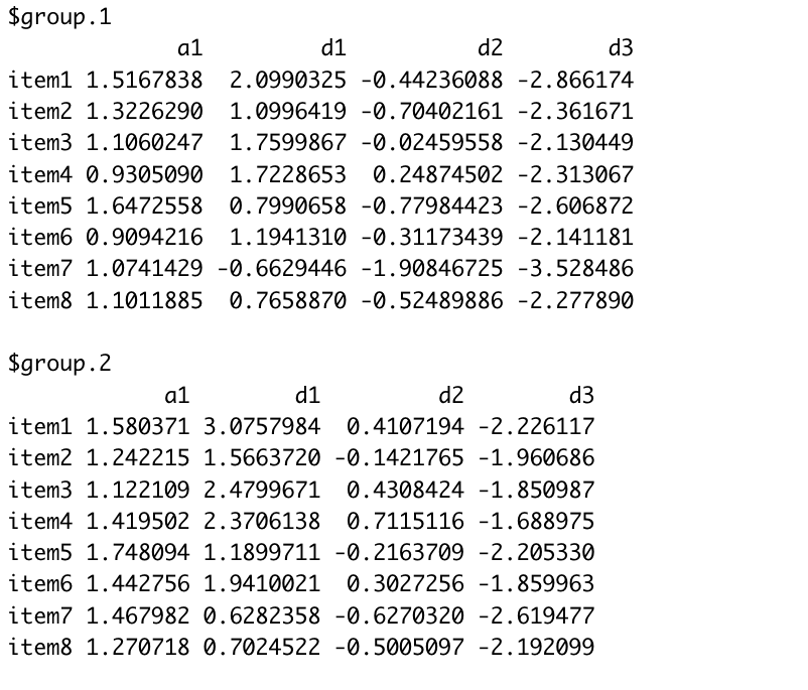
## 

## Paso 2. Extraer los parámetros del modelo

• Extraer los parámetros del modelo en mirt

# Extraer parámetros del modelo  
mirt.parms <- get\_model\_parms(config.mod)  
  
*## Comprobar el output*mirt.parms$est

## Paso 2. Extraer los parámetros del modelo



## Paso 3. Correr un análisis FDI robusto

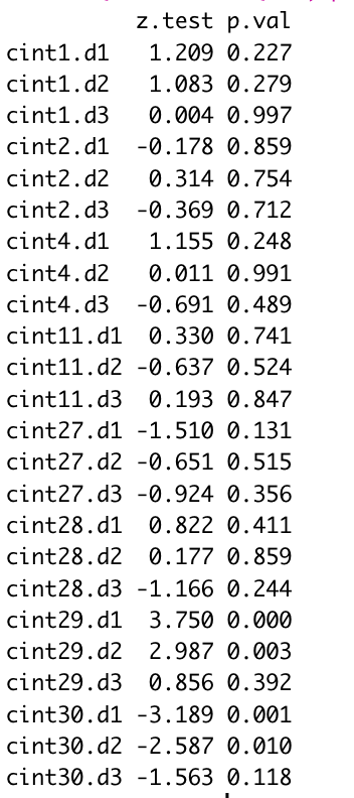
• rdif es una función interna para la estimación

• rdif\_z\_test y rdif\_chisq\_test para un output fácil de entender para los usuarios

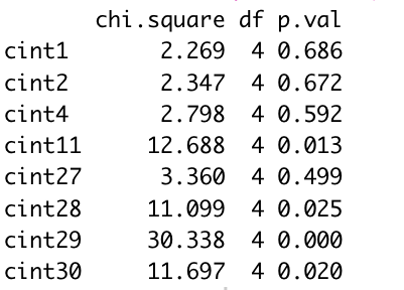
# "raw" ouput con ponderaciones  
rdif(mirt.parms, par = "intercepts")  
  
# z\_test para cada parámetro  
rdif\_z\_test(mirt.parms, par = "intercepts")  
  
# Pruebas chi\_sq para DIF a nivel de ítem (incluyendo las pendientes)  
rdif\_chisq\_test(mirt.parms)

## 

## Paso 3. Correr un análisis FDI robusto



## Paso 3. Correr un análisis FDI robusto



## Paso 3. Correr un análisis FDI robusto

• Nótese que en la prueba LR sólo cint29 y cint30 resultaron tener FDI

• Estudiaremos juntos el FDI en las pendientes de los ítems durante el taller

# Análisis DIF para pendientes de los ítems con mirt  
strong.invariance <- c("free\_mean", "free\_var", "slopes", "intercepts")  
strong.mod <- multipleGroup(depression\_items,  
 group = gender,  
 itemtype = "graded",  
 invariance = strong.invariance)  
  
DIF(strong.mod,  
 which.par = c("a1"),  
 scheme = "drop")  
  
# Y con robustDIF  
rdif\_z\_test(mirt.parms, par = "slopes")

## 

## Resumen del ejemplo

• Conclusiones similares a las de la prueba LR, pero no exactamente iguales

– Prueba de los umbrales de los ítems encontró los mismos ítems que la prueba LR

– La prueba de chi-cuadrado encontró 2 ítems adicionales con FDI (debido a las pendientes)

• Próximos pasos

– Puede seguirse con un modelo de invarianza parcial como antes para examinar los efectos a nivel de ítem.

– Podría considerarse revisar u omitir ítems …

• Otras preocupaciones

– Diferentes métodos de FDI conducen a conclusiones distintas (en general)

– ¿Afecta todo esto a las conclusiones sobre el impacto?

# FTD

• Funcionamiento diferencial de la prueba: ¿afecta el FDI a las conclusiones sobre el impacto?

## Resumiendo dónde estamos

• El análisis de FDI es acerca de los ítems

– Útil para el desarrollo de pruebas

• El análisis de FDI no proporciona una manera directa de inferir si el FDI afecta las conclusiones sobre el impacto

– ¡A menudo lo que nos importa en la investigación!

## Uso de FDI robusto para abordar el impacto

• Hasta ahora nos hemos centrado en usar una estimación robusta de los parámetros de escala (impacto) como una forma de probar el FDI en ítems individuales

• También podemos comparar la estimación robusta con una estimación “ingenua” que surgiría si ignoramos el FDI

– por ejemplo, la estimación de máxima verosimilitud (EMV) del impacto, basada en el mismo procedimiento de escalamiento

• Si las dos estimaciones dan el “mismo” resultado, entonces el FDI no afecta las conclusiones sobre el impacto

– es decir, si ignoramos el FDI, llegaríamos a la misma conclusión sobre cómo difieren los grupos

## Lógica de la prueba

• La lógica de esta prueba es similar a la prueba de especificación de Hausman

• Bajo la hipótesis nula, tanto la estimación robusta como la EMV son estimaciones consistentes (sin sesgo) del “verdadero” impacto

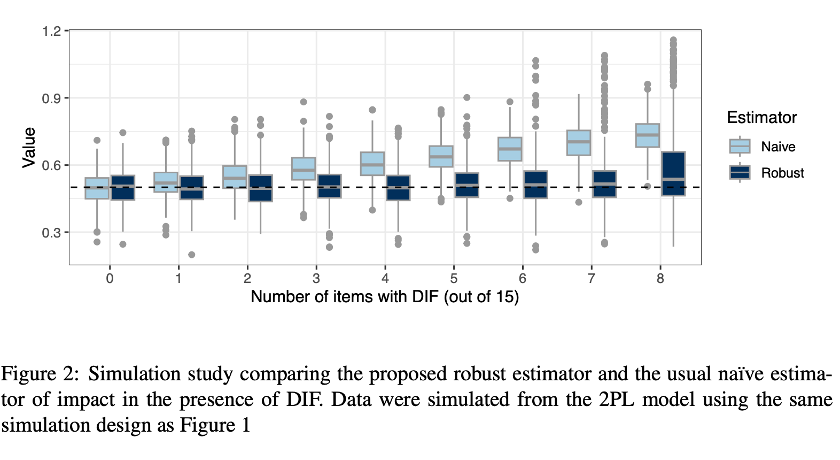
– La EMV es más eficiente, pero esto no es muy importante para nosotros

• Bajo la hipótesis alternativa, ambas pueden ser inconsistentes, pero la estimación robusta será menos sesgada

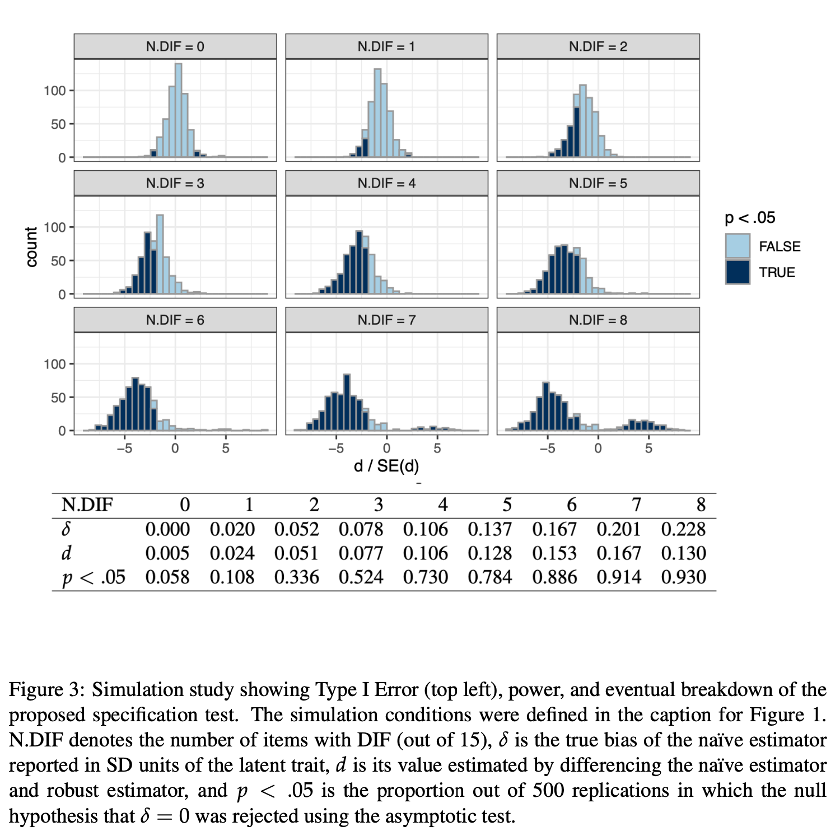
– Asumiendo < 50% de los ítems exhiben FDI

• Consecuentemente, la diferencia entre las estimaciones se puede usar para probar el efecto del FDI en el impacto

## Estudios de simulación



## Estudios de simulación



## Relación con la IM

• En IM, probamos si (un subconjunto de) los parámetros de los ítems son iguales entre grupos

• Podemos rechazar la IM con poco efecto en el impacto

– FDI en un solo ítem puede ser insignificante cuando se promedia el de muchos ítems

– El FDI en direcciones opuestas puede cancelarse entre los ítems

• En la prueba de Funcionamiento Diferencial de la Prueba (FDP)

– No probamos ningún parámetro del ítem (aunque podríamos disminuir su peso)

– Solo probamos si dos estimaciones de impacto son iguales

– ¡Tenga en cuenta que aún puede haber ítems con FDI incluso si el impacto no se ve afectado!

## Implicaciones para la práctica

• Si todo lo que queremos es comparar las puntuaciones de los grupos en las pruebas:

– Usando un enfoque robusto, podemos probar el FDP sin *antes* tener que hacer un análisis de FDI ítem por ítem

– Si no hay FDP, podemos proceder con las comparaciones entre grupos sin hacer análisis de FDI

– Si hay FDP, podemos seguir con análisis ítem por ítem, revisiones de pruebas, etc., antes de hacer comparaciones

# Volver al ejemplo

delta\_test(mirt.parms)

• Las estimaciones ingenuas y robustas de las diferencias de género promedio en depresión no difieren

# Resumen

• En el desarrollo de pruebas, casi siempre queremos saber sobre el FDI a nivel de ítem

• En contextos de investigación, a veces solo nos importa si las comparaciones entre grupos están sesgadas o no

• Usando un escalamiento robusto, podemos hacer inferencias sobre el FDP antes de realizar un análisis de FDI ítem por ítem

– Truco: comparar dos estimaciones de impacto, ingenua y robusta

• A diferencia de las pruebas de IM, no estamos probando si todos los ítems están libres de FDI

– Solo estamos probando si el FDI afecta las conclusiones sobre el impacto

• Si concluimos que no hay FDP, puede o no haber FDI

– ¡Si realmente queremos saber sobre los ítems individuales, necesitamos hacer el análisis de FDI!

# Conclusión

## Lo que hemos cubierto hoy

• Escalado basado en la Teoría de Respuesta al Ítem y su relación con el FDI

– Más información sobre el escalado en el apéndice

• Escalamiento robusto

– Vea Halpin 2022 para detalles técnicos

• Pruebas de FDI basadas en escalamiento robusto

– rdif\_z\_test y rdif\_chisq\_test

• Pruebas de FDP (impacto) basadas en escalamiento robusto

– delta\_test

• Ejemplo práctico

## Advertencias y futuras direcciones

• robustDIF está en las primeras etapas de desarrollo

– ¡Acabamos de agregar soporte para datos categóricos la semana pasada!

– Trabajando en múltiples grupos este invierno

– ¡Seguro que habrá muchos errores y problemas!

• Por favor, no dude en contactarme con preguntas sobre el software o ideas para nuevos desarrollos

– peter.halpin@unc.edu

## Referencias

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

## Referencias

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

# Apéndice

## Más sobre el escalado

• Especificar el modelo IRT de 2PL en forma de "pendiente-intersección"

• Grupo 1

• Grupo 2

• El escalado implica resolver para y

• Sabemos que esta relación se mantiene para alguna elección de y porque los modelos IRT solo se identifican hasta una transformación lineal de.

## Más sobre el escalado

• Sustituyendo en las ecuaciones de escalado y haciendo el álgebra da lugar a::

–

–

• Tomar la media sobre los ítems da el escalado “ingenuo” usual (en forma de pendiente-intersección)

–

–

• En el diseño CINEG, dejamos que los parámetros de ítem en el grupo de referencia sustituyan a los parámetros de ítem “no escalados”

–

–

## Ejemplo con 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 (not 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)