

La imparcialidad y la detección de sesgo en Pruebas y Evaluación Educativa

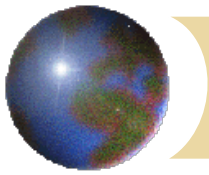
*Discurso de apertura del Seminario de
Evaluación Equitativa en Diversos Contextos*

Kurt F. Geisinger, Bueros Center for Testing

Universidad de Nebraska-Lincoln

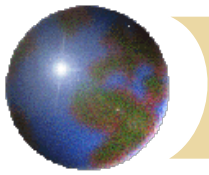
22 de Septiembre de 2016

BUROS
CENTER FOR TESTING



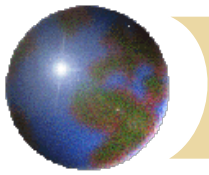
Introducción

- ✚ ¿Por qué se desea la imparcialidad?
- ✚ Hay demasiadas razones sociales para buscar imparcialidad en la evaluación
- ✚ Hay demasiadas razones personales para buscar imparcialidad en la evaluación
- ✚ Debemos hacer lo que esté a nuestro alcance
- ✚ Mi propia experiencia
- ✚ Motivaciones para trabajar hacia la imparcialidad



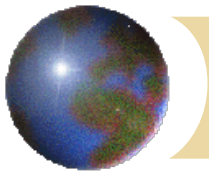
Imparcialidad: Problemas con el Constructo

- ✚ A veces vistos como opuestos
- ✚ Imparcialidad: ¿Cómo la prueba se administra y como se utilizan los resultados?
- ✚ Sesgo: Una características de los puntajes o resultados de la prueba
- ✚ El sesgo no solo son las diferencias en los puntajes promedio o resultados de las evaluaciones, pero podría indicar una necesidad de ver más allá...



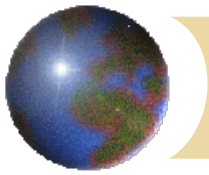
Definiciones de Psicología Organizacional

- ❖ La imparcialidad es un constructo social que se asocia al reconocimiento de los recursos escasos
- ❖ La equidad considera los resultados de la contratación o de las decisiones de admisión posteriores a que las decisiones iniciales sean hechas para validar la conveniencia de esas decisiones a través de los grupos



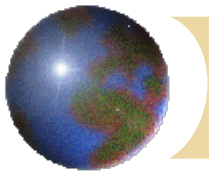
Sesgo Predictivo

- Los primeros intentos de estudios de sesgo (impacto de los puntajes)
- Lo siguió el Acta de los Derechos Civiles de EEUU de 1964
- Las pruebas de admisión para educación superior y medidas de empleo
- Los casos legales que han moldeado esta discusión en EEUU
- Los investigadores pioneros sabían que algunos grupos minoritarios tuvieron resultados más bajos en dichas evaluaciones, pero consideraban que tales puntajes bajos predecirían resultados más altos que en los grupos mayoritarios.



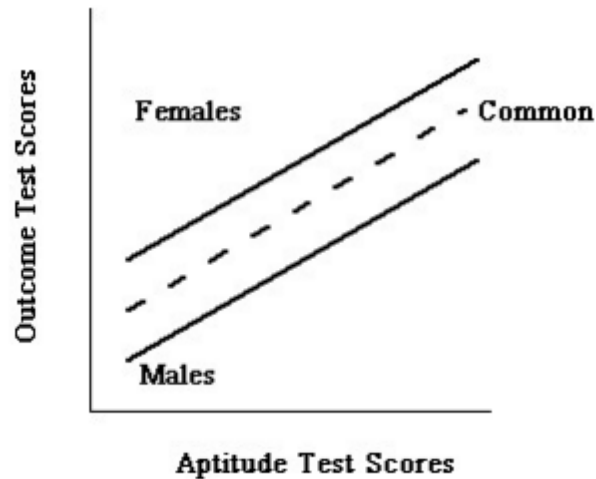
Situaciones Predictivas: Educación Superior

- ✚ Líneas de Regresión de dos grupos estudiados.
- ✚ Las líneas pueden diferir en magnitud (errores), pendiente o intercepto (Modelo de Cleary, 1968)
- ✚ Los investigadores pioneros esperaban en mayor medida sesgo en el intercepto

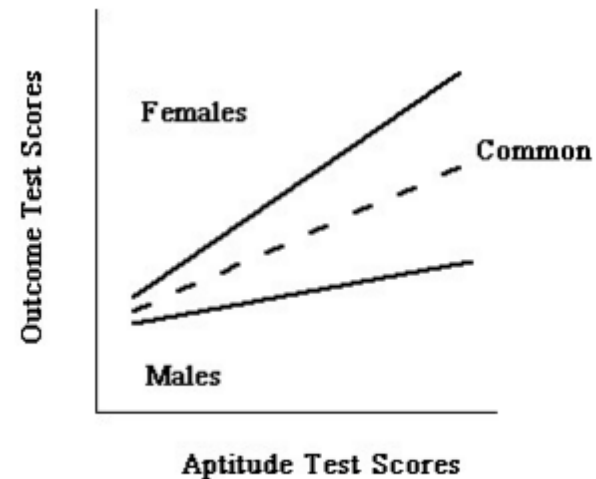


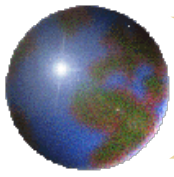
Tipos de Sesgo Predictivo

**Sesgo en el Intercepto-
semejante pero validez
diferente**

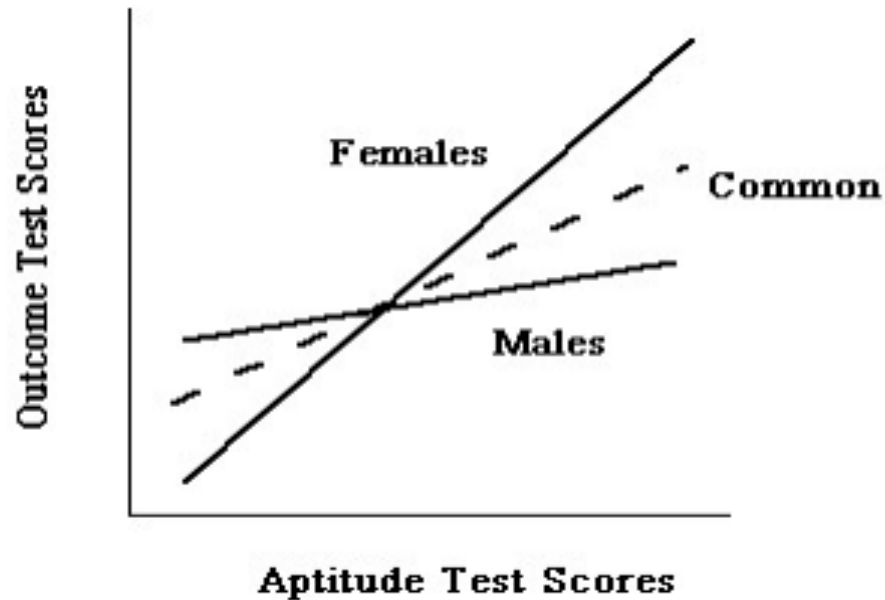


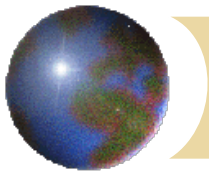
**Sesgo en la pendiente-
validez diferencial**





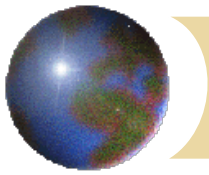
Un Segundo Caso de Sesgo de la Pendiente





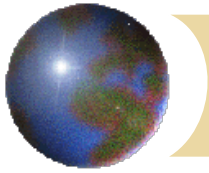
Una Aproximación por Regresión Múltiple a la Detección del Sesgo

1. Criterio de predicción con las pruebas de admisión
2. Incorporar 2^{do} y 3^{er} predictor: grupo de membresía y grupo x interacción del predictor
3. ¿La segunda revisión se agrega a la varianza predecida?
4. Si la segunda revisión no agrega significativamente, la prueba es imparcial para estos grupos



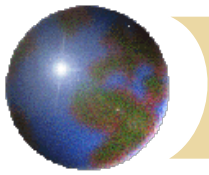
Una Aproximación por Regresión Múltiple a la Detección del Sesgo

5. Si la segunda revisión agrega significativamente, la prueba está sesgada para estos grupos
 6. Si está sesgada, retirar el término de interacción; si eso disminuye la varianza explicada para la significancia, existe sesgo en la pendiente
 7. Si no hay una gran diferencia, entonces existe sesgo en el intercepto
- Se puede utilizar modelos de ecuaciones estructurales si hay diferencias muy grandes en la confiabilidad



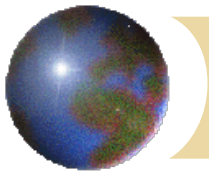
Sesgo Predictivo (Final)

- ✚ No ha sido comúnmente encontrado en las investigaciones de evaluación
- ✚ Cuando llega a ocurrir, el sesgo en el intercepto es mucho más común que el sesgo en la pendiente



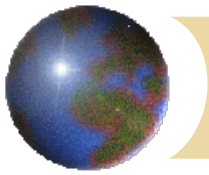
Paneles de Revisión de la Imparcialidad

- ✚ Panel de expertos para revisar los ítems de la prueba
- ✚ Los miembros inspeccionan los items de la prueba (preguntas de la prueba) y fomentan la identificación de materiales que los revisores consideren que puedan resultar en mediciones menos válidas para algunos grupos demográficos de evaluados –material inapropiado para un grupo pero no para otro
- ✚ Realizados inicialmente antes del uso de la prueba o inclusive en pre-prueba



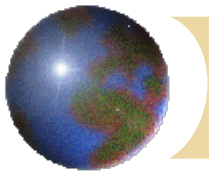
Propósito de las Revisiones Imparciales

- ✚ Búsqueda de cualquier tipo de variables que pudieran influencia inapropiadamente los resultados de la evaluación
- ✚ Se espera que el uso de estas revisiones elimine o minimice las fuentes de varianza de las puntuaciones no atribuida al contenido o constructo que está siendo evaluado
- ✚ Términos ofensivos o comentarios hacia un grupo
- ✚ Lenguaje vergonzoso si es hecho público
- ✚ Estereotipamiento



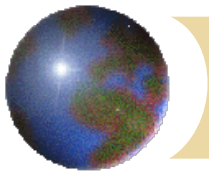
Naturaleza de los Grupos Imparciales

- ✚ Pueden incluir miembros de más de un grupo "minoritario"
- ✚ Incluye personas conocedoras acerca del contenido de la prueba
- ✚ Búsqueda de ítems problemáticos
 - ✚ Pueden ser eliminados
 - ✚ Pueden ser cambiados y re-revisados



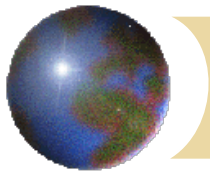
Reglas Generales para Revisiones Imparciales

1. Evitar fuentes cognoscitivas de varianza irrelevante del constructo.
 2. Evitar fuentes afectivas de varianza irrelevante del constructo
 3. Evitar fuentes físicas de varianza del constructo.
- ✚ Los paneles de imparcialidad necesitan considerar si el conocimiento es visto como desigual a través de los grupos
 - ✚ Revisiones lógicas de los ítems



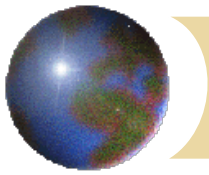
Revisiones Imparciales

- ✚ Objetivo: “Identificar factores irrelevantes del constructo”
 - ✚ Carga verbal
 - ✚ Vocabulario
- ✚ Necesidad para seguir el protocolo bien descrito
 - ✚ ¿Es esto sólo exactitud política?
 - ✚ ¿Por qué el lenguaje no discriminatorio es tan importante?
 - ✚ ¿Cómo tales procesos impactan las habilidades y contenidos de las pruebas?



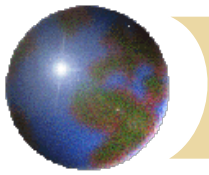
Resultados de las Revisiones Imparciales

- ✚ Ha sido bien conocido que la relación entre los juicios de los revisores de la imparcialidad del ítem y los hallazgos empíricos del sesgo del ítem, es en el mayor de los casos, “débil” (Zieky, 2013, p. 301).
- ✚ Evitar el estereotipamiento y comentarios vergonzosos (que pudieran terminar en la prensa)



Funcionamiento Diferencial del Ítem (dif)

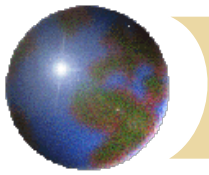
- ✚ Revisión ítem por ítem del desempeño
- ✚ Ajustados para puntuaciones de criterio (usualmente en puntajes promedio de las pruebas-ajustados posteriormente)
- ✚ Usualmente hechos durante la pre-evaluación y puntaje operacional
- ✚ Los resultados pueden ser enviados al Panel de Revisión de Imparcialidad
- ✚ Pero fácilmente salir con razones de diferencias *ex post facto*



Servicio Educacional de Pruebas

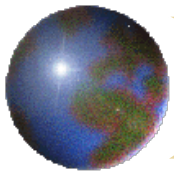
Graduaciones dif

- ❖ La categoría A contiene las preguntas con una poca o nula diferencia entre los grupos igualados.
- ❖ La categoría B contiene preguntas con una pequeña o moderada diferencia.
- ❖ La categoría C contiene preguntas con una gran diferencia.

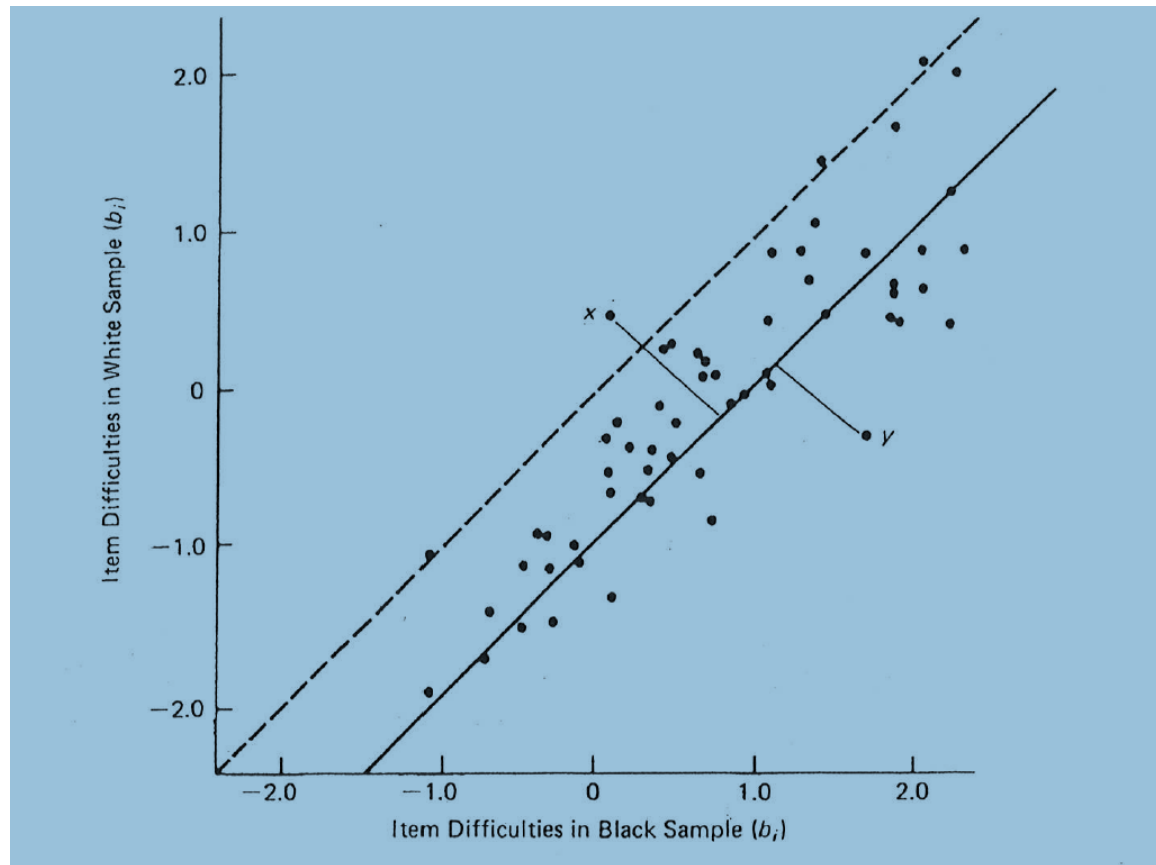


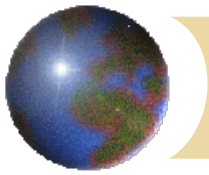
El uso del DIF

- ✚ Muy prevalente a gran escala, evaluación con grandes expectativas
- ✚ Mucho menos prevalente en evaluación organizacional dónde la creencia es más hacia que las pruebas son generalmente imparciales



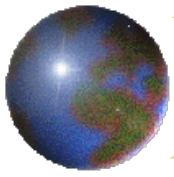
Aproximación Inicial DIF: Angoff



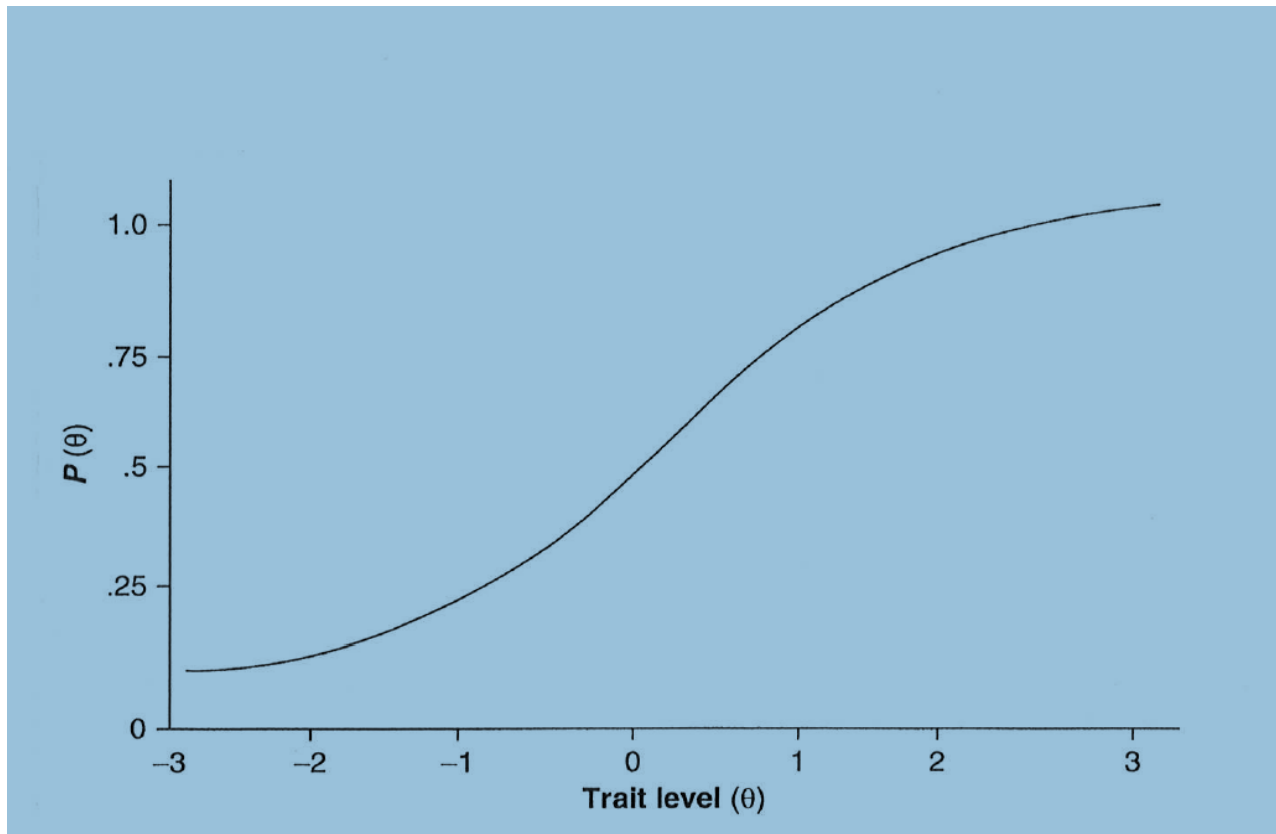


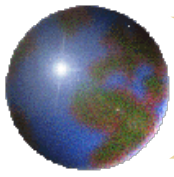
Procedimiento Mantel-Haenszel

- ✚ Procedimiento chi-cuadrada
- ✚ Divide tanto la referencia como los grupos focales en rangos de puntuaciones (4-5)
- ✚ Compara los momios de responder corretamente la pregunta para los individuos en cada grupo a traves de 4-5 bandas.
- ✚ Procedimiento muy comúnmente empleado ya que está ampliamente disponible.
- ✚ Requiere 500 en referencia y 200 en focal

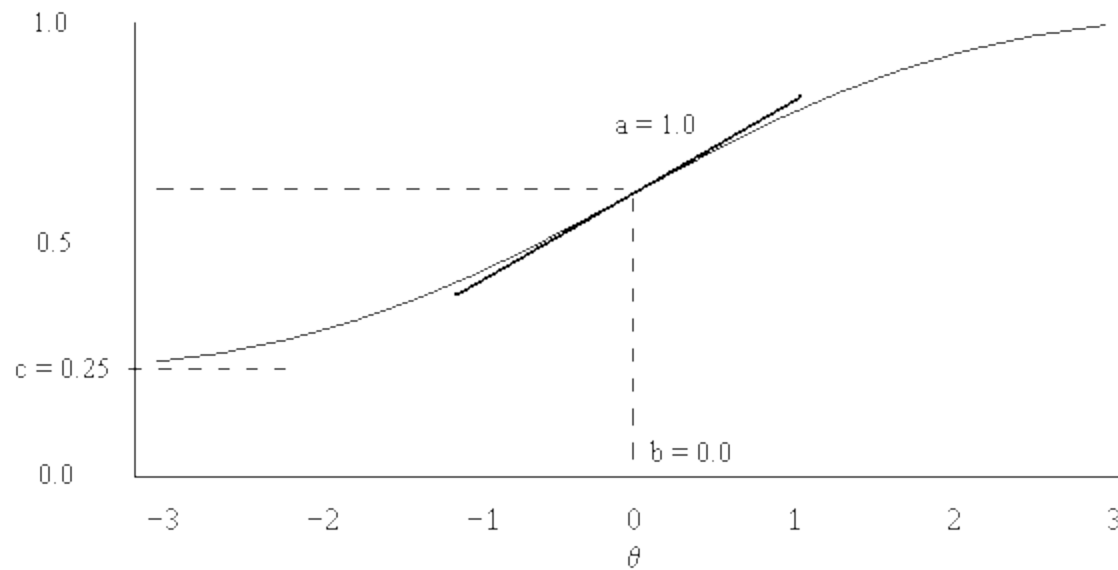


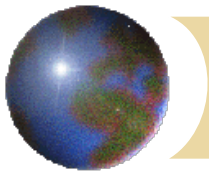
El Model de Teoría de Respuesta al Ítem





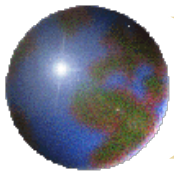
Los Tres Parámetros





Los Tres Parámetros Comunes (TRI)

- ✚ B— dificultad del ítem (utilizado en todos los modelos de TRI y es el único parámetro en Rasch)
- ✚ A—discriminación del ítem (pendiente de la línea)
- ✚ C—Probabilidad de acertar sin tener el conocimiento también llamado parámetro de adivinación—cuando la línea cruza el eje Y, empleado en el modelo de 3 parámetros.



Modelo TRI con DIF Uniforme

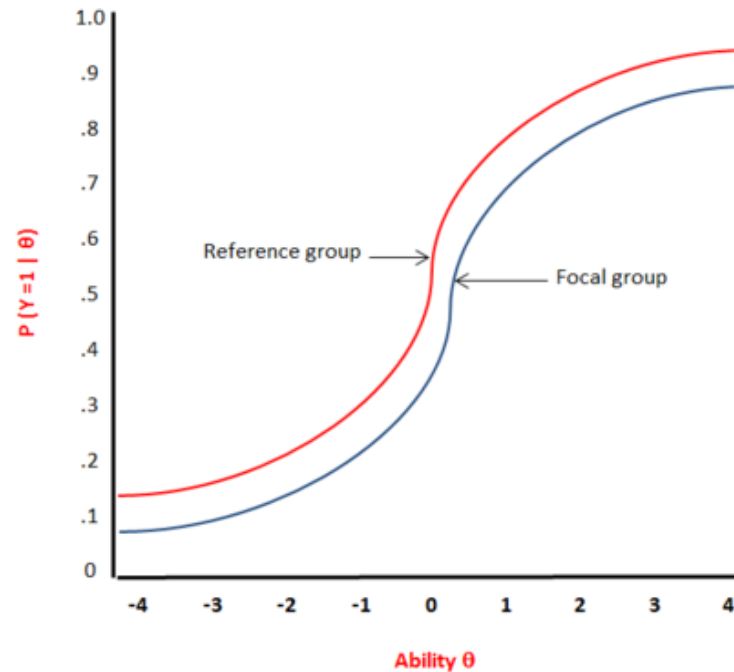
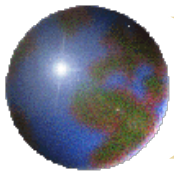


Figure 1. ICC for Uniform DIF



Modelo TRI con DIF No-Uniforme

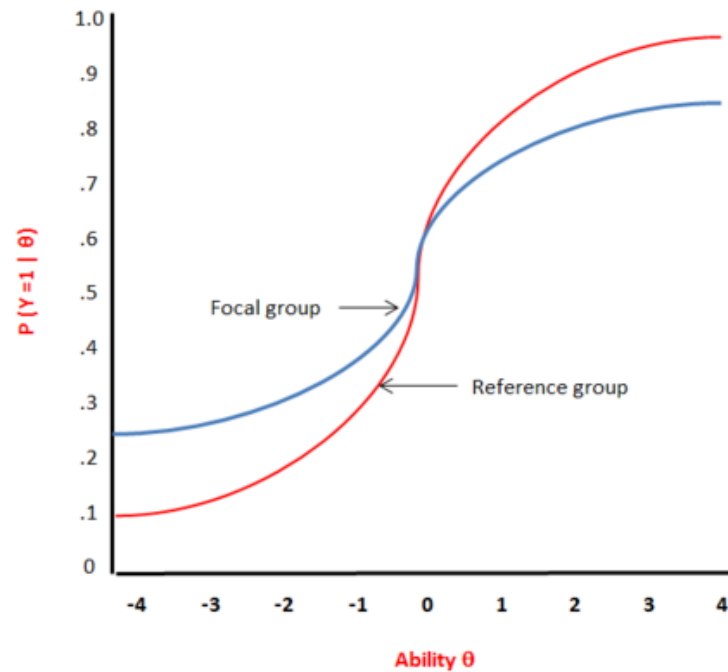
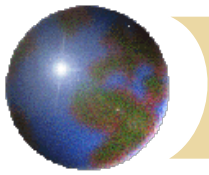
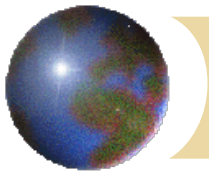


Figure 2. ICC for Nonuniform DIF



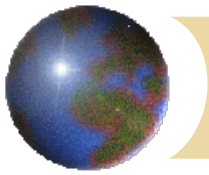
Reseña del DIF

- ✚ Un componente necesario en la construcción de pruebas
- ✚ ¿Cuántos grupos para hacerlo?
- ✚ A veces es difícil obtener cifras mínimas, especialmente en las pruebas preliminares
- ✚ ¿Cuál es el criterio correcto (p.e., la puntuación total?) - Puede tener un impacto enorme!



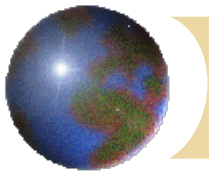
Sesgo en la Validez de Constructo

- ✚ No es muy común en las pruebas educativas
- ✚ ¿Se evaluó el constructo subyacente por una medida igual para ambos grupos?
- ✚ Algunos constructos son diferentes en diferentes grupos
- ✚ Algunas medidas adaptadas pueden no medir el constructo de manera similar para ambos grupos
- ✚ La investigación es necesaria para comparar los resultados



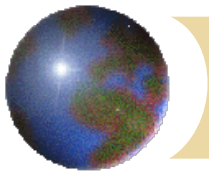
Imparcialidad a través de los Idiomas y Culturas

- ✚ Las sociedades son cada vez más diversas
- ✚ Muchas culturas presentes e idiomas hablados
- ✚ La conversión es más que la traducción
- ✚ Deben adaptarse los ítems para idiomas y culturas – conocido como adaptación de pruebas
- ✚ Necesidad de equipos de expertos para lograrlo
- ✚ Posible construcción simultánea



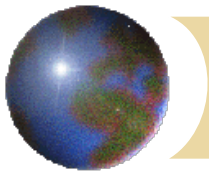
Pasos en la Adaptación de Pruebas

- ✚ Un equipo independiente de expertos ha de evaluar si el idioma, contenido, formato y otros aspectos de los dos exámenes parecen comparables
- ✚ Cambios realizados
- ✚ Pruebas de campo-probar: validez y confiabilidad
- ✚ Estudio que lo vincule



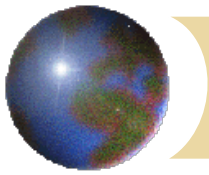
Evaluando Estudiantes con Discapacidades

- ✚ Área en boga de investigación debido a las leyes de EEUU y convenciones de la ONU
- ✚ Comenzado con las pruebas de admisión
- ✚ Adaptaciones necesarias para permitir evaluar a las personas con discapacidades
- ✚ Intentar mantener pruebas válidas
- ✚ Investigación de validación indica la validez sustancial en el uso de medidas adaptadas



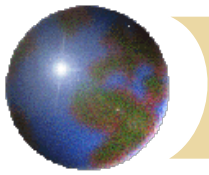
Adaptaciones

- ⊕ Retirar varianza irrelevante del constructo
- ⊕ Proporcionar acceso a los materiales de prueba que de otro modo se puede perder
- ⊕ 3 Atributos
 - *Presentación alterada o modo de respuesta para proporcionar acceso
 - *No cambia el constructo evaluado
 - *"El acceso diferencial": la puntuación aumenta para aquellos que reciben la adaptación
 - *Phillips preguntó cuan normal se beneficiarían los examinados si recibieron la adaptación
 - *La confiabilidad y la validez de la determinación de la adaptación



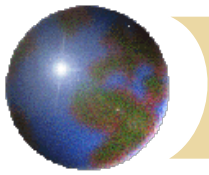
Administración/Presentación de la Prueba

- ✚ De tipo grande, lectura en voz alta, computadoras, calculadoras, problemas de ritmo, idioma
- ✚ ¿Cómo lee normalmente el estudiante?
 - *¿De qué manera? (p.e., a gran tipo)
 - *¿Cuán efectivo? (a nivel de grado?)
- ✚ Puede el estudiante seguir instrucciones?
- ✚ ¿El estudiante normalmente utilizan computadora?
- ✚ ¿Cómo evaluarlo en el aula?



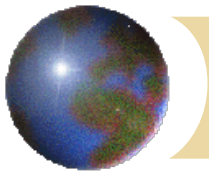
Modificaciones

- ⊕ Se cambia el constructo que se está evaluando
- ⊕ Implica un cambio significativo en la prueba
 - *A partir de la lectura para leer algo que tiene a la persona
 - *A partir de una prueba altamente acelerada a tiempo prolongado
- ⊕ No interpretar lo mismo como “puntuaciones normales”
- ⊕ La única alternativa puede ser no evaluarlo



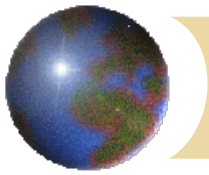
Constructos sin Cambios

- ✚ La adaptación elimina factores de confusión, dejando al constructo que deben evaluarse
- ✚ Ejemplo: Eliminación de la lectura en las pruebas de matemáticas
- ✚ La eliminación de la velocidad de las pruebas de aptitud
- ✚ ¿Qué piensan sobre estos cambios?



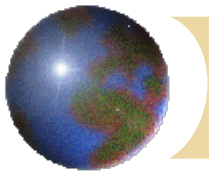
Algunas Conclusiones

- ✚ Podemos evaluar las pruebas y sustituir las formas adaptadas para hacerlas pruebas más válidas para todos.
- ✚ Tenemos que estar especialmente preocupados por los problemas de idioma
- ✚ Tenemos que avanzar a una mejor adaptación y construir pruebas con diseño Universal



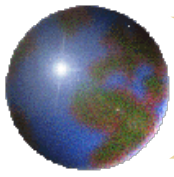
Conclusiones Generales

- ✚ Algunas diferencias entre los grupos
diferencias son exactamente eso:
diferencias reales
- ✚ Muchas diferencias pueden ser
adaptadas
- ✚ Las preguntas clave para los
investigadores, determinarán si las
diferencias son reales y cuales pueden
ser adaptadas



Gracias

- ❖ Después de todo, nuestro objetivo debe ser el éxito para todos los estudiantes!
- ❖ Construyamos puentes para su éxito, no barreras para evitar que lo tengan.



Resumen

