Parte 2: Funcionamiento Diferencial de los Ítems

Peter F. Halpin

## Descripción General del Taller

* Parte 1. Introducción + análisis factorial + IM
* Parte 2. TRI + FDI 🡨
* Parte 3. Escalamiento robusto + FDI + DTF

## Descripción General de la Parte 2

* Revisar los objetivos del análisis de FDI
* Datos binarios TRI (2PL)
* TRI para datos categóricos ordenados (GRM)
  + El Modelo de respuesta Graduado es el análogo TRI del análisis factorial categórico
* Pruebas de FDI usando la prueba de razón de verosimilitudes
  + Selección de ítems ancla
* 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 (traducido, editable)
* Código: Solo el código de estas diapositivas

# Objetivos del análisis de FDI

* Aseguramiento de la calidad ítem por ítem

## Recapitulación

* Los modelos psicométricos plantean dos explicaciones no mutuamente excluyentes de por qué la distribución de las puntuaciones de las pruebas puede diferir entre grupos de encuestados
  1. Impacto: los grupos difieren en el rasgo que se está midiendo
  2. FDI: la medida está sesgada con respecto a la pertenencia a un grupo
* El objetivo del análisis de FDI es detectar ítems sesgados sin hacer suposiciones sobre el impacto

## ¿FDI como seguimiento a IM?

* En el marco de análisis factorial
  + IM: prueba de la hipótesis de "no FDI" sobre todos los ítems
  + FDI: procedimientos para el seguimiento del rechazo de IM
    - También llamado IM parcial
    - Vea semTools::partialInvariance()

## ¿FDI como seguimiento a IM?

* En el marco de TRI
  + El objetivo de FDI es tomar decisiones a nivel de ítem (mantener, revisar, omitir)
  + Los niveles de IM no se traducen directamente en decisiones a nivel de ítem
  + Por lo tanto, los analistas de FDI generalmente omiten IM (vea Thissen, 2023)

## Visión general

* Modelo de medición genérico: IM débil / métrica

|  |
| --- |

## Visión general

* Modelo de medición genérico: IM fuerte / escalar

|  |
| --- |

## Visión General

* Modelo de medición genérico: IM Estricta

|  |
| --- |

## Visión General

* Modelo de medición genérico: FDI ítem por ítem

|  |
| --- |

## Visión General

* Modelo de medición genérico: FDI ítem por ítem

|  |
| --- |

## Visión General

* Modelo de medición genérico: FDI ítem por ítem

|  |
| --- |

## Resumen

* El IM y el FDI tratan el mismo tema
* Pero de manera diferente
  + IM: invarianza de parámetros entre grupos
  + FDI: invarianza de ítems entre grupos
* Conceptualmente, podrían combinarse en una "gran teoría"
* En la práctica, diferentes modelos, diferente software, diferentes entornos de investigación, diferentes tradiciones …

# TRI

* Datos binarios, luego datos categóricos ordenados

## TRI

* Hay muchos modelos de TRI
* Nos enfocaremos en dos modelos análogos al análisis factorial para datos categóricos
* En términos de análisis de FDI, los mismos principios aplican independientemente del modelo que se utilice

## Nota sobre terminología

* TRI fue desarrollado en el contexto de pruebas educativas
* Mucha de la terminología del modelo refleja este contexto
  + p. ej., hablamos sobre la dificultad de los ítems, la habilidad de los respondientes
* Otra terminología es más adecuada en otros contextos
  + p. ej., la severidad de los síntomas, la depresión de los respondientes
* Es común usar diferente terminología dependiendo del contexto

## Nota sobre símbolos matemáticos

* En TRI, usualmente la letra “” denota el rasgo latente (habilidad)
  + Utilizaré esta terminología para ser consistente con el software de TRI
* Como en el análisis factorial, usualmente asumimos cuando consideramos un solo grupo o población
  + Abordaremos cómo estimar el impacto más adelante

## El modelo logístico de 2 parámetros (2PL)

* El 2PL es aplicable a respuestas de ítems binarios, por ejemplo:
  + correcto / incorrecto
  + sí / no
  + aprobado / no aprobado
* También es un bloque de construcción para modelos con > 2 categorías de respuesta ordenadas

## Funciones de respuesta al ítem (FRI)

* Es costumbre presentar los modelos de TRI en términos del modelo de medición para cada ítem:
* Para datos binarios:
* Esto se llama la función de respuesta al ítem (FRI)
* Describe cómo la probabilidad de aprobar un ítem depende del nivel del rasgo que se está midiendo

## La FRI de 2PL

* Esto es solo la función logística de la regresión logística
* se llama el parámetro de discriminación del ítem
  + Pendiente de la regresión logística, similar a las cargas factoriales
* se llama el parámetro de dificultad del ítem
  + Intercepto re-escalado

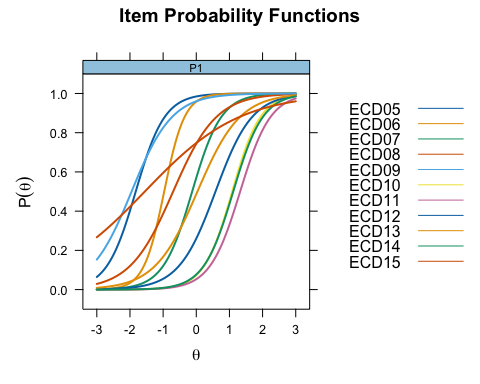
## Ejemplos de FRI de 2PL

library(mirt)

Loading required package: stats4

Loading required package: lattice

path\_to\_data <- "ECDI\_learning.csv"   
learning\_items <- read.csv(path\_to\_data)  
  
# Run the 2Pl model analysis  
twoPL <- mirt(learning\_items, model = 1, itemtype = "2PL", verbose = F)  
plot(twoPL, type = "trace", facet = F, theta\_lim = c(-3, 3), lwd = 2)

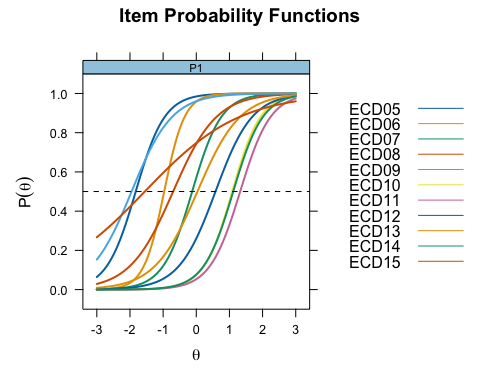


## Interpretación de los parámetros del modelo

* El parámetro se denomina **dificultad del ítem**
* Nota que implica
* Por lo tanto, la dificultad del ítem es el valor de en el cual la probabilidad de aprobar el ítem es igual a 1/2
* Los encuestados con habilidad por encima del nivel de dificultad del ítem tienen una probabilidad > 1/2 de responder correctamente al ítem, y viceversa

## Interpretación de los parámetros del modelo

plot(twoPL, type = "trace", facet = F, theta\_lim = c(-3, 3), lwd = 2,  
 panel=function(...) {panel.xyplot(...)  
 panel.abline(h=1/2, lty = 2)})



* Dificultad es el nivel del rasgo () requerido para tener una probabilidad ≥ 1/2 de aprobar el ítem

## Interpretación de los parámetros del modelo

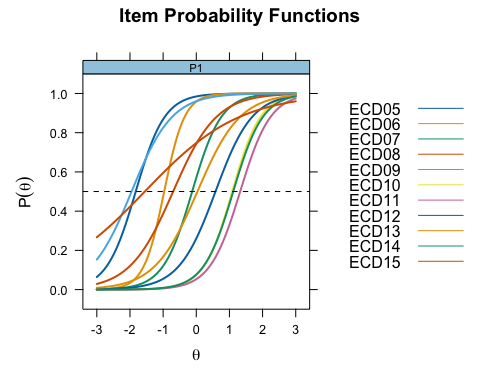
* El parámetro se llama **discriminación del ítem**
* Los ítems con mayor discriminación tienen pendientes más pronunciadas, y relación más fuerte con el rasgo latente:
* La tasa de cambio de la FRI es mayor para los ítems que más discriminan

## Interpretación de los parámetros del modelo

* Si nuevamente sustituimos , tenemos
* Para valores "cercanos a" el parámetro de dificultad, el parámetro de discriminación es proporcional a la pendiente

## Interpretación de los parámetros del modelo

plot(twoPL, type = "trace", facet = F, theta\_lim = c(-3, 3), lwd = 2,  
 panel=function(...) {panel.xyplot(...)  
 panel.abline(h=1/2, lty = 2)})



* Discriminación es (proporcional a) la pendiente de la curva donde se intersecta con .

## Resumen

* El modelo TRI de 2PL:
  + Se utiliza para modelar la probabilidad de aprobar un ítem binario
  + El parámetro de dificultad describe el nivel del constructo objetivo en el cual la probabilidad de aprobación = 1/2
  + El parámetro de discriminación describe cuán fuertemente cada ítem está relacionado con el constructo objetivo

# Otros Conceptos de TRI

* No requeridos para el análisis de FDI, pero relevantes para entender la teoría

## Información en TRI

* Deje que denote la estimación de máxima verosimilitud (EMV) de
* En términos prácticos, es la "mejor" estimación del rasgo que podemos obtener de una evaluación
* El error estándar de la EMV, describe con qué precisión podemos estimar
* *Información* definida como la precisión de la EMV,
  + No es fácil de interpretar, pero valores más grandes significan estimaciones más precisas

## Información en TRI

* Una de las principales contribuciones de TRI es modelar cómo la información depende de los parámetros de los ítems de la prueba
* ¡Esto proporciona una buena teoría para el desarrollo de pruebas! Nos dice cómo construir pruebas con un nivel deseado de precisión / información
* Abordamos la información aquí por completitud, pero no es requerida para el análisis de FDI

## Función de Información del Ítem (FII)

* La función de información del ítem (FII) es la precisión que resulta al estimar el rasgo latente usando un solo ítem
* En la práctica, nunca usaríamos solo un ítem en una prueba
* Pero, podemos construir la función de información de toda la prueba a partir de la de cada ítem individual
* Por lo tanto, comenzamos con la FII y luego usamos eso para obtener la función de información de la prueba (FIP)

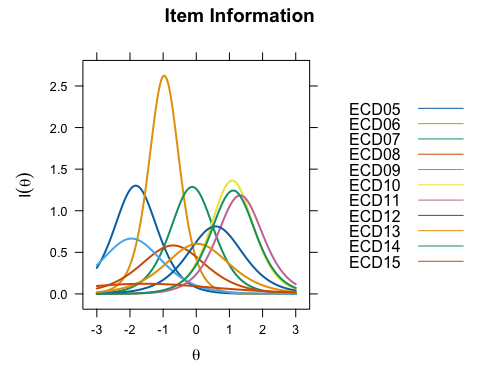
## Función de Información del Ítem (FII)

Para el 2PL, la FII es:

* La información aumenta con
  + Los ítems que están más fuertemente relacionados con el rasgo latente (que más discriminan) proporcionan más información sobre el rasgo
* El máximo de la FII para cada ítem ocurre cuando
  + Cada ítem proporciona la mayor información en su nivel de dificultad

## FIIs

plot(twoPL, type = "infotrace", facet = F, theta\_lim = c(-3, 3), lwd = 2)



* La ubicación del pico es ; la altura del pico es

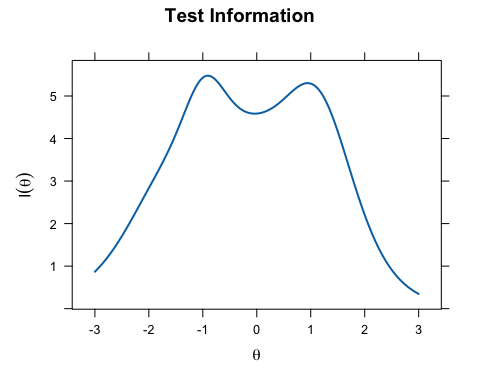
## Función de Información de la Prueba (FIP)

La FIP se obtiene sumando las funciones de información de todos los ítems de una prueba

* Este resultado se sigue directamente del supuesto de independencia condicional

## FIP

plot(twoPL, type = "info", theta\_lim = c(-3, 3), lwd = 2)



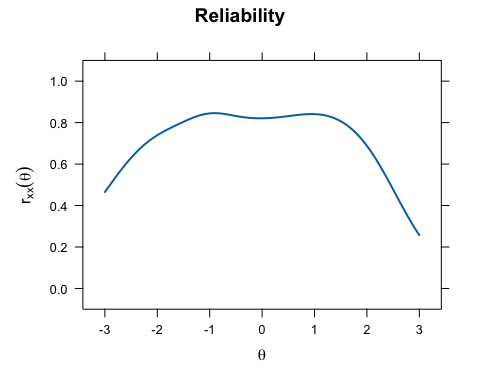
* Útil para comparar diferentes pruebas, pero los valores no se interpretan fácilmente…

## Confiabilidad

* La FIP se puede convertir en una función de confiabilidad para la puntuación total (vea Nicewander, 2018):
* Promediando esta función sobre los valores de proporciona un coeficiente de confiabilidad "marginal"
  + Interpretado como la proporción de variabilidad en la puntuación total que está asociada con el rasgo (como el alfa de Cronbach)

## Confiabilidad

plot(twoPL, type = "rxx", theta\_lim = c(-3, 3), lwd = 2)



Confiabilidad marginal de la puntuación total:

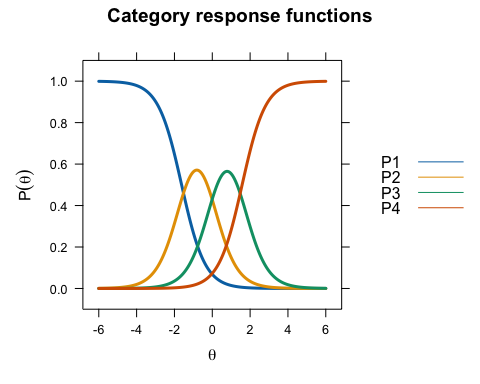
marginal\_rxx(twoPL)

[1] 0.8148199

## Resumen

* Un concepto central en TRI es la información (de Fisher), que es la precisión con la que se puede estimar el constructo objetivo
* Las funciones de información de los ítems describen la información proporcionada por cada ítem
* La información de la prueba es la suma de la información de los ítems
* La confiabilidad de la puntuación total se puede derivar de la información de la prueba
* La información es útil para comparar diferentes ítems / pruebas, pero la confiabilidad es más fácil de interpretar

# El Modelo de Respuesta Graduada (MRG)



## Datos categóricos ordenados

* Permita que la respuesta del ítem tome valores , donde el número de categorías para el ítem
  + p. ej., si un ítem tiene 4 posibles categorías de respuesta, y son las 4 categorías de respuesta
* No importa cómo etiquetemos las categorías siempre que estén ordenadas
* Para CINT: nunca (0), raramente (1), a veces (2), o casi siempre (3)

## Funciones de respuesta al ítem

La *función de respuesta acumulativa* es la probabilidad de aprobar la categoría o superior, condicional a

* Por definición, , , y
* *se llaman las funciones de respuesta de categoría de ítem (FRCI)*

## MRG

* MRG asume un modelo 2PL para las funciones de respuesta acumulativas
* Cada ítem tiene solo una discriminación (suposición de probabilidades proporcionales)
* Cada categoría de respuesta tiene su propio parámetro de dificultad, ahora llamado parámetro de "umbral"

## MRG

* A partir de la función de respuesta acumulativa, derivamos las FRCI
* Estas dan la probabilidad de aprobar cada categoría
* Note que esto se reduce a 2PL para

## MRG: Ejemplo

library(mirt)  
  
# Load data and separate depression items  
cint <- read.csv("cint\_data.csv")  
depression\_names <- c("cint1", "cint2", "cint4", "cint11",   
 "cint27", "cint28", "cint29", "cint30")  
depression\_items <- cint[, depression\_names]  
  
# Run GRM model  
grm <- mirt(depression\_items,   
 itemtype = "graded")  
  
# per item plots   
itemplot(grm,   
 item = 1,   
 type = "threshold",   
 main = "Cumulative response functions")  
  
itemplot(grm,   
 item = 1,   
 type = "trace",   
 main = "Category response functions")  
  
# Plotting all test items   
plot(grm,   
 type = "itemscore",   
 main = "Expected item score functions",   
 facet = F)

## MRG: Ejemplo

|  | * Gráficos para CINT.1 (“Se siente triste o deprimido”) * Las funciones de respuesta acumulativa (arriba) no se informan habitualmente, se muestran aquí para ilustrar la suposición de 2PL * Las FRCI (abajo) suelen informarse, muestran la probabilidad de aprobar cada categoría * Los valores de los umbrales de los ítems se muestran con líneas verticales discontinuas   + “dificultades” en las funciones de respuesta acumulativa   + “umbrales” o límites de categoría en FRCI |
| --- | --- |

## MRG: Ejemplo

|  | * Trazar la puntuación esperada del ítem es una forma de simplificar la presentación de toda la evaluación * Muestra cómo la respuesta esperada (0-3) depende del rasgo medido * Calculado como * De izquierda a derecha, ítems que fueron “más fáciles” en el sentido de que se esperaban puntuaciones de ítem más altas para valores más bajos de depresión |
| --- | --- |

## MRG: Ejemplo

coef(grm, IRTpars = T, simplify = T)

$items  
 a b1 b2 b3  
cint1 1.612 -1.622 -0.011 1.578  
cint2 1.286 -1.049 0.306 1.665  
cint4 1.129 -1.879 -0.195 1.758  
cint11 1.219 -1.688 -0.402 1.608  
cint27 1.692 -0.594 0.279 1.404  
cint28 1.213 -1.292 -0.001 1.641  
cint29 1.351 -0.004 0.889 2.217  
cint30 1.155 -0.630 0.435 1.910  
  
$means  
F1   
 0   
  
$cov  
 F1  
F1 1

* Generalmente, se presentan gráficos para resumir modelos, en lugar de tablas de coeficientes
* Las FII, la FIP y la confiabilidad se proporcionan en el Apéndice

## Resumen

* El MRG es ampliamente utilizado para respuestas categóricas ordenadas
* Las funciones de respuesta acumulativa se modelan utilizando 2PL
  + Cada ítem tiene la misma discriminación para todas las categorías, suposición de probabilidades proporcionales
  + Cada ítem tiene un parámetro de umbral diferente para cada categoría
* Las FRCI se derivan de las funciones de respuesta acumulativa
  + Estas dan la probabilidad de aprobar cada categoría

# FDI

## Descripción General

* El objetivo del análisis de FDI es detectar ítems sesgados sin hacer suposiciones sobre el impacto
* Entonces, queremos probar si los parámetros de los ítems difieren entre grupos
* Hay muchas formas de hacer esto, pero para el ejemplo nos centraremos en la prueba de razón de verosimilitud para modelos anidados
  + Mismo enfoque utilizado para el análisis factorial
* Si un ítem muestra FDI, debe ser investigado (por ejemplo, revisar, omitir)
* Suena lo suficientemente simple, pero…

## FDI son dos problemas interrelacionados

* El problema más obvio: Inferir si los parámetros de los ítems difieren como función de alguna(s) variable(s) externa(s)
* Para fines ilustrativos, considere la prueba de Wald de Lord para el parámetro de dificultad de 2PL en grupos g =
* (Utilizaremos la prueba de razón de verosimilitud más adelante, pero la prueba de Lord es simple de interpretar)

## FDI son dos problemas interrelacionados

* El problema menos obvio: Los modelos TRI están identificados solo hasta una transformación lineal del rasgo latente
* Esto significa que los parámetros de los ítems y el rasgo latente pueden transformarse linealmente sin cambiar las FRI
* Deje que , , y :
* Esta es la razón técnica por la que necesitamos establecer la escala del rasgo latente al estimar modelos psicométricos
* Establecer que implica que y , lo cual resuelve el problema

## Implicaciones de la escala para la prueba de Lord

* Si escalamos el rasgo latente para tener la misma media y varianza en ambos grupos, esto tiene implicaciones para probar los parámetros del modelo
* Las transformaciones de escala son

Introduciendo los parámetros de los ítems reescalados en la prueba de Lord

Conclusión: Si hay impacto en la media o la varianza del rasgo latente, la prueba de Lord está sesgada

## ¿Cómo resolvemos el problema de escala?

* Paso 1. Escalar arbitrariamente el rasgo latente en el "grupo de referencia"
  + Justificado porque los modelos TRI para un solo grupo están identificados solo hasta una transformación lineal del rasgo latente
* Paso 2. Suponer que (algunos de) los parámetros de los ítems son iguales entre grupos
  + Estos ítems se llaman anclas
  + Basta para escalar el rasgo latente en el grupo de comparación}
  + p. ej., establecer para al menos 2 ítems y resolver para y en la diapositiva anterior
* Conclusión: Necesitamos saber los ítems sin FDI (anclas) para escalar el rasgo latente

## La “naturaleza circular” del FDI

* El problema descrito ha sido referido como la naturaleza circular del FDI (Angoff, 198)
  + Queremos comparar el valor de los parámetros del modelo entre grupos
  + Para hacer esto debemos escalar el rasgo latente en ambos grupos
  + Para escalar el rasgo latente, debemos suponer que algunos parámetros del modelo son iguales entre grupos
  + ¡Pero esto es lo que queríamos probar en primer lugar!

## Ítems Ancla

* En la práctica, el problema se resuelve eligiendo un "conjunto ancla" de ítems
* Las anclas son ítems que tratamos como libres de FDI al probar otros ítems para FDI
* Hay muchas estrategias, heurísticas, etc., para elegir anclas
* Todas estas tienen falencias: la selección de ítems ancla es una limitación de los métodos tradicionales para el análisis de FDI
  + Más sobre esto en la última parte de este taller

## Purificación y Refinamiento en Dos Etapas

* Un enfoque ampliamente utilizado para elegir anclas es llamado purificación y refinamiento
  + Etapa 1. Probar cada ítem asumiendo que todos los demás ítems son anclas. Antes de comenzar la etapa dos, eliminar cualquier ítem con FDI del conjunto ancla ("purificación")
  + Etapa 2. Probar los ítems sin FDI nuevamente, usando el conjunto ancla purificado
* Se puede repetir según se desee
* El procedimiento es exploratorio, implica muchas pruebas de FDI, ajustando modelos con diferentes restricciones, …

# Prueba de FDI con la Prueba de Razón de Verosimilitud (LR)

## La Prueba LR

La prueba LR usa un modelo TRI multi-grupo para probar si los parámetros de un ítem difieren entre grupos

* Esta prueba es aplicable a cualquier modelo de TRI
  + Se centrará en 2PL por simplicidad, pero la ilustración usará GRM
* Este enfoque está sujeto a inquietudes sobre la selección de ítems ancla mencionada anteriormente
  + Necesitamos elegir un conjunto de ítems que se cree no tienen FDI al probar cuáles ítems sí tienen FDI

## Prueba LR para 2PL

Escriba el 2PL en dos grupos de la siguiente manera:

* El segundo subíndice en las FRI y el parámetros de los ítems indica el grupo de referencia (0) o el grupo de comparación (1)
* En el grupo de referencia, escalamos el rasgo latente arbitrariamente
  + Usualmente, estandarizado para tener y
* En el grupo de comparación estimamos la media y la varianza del rasgo latente
  + La justificación para esta configuración se discutió anteriormente al abordar el problema de escalado multi-grupo

## Prueba LR para 2PL

* Para aplicar la prueba LR, estimamos los siguientes dos modelos
* **Modelo 1**: El modelo anidado (más pequeño) se obtiene estableciendo todos los parámetros de los ítems iguales entre grupos
* Igual que la **invarianza fuerte** en el análisis factorial
* **Modelo 2**: El modelo anidante (más grande) se obtiene permitiendo que los parámetros del ítem focal varíen entre grupos:
* Note que no estamos requiriendo que los ítems sean desiguales, pueden ser iguales o desiguales, y simplemente permitimos que sean estimados libremente en cada grupo
* El software automatiza el ajuste de estos modelos ítem por ítem

## Prueba LR para 2PL

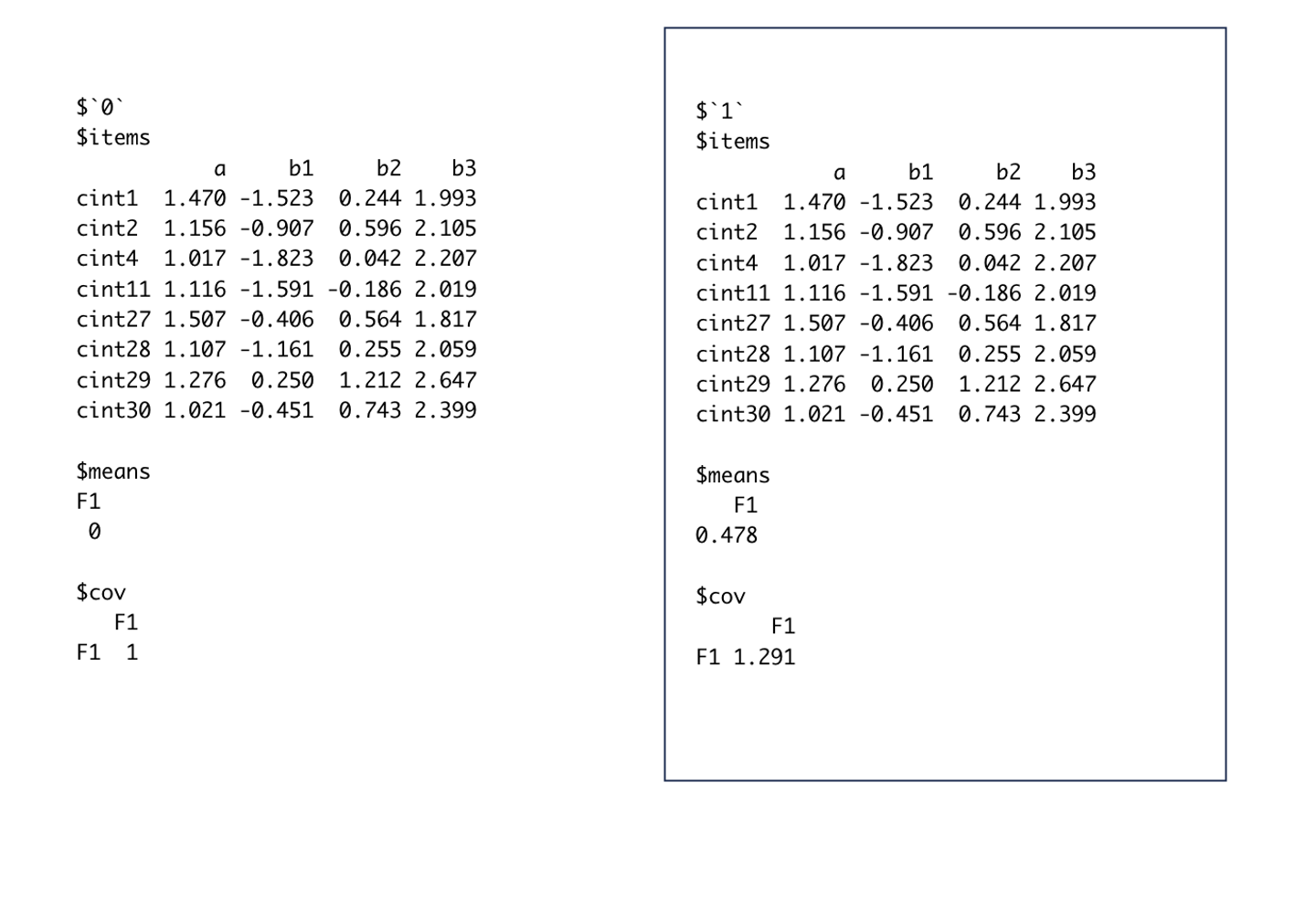
* La prueba LR procede comparando la verosimilitud del modelo anidado con el modelo anidante
* Cuando las restricciones impuestas por el modelo anidado son válidas (es decir, si no hay FDI en el ítem), esta prueba tiene una distribución chi-cuadrado con grados de libertad igual al número de parámetros restringidos
* Si la prueba LR de FDI es significativa, concluimos que el ítem está sesgado
* Si no, entonces concluimos que el ítem no está sesgado

## Invarianza Estricta: Código

* Paso 1. Estimar un modelo en el que las pendientes y los interceptos de los ítems son invariantes entre grupos (invarianza fuerte)

# Groups need to be a factor   
gender <- factor(cint$cfemale)  
  
# Invariance constraints used by mirt  
strong.invariance <- c("free\_mean", "free\_var", "slopes", "intercepts")  
  
# Estimate model (can request SE using SE = T)  
strong.mod <- multipleGroup(depression\_items,   
 group = gender,   
 itemtype = "graded",  
 invariance = strong.invariance)  
  
# View output  
coef(strong.mod, IRTpars = T, simplify = T)

## Prueba de FDI usando la prueba LR: Resultados



## Prueba de FDI usando la prueba LR: Código

* Paso 2. Ejecutar análisis de FDI (sin purificación)

DIF(strong.mod,   
 which.par = c("a1", "d1", "d2", "d3"), # <- mirt notation  
 scheme = "drop") # <- drop item constraints

groups converged AIC SABIC HQ BIC X2 df p  
cint1 0,1 TRUE 3.113 9.344 10.370 22.047 4.887 4 0.299  
cint2 0,1 TRUE 4.692 10.923 11.949 23.626 3.308 4 0.508  
cint4 0,1 TRUE 3.169 9.400 10.425 22.102 4.831 4 0.305  
cint11 0,1 TRUE 1.948 8.178 9.204 20.881 6.052 4 0.195  
cint27 0,1 TRUE 0.629 6.860 7.886 19.563 7.371 4 0.118  
cint28 0,1 TRUE 3.090 9.321 10.347 22.024 4.91 4 0.297  
cint29 0,1 TRUE -23.411 -17.180 -16.154 -4.477 31.411 4 0  
cint30 0,1 TRUE -9.195 -2.964 -1.939 9.738 17.195 4 0.002

## Prueba de FDI usando la prueba LR: Código

* Paso 2. Ejecutar análisis de FDI (con purificación)

DIF(strong.mod,   
 which.par = c("a1", "d1", "d2", "d3"),   
 scheme = "drop\_sequential", #<- different scheme  
 seq\_stat = .05, # <- Type I Error rate for DIF  
 max\_run = 2) # <- two stages only

Checking for DIF in 6 more items  
Computing final DIF estimates...

groups converged AIC SABIC HQ BIC X2 df p  
cint29 0,1 TRUE -18.863 -12.632 -11.606 0.071 26.863 4 0  
cint30 0,1 TRUE -4.647 1.584 2.610 14.286 12.647 4 0.013

## Resumen del ejemplo

* El análisis de FDI identificó dos ítems que estaban sesgados con respecto al género
  + CINT 29: “Ir a su habitación y llorar”
  + CINT 30: “Sentirse inquieto y caminar de un lado a otro”
* Más preguntas:
  + ¿Dirección y tamaño del efecto?
  + ¿Afecta el FIÍ las conclusiones sobre el impacto?
* Una forma de investigar estas preguntas:
  + Ajustar un modelo que permite que los ítems con FDI varíen entre grupos

## Análisis de seguimiento: Código

* Ajustar el modelo de invariancia parcial

# Invariance constraints  
partial.invariance <- c("free\_mean", "free\_var",   
 "cint1", "cint2", "cint4", "cint27", "cint28")  
  
# Estimate model  
partial.mod <- multipleGroup(depression\_items,   
 group = gender,   
 itemtype = "graded",  
 invariance = partial.invariance)  
  
  
# Plot IRFs of biased items  
itemplot(partial.mod, type = "score", item = "cint29", main = "CINT 29")  
itemplot(partial.mod, type = "score", item = "cint29", main = "CINT 30")  
  
# Examine parameter estimates  
coef(partial.mod, IRTpars = T, simplify = T)

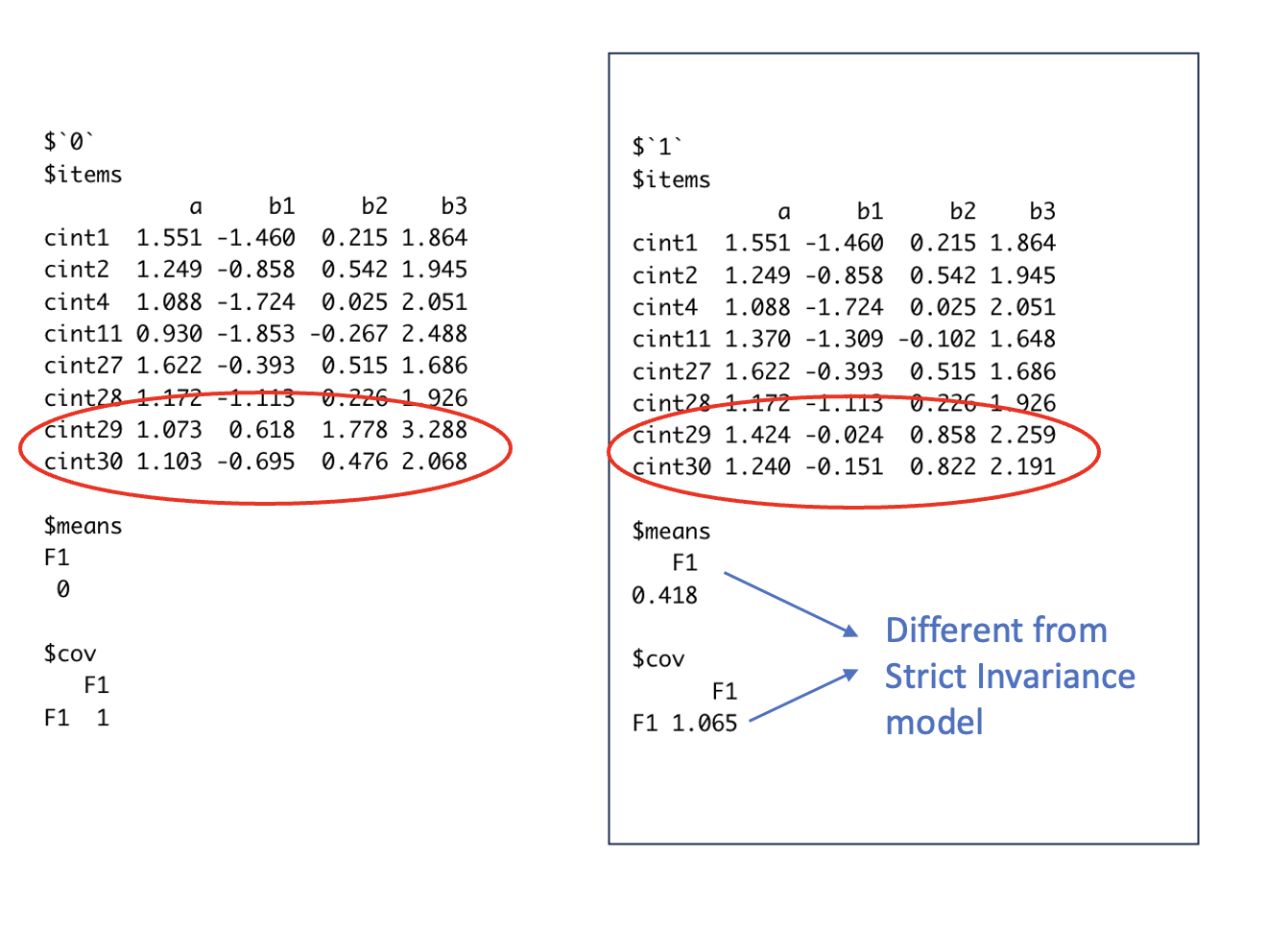
## Análisis de seguimiento: Resultados

* Se esperaba que las mujeres reportaran puntuaciones más altas en CINT 29 (llora)
* Se esperaba que los hombres reportaran puntuaciones más altas en CINT 30 (paces)
* Esto es después de controlar por el nivel de depresión

| A graph of a line  Description automatically generated  A graph of a function  Description automatically generated |  |
| --- | --- |

|  |  |
| --- | --- |

## Análisis de seguimiento: Resultados



## Resumen del ejemplo

* El análisis de FDI identificó dos ítems que estaban sesgados con respecto al género
  + CINT 29: Ir a su habitación y llorar
  + CINT 30: Sentirse inquieto y caminar de un lado a otro
* En CINT 29, se esperaba que las mujeres reportaran puntuaciones más altas que los hombres, incluso si tenían el mismo nivel de depresión
* En CINT 30, lo contrario fue cierto
* Las diferencias de medias en depresión cambiaron cuando se permitió que los ítems con FDI variaran entre grupos (invariancia parcial)
  + Las diferencias de medias se redujeron aproximadamente .06 SD
* Una limitación de los métodos actuales es que no podemos probar directamente si el FDI afecta las conclusiones sobre el impacto
  + Más sobre este tema en la parte 3

# Próximos pasos

## Resumen

* Hemos visto cómo probar el FDI utilizando modelos TRI (2PL GRM)
* En nuestro ejemplo, encontramos dos ítems sesgados con respecto al género
  + Estos ítems podrían considerarse para revisión o eliminación
* Discutimos dos limitaciones del análisis de FDI
  + Selección de ítems ancla
  + No hay prueba directa de si el FDI afecta el impacto

## Lo que haremos a continuación

* Nuevos procedimientos para abordar el FDI y el DTF
* No requieren la selección de ítems ancla
* Garantizados para funcionar si < 50% de los ítems muestran FDI
  + Diagnósticos disponibles cuando se enfrentan a proporciones mayores de ítems sesgados
* Se pueden utilizar para probar si el FDI afecta el impacto (¡sin tener que probar primero el FDI en cada ítem!)
* Fácil de implementar

## Referencias

Angoff, W. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. Berk (Ed.), Handbook of Methods for Detecting Test Bias (pp. 96–116). The Johns Hopkins Press.

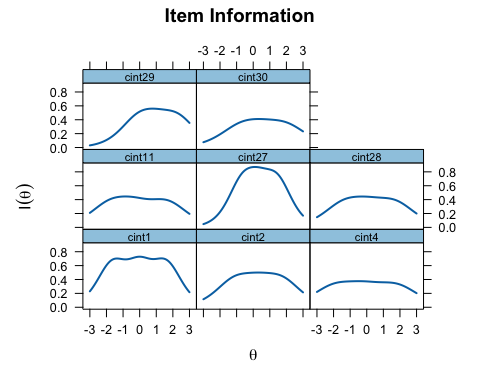
Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. Psychological Methods, 23(2), 351–362. https://doi.org/10.1037/met0000132

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), Differential Item Functioning (pp. 67–113). Lawrence Erlbaum Associates.

# Apéndice

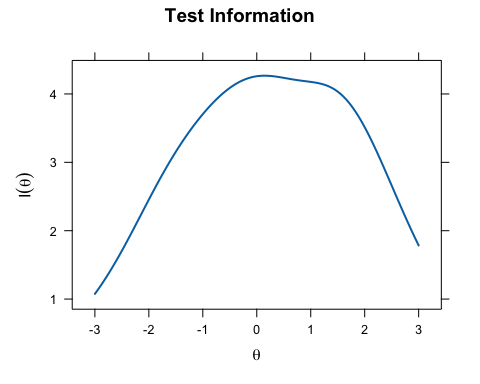
## Ejemplo: FII

plot(grm, type = "infotrace", theta\_lim = c(-3,3), lwd = 2)



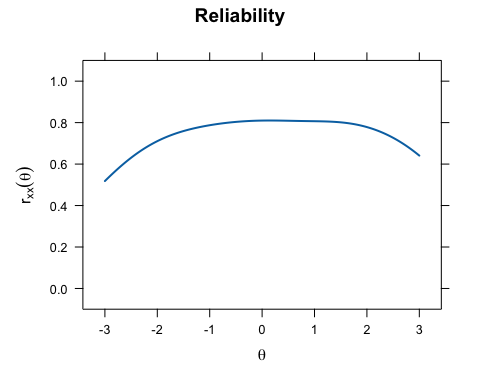
## Ejemplo: FIP

plot(grm, type = "info", theta\_lim = c(-3,3), lwd = 2)



## Ejemplo: Confiabilidad

plot(grm, type = "rxx", theta\_lim = c(-3,3), lwd = 2)



marginal\_rxx(grm)

[1] 0.7944443

## Ejemplo: Ajuste de ítem

itemfit(grm)

item S\_X2 df.S\_X2 RMSEA.S\_X2 p.S\_X2  
1 cint1 37.080 39 0.000 0.558  
2 cint2 51.364 44 0.014 0.207  
3 cint4 40.499 43 0.000 0.580  
4 cint11 59.168 43 0.021 0.051  
5 cint27 81.114 40 0.035 0.000  
6 cint28 41.579 43 0.000 0.533  
7 cint29 55.568 43 0.019 0.095  
8 cint30 39.844 46 0.000 0.727