



El papel de la evaluación en la promoción de la justicia social

3rd International Virtual Meeting:
Teaching, Learning & Assessment In Higher Education

Dylan Wiliam (@dylanwiliam)

www.dylanwiliam.org

Resumen

2

- ¿Que es la evaluación?
- ¿Para que es la evaluación?

Un marco para pensar en el cambio

No hay soluciones perfectas, compensaciones

4

- Propuestas usuales para el cambio
 - claro sobre lo que hizo mal el enfoque anterior y cómo la propuesta mejorará esos aspectos
 - guardar silencio sobre las cosas que el antiguo enfoque hizo bien y que la nueva propuesta hará menos bien
- Las propuestas de cambio deben responder 2 preguntas:
 - “¿Qué será mejor si se hacen los cambios?”
 - “¿Qué será peor si se hacen los cambios?”
- Si la respuesta a la segunda pregunta es "nada", entonces el proponente debe pensar de nuevo.
- Siempre habrán compensaciones
 - La pregunta es si son explícitos o no

Evaluación

Evaluación

6

- La evaluación es un procedimiento para sacar inferencias
 - Damos a los alumnos cosas que hacer
 - Identificamos la evidencia
 - Sacamos conclusiones

Evolución de la idea de validación

- Una propiedad de ...
 - un test
 - puntajes de los estudiantes en una prueba
 - inferencias extraídas sobre la base de los resultados de las pruebas
- “Uno valida no una prueba sino una interpretación de los datos que surgen de un procedimiento específico” (Cronbach, 1971)

Permutaciones preposicionales

8

- Evaluación del aprendizaje
 - Status
 - Competencia
 - Ranking
 - Predicción
 - Accountability
- Evaluación para el aprendizaje
 - Mejorar el aprendizaje
- Evaluación como aprendizaje
 - El rol del aprendiz

Evaluación por (y como) aprendizaje

Evaluación para el aprendizaje (Mittler, 1973)

10

- La evaluación para el aprendizaje es un concepto más amplio que la evaluación formativa
 - Evaluación para la motivación
 - Evaluación para la práctica de recuperación
 - Eficaz incluso cuando las pruebas no están marcadas
 - Efecto de hipercorrección

¿Cómo puede la evaluación mejorar el aprendizaje ?

11

¿Anunciado?	¿Entregado?	¿Anotado?	¿Usado?	
				Evaluación para motivación
				Práctica de recuperación
				Correctivos instruccionales
				Evaluación formativa

Desempaquetando la Evaluación Formativa

12

	Hacia donde va el aprendiz	Donde el aprendiz esta ahora	Como llevar al aprendiz ahí
Docente	Aclarar, compartir y comprender las intenciones de aprendizaje	Obteniendo evidencia	Proporcionar comentarios que hagan avanzar a los alumnos
Par		Activar a los estudiantes como aprendizajerecursos unos para otros	
Estudiante		Activar a los estudiantes como dueños de su propio aprendizaje	

Desempaquetando la Evaluación Formativa

13

	Hacia donde va el aprendiz	Donde el aprendiz esta ahora	Como llevar al aprendiz ahí
Docente	Aclarar, compartir y comprender las intenciones de aprendizaje	Obteniendo evidencia	Proporcionar comentarios que hagan avanzar a los alumnos
Par		Activar a los estudiantes como recursos unos para otros	
Estudiante		Activar a los estudiantes como dueños de su propio aprendizaje	

Intenciones de aprendizaje y criterios de éxito

14

- Intenciones de aprendizaje
 - descripciones del *aprendizaje previsto*
 - útil para la *planificación*
 - principalmente útil para los profesores
- Criterio de éxito
 - descripciones del desempeño en tareas de aprendizaje
 - principalmente útil para *evaluar*
 - útil tanto para profesores como para estudiantes

Obteniendo evidencia

15

- Dos buenas razones para hacer una pregunta
 - Porque pensando
 - Proporcionar datos que informen la instrucción.
- Mejor evidencia
 - Más profunda
 - Más amplia
- *Creando, capitalizando momentos de contingencia*

Feedback

16

- El propósito de la retroalimentación es mejorar al estudiante, no el trabajo.
- Lo único que importa con la retroalimentación es lo que los estudiantes hacen con ella.
- Si tu retroalimentación te está dando más de lo que quieres, es una buena retroalimentación.
- La retroalimentación debería ser más trabajo para el receptor que para el donante.

Aprendizaje Cooperativo y Colaborativo

17

- Asunto clave: ¿Quién decide el objetivo?
 - Estudiantes: “aprendizaje colaborativo”
 - Docentes: “aprendizaje cooperativo”
- El propósito de la colaboración y cooperación
 - Para producir la mejor solución
 - Para mejorar la capacidad de las personas para colaborar
 - Maximizar el aprendizaje del grupo.
 - Maximizar el aprendizaje de cada miembro del grupo.

Los estudiantes como recursos unos para otros

18

- Metas grupales:
 - para que los estudiantes trabajen como un grupo, no solo en un grupo
- Accountability individual:
 - los mejores esfuerzos de aprendizaje de cada miembro del grupo deben ser necesarios para que el grupo tenga éxito, y
 - el desempeño de cada miembro del grupo debe ser claramente visible y cuantificable para los otros miembros del grupo

Los estudiantes como dueños de su propio aprendizaje

19

- Estudiantes evaluando su propio trabajo:
 - Con rúbricas
 - Con ejemplares
- Auto-evaluación del entendimiento:
 - Portafolio de aprendizaje
 - más/menos/interesante
 - Pruebas de práctica

Evaluación del aprendizaje

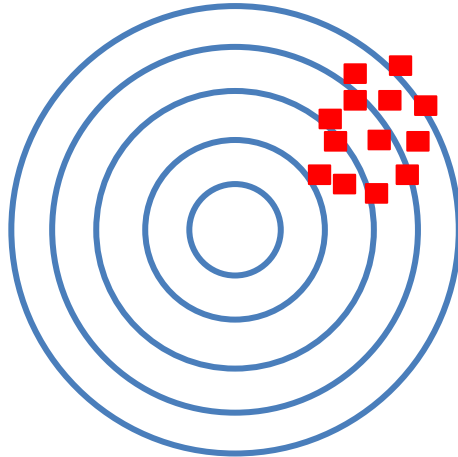
Fiabilidad

21

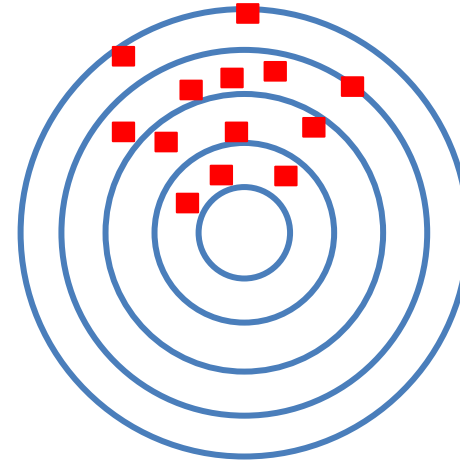
- ¿Validez y fiabilidad?

Una común – pero limitada – metáfora

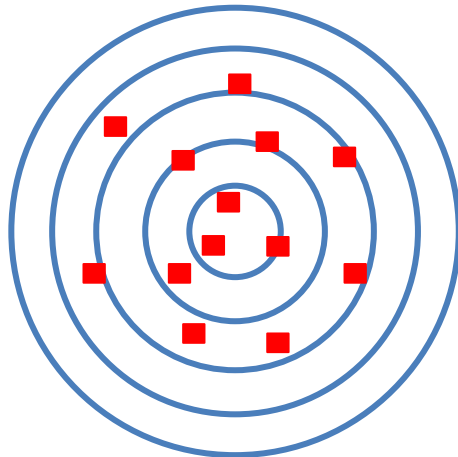
22



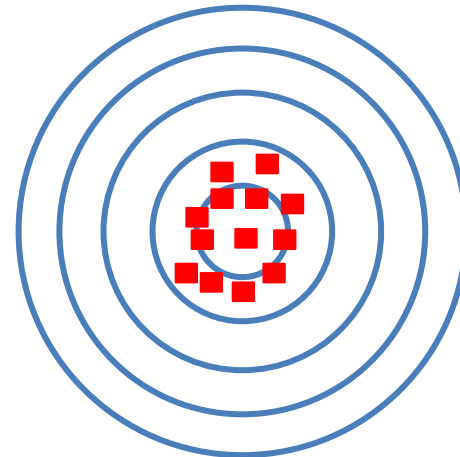
Fiable pero no valido



Ni fiable ni valido



Valido pero no fiable



Tanto valido como fiable

Fiabilidad

23

- ¿Validez y fiabilidad?
- ¿Validez versus fiabilidad?

Resolviendo la paradoja validez-confiabilidad

- La paradoja en breve
 - La fiabilidad es un pre-requisito previo para la validez
 - Los puntajes no confiables no pueden respaldar inferencias válidas
 - Fiabilidad como enemigo de la validez
 - Los intentos de aumentar la confiabilidad generalmente reducen la validez
- La resolución
 - Aumentar la fiabilidad
 - Fortalece ciertas inferencias
 - Debilita otras inferencias

La metáfora de la “iluminación escénica”: Floodlight



La metáfora de la “iluminación escénica”: Spotlight



Fiabilidad

27

- ¿Validez y fiabilidad?
- ¿Validez versus confiabilidad?
- Validez *incluyendo* la fiabilidad

Dos amenazas principales para las interpretaciones válidas

28

La evaluación es "demasiado grande"

- Las puntuaciones dependen de cosas que no deberían
 - Factores irrelevantes
 - Escritura
 - Suerte
 - Buenos/malos días
 - Quien califica los trabajos
 - Las tareas particulares establecidas
- Varianza irrelevante del constructo

La evaluación es "demasiado pequeña"

- Las puntuaciones no dependen de las cosas que deberían
 - Evaluar solo cosas que son fáciles de evaluar
- Sub-representación del constructo.

- Implicancias
 - No hay tal cosa como una evaluación válida
 - La validez es una propiedad de las inferencias
 - La fiabilidad es parte de la validez
 - Específicamente, el componente aleatorio de la varianza irrelevante para el constructo
 - No hay tal cosa como una “evaluación sesgada”
 - Una prueba prueba lo que una prueba prueba
 - No existe como tal la evaluación formativa
 - los términos formativa y sumativa son propiedades de inferencias, no evaluaciones

Constructos y Equidad: Ejemplos

Evaluación de la historia

- En las pruebas de opción múltiple de la historia, los hombres superan a las mujeres
- En las pruebas de respuesta construida de la historia, las mujeres superan a los hombres
- Interpretaciones comunes
 - a. Las pruebas de opción múltiple están sesgadas contra las mujeres
 - b. Las pruebas de respuesta construida están sesgadas contra los hombres

Definición de constructo

- “Saber hechos y fechas”
 - Las pruebas de opción múltiple son ideales porque es posible evaluar muchos hechos y fechas.
 - Las pruebas de respuesta construida evalúan cosas que no deberían, como habilidades lingüísticas, escritura a mano.
 - Las pruebas de respuesta construida hacen menos preguntas, lo que aumenta el papel de la suerte.
- “Describir y explicar eventos históricos”
 - Las pruebas de opción múltiple son inadecuadas (las puntuaciones no dependen de las cosas que deberían).
 - Las pruebas de respuesta construida son esenciales (por lo que los puntajes dependen de las cosas que deberían)

Evaluación de desempeño

- Requisitos clave
 - La evaluación es directa y no por correlatos
 - demostración del desempeño de la tarea
- Ejemplos
 - Test de conducción
 - Examen Clínico Estructurado Objetivo
- Complementos adicionales
 - Auténtico
 - significativo para el estudiante
 - Atractivo
 - aplicado

Compensaciones en la evaluación del desempeño

- Pros

- Evaluar las capacidades que no se pueden evaluar de otras maneras
- Atractivo y motivador para los estudiantes
- Validez aparente

- Contras

- La definición de construcción es desafiante
- Comparabilidad
- Divulgación
- Generalizabilidad
 - Puntuación
 - Interacción alumno x tarea

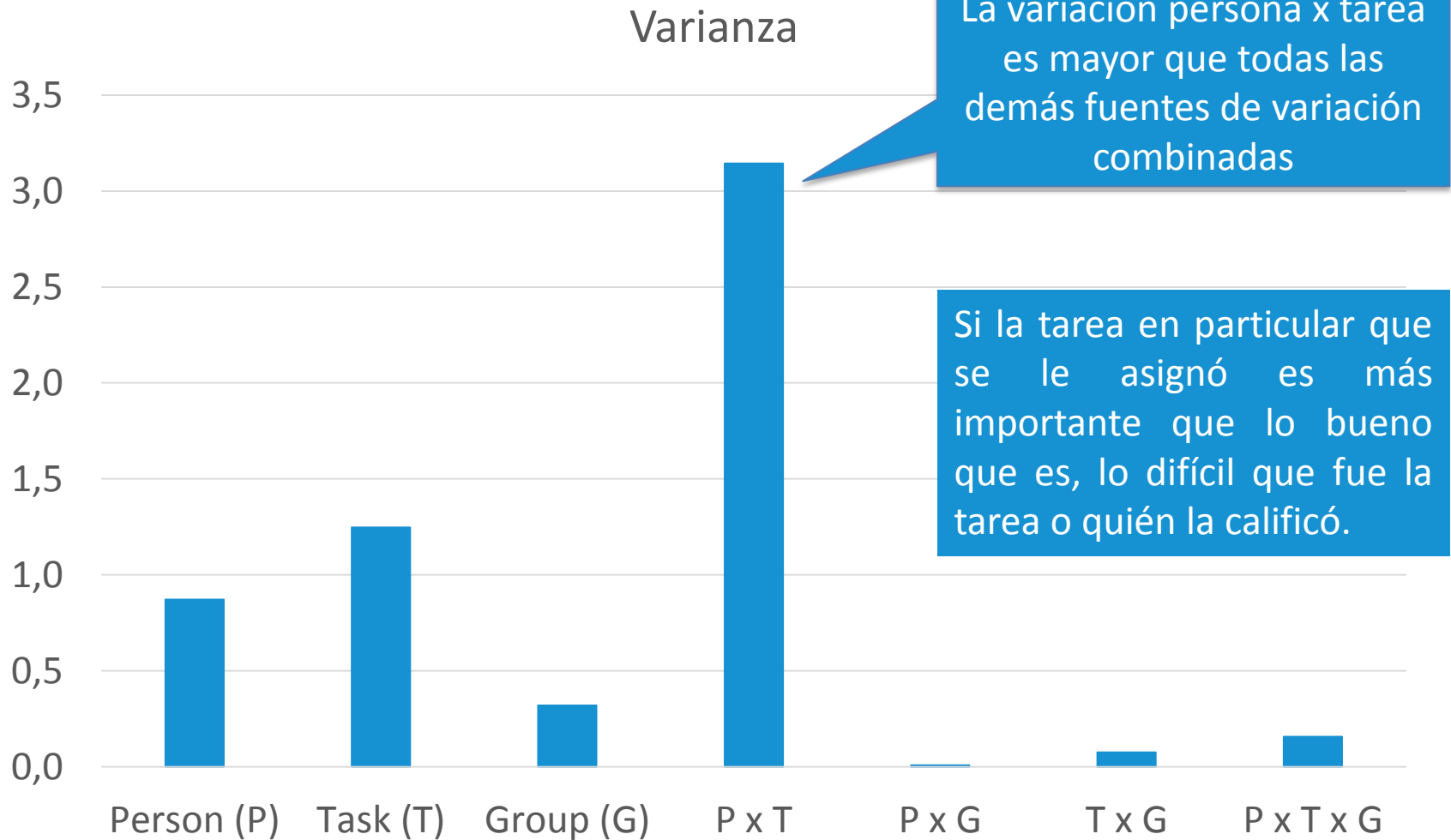
Estudio de Caso: Evaluación de las habilidades de manejo de pacientes

35

- 200 estudiantes de medicina seleccionados al azar de un grupo más grande que había completado una evaluación de desempeño de habilidades de manejo de pacientes en línea
- Tres grupos de cuatro médicos de atención primaria de la facultad universitaria evaluaron 100 de las respuestas de los estudiantes
- Resultados utilizados para crear tres algoritmos de puntuación para cada tarea
- Algoritmos utilizados para puntuar las respuestas de los otros 100 alumnos

Resultados

36



Compensaciones en el diseño de la evaluación

- Distribuido
 - De modo que la recopilación de pruebas no se lleva a cabo en su totalidad al final
- Sinóptico
 - Para que el aprendizaje se tenga que acumular
- Extensivo
 - Para que todos los aspectos importantes estén cubiertos (amplitud y profundidad)
- Manejable
 - Para que los costos sean proporcionales a los beneficios.
- Confiado
 - Para que las partes interesadas tengan fe en los resultados

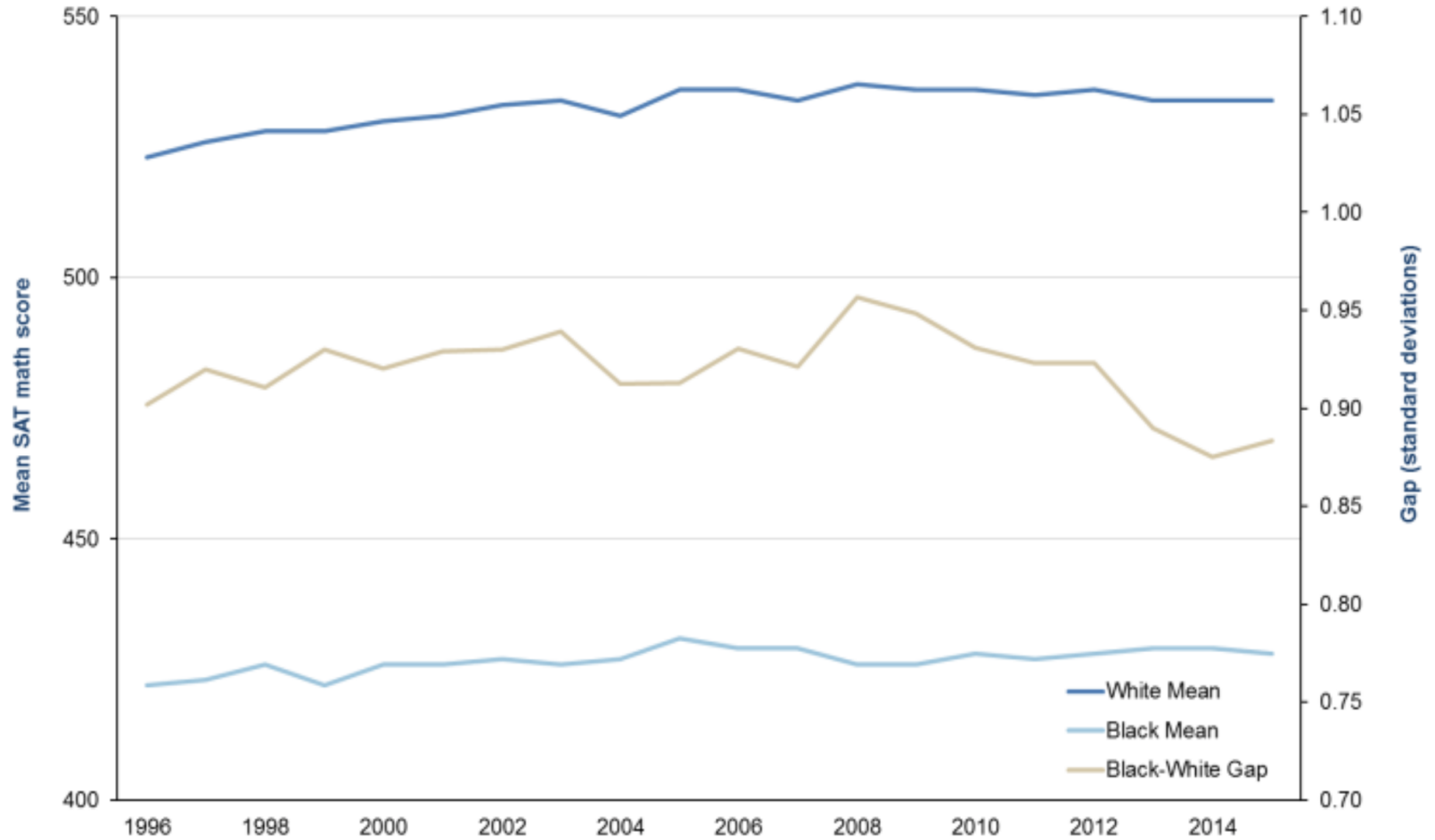
Selección para Educación Superior

Selección para la Educación Superior: El SAT

- “La prueba más investigada del mundo.” (College Board, 2009)
- Correlación
 - GPA de primer año: 0.44 to 0.62 (Hezlett et al., 2001)
 - GPA Acumulado: 0.35 to 0.45 (Hezlett et al., 2001)
 - GPA comparable al escolar (Kobrin, Camara, & Milewski, 2002)
- Todavía mejor al compararse con HSGPA
 - Estudiantes hispánicos 0.42
 - Estudiantes blancos 0.48
 - Estudiantes Asiático-americanos 0.55
 - Estudiantes Afro-americanos 0.55

¿Es justo el SAT?

Black-white SAT math achievement gap over time



Causas potenciales

- Diferencias muestrales
- Varianza irrelevante para el constructo
 - “Se ha demostrado que el SAT está sesgado cultural y estadísticamente en contra de los afroamericanos, los hispanoamericanos y los asiáticoamericanos.” (Freedle, 2003 p.1)
- Varianza relevante para el construct

Predicción del GPA del Primer año en 23 Universidades

	HSGPA		SAT		HSGPA+SAT	
	Mujer	Hombre	Mujer	Hombre	Mujer	Hombre
Afro-Americano	0.09	0.20	0.01	0.22	-0.01	0.16
Nativo Americano	0.13	0.24	0.06	0.32	0.07	0.28
Asiático-Americano	0.03	0.08	-0.01	0.15	-0.03	0.07
Hispánico	0.23	0.31	0.03	0.20	0.04	0.20
Blanco	-0.11	-0.03	-0.11	0.06	-0.09	0.05
Otro	-0.09	0.04	-0.13	0.03	-0.12	0.04

La relevancia de la construcción depende de las construcciones!

- ¿Cual es propósito SAT?
 - ¿Para predecir el rendimiento de los estudiantes en los programas universitarios existentes?
 - ¿Para predecir la capacidad de los estudiantes para prosperar en la educación superior?
 - ¿Para indicar qué estudiantes pueden necesitar apoyo adicional para prosperar?

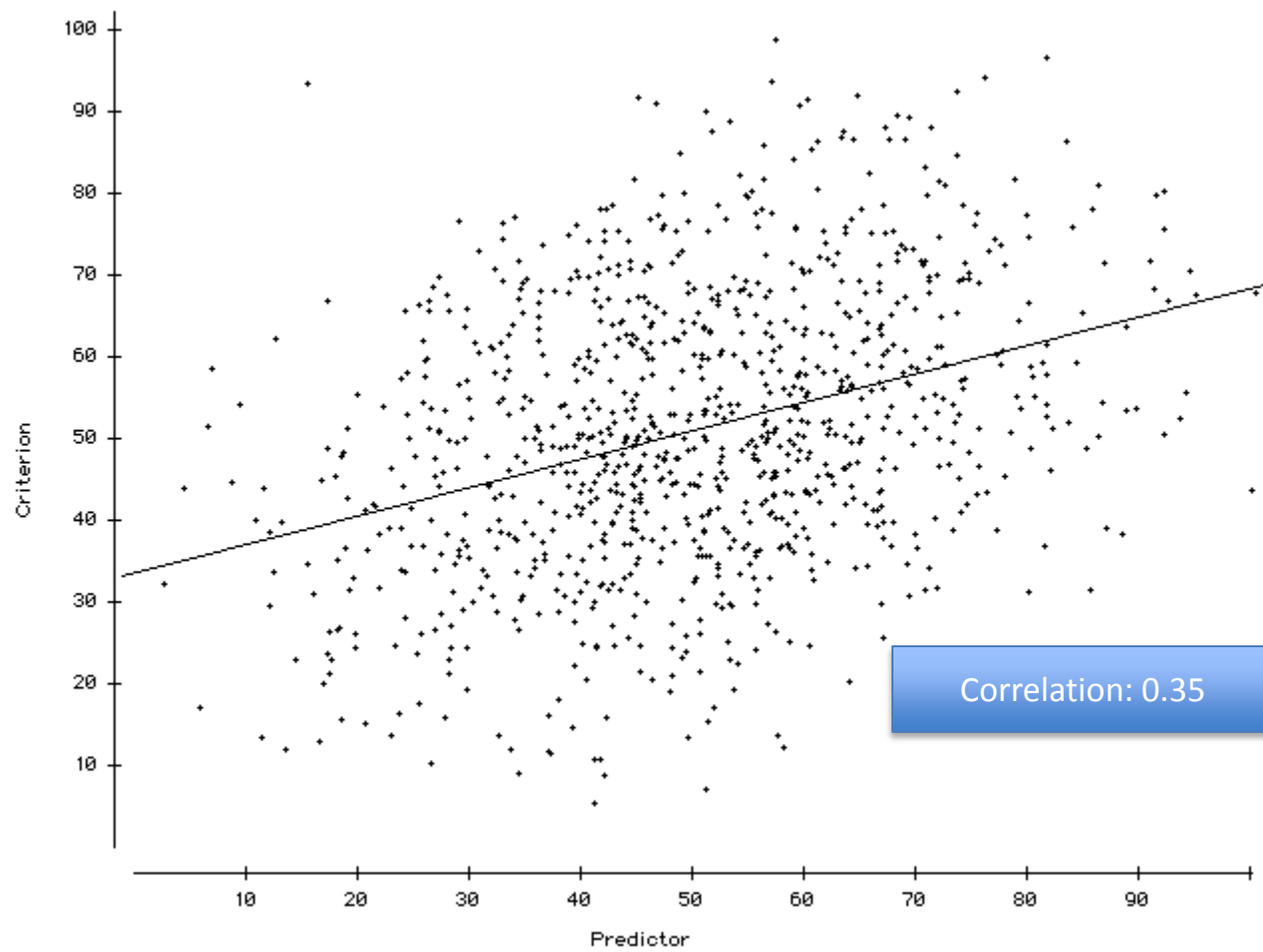
**Un estudio de caso
ilustrativo: Acceso a la
Medicina**

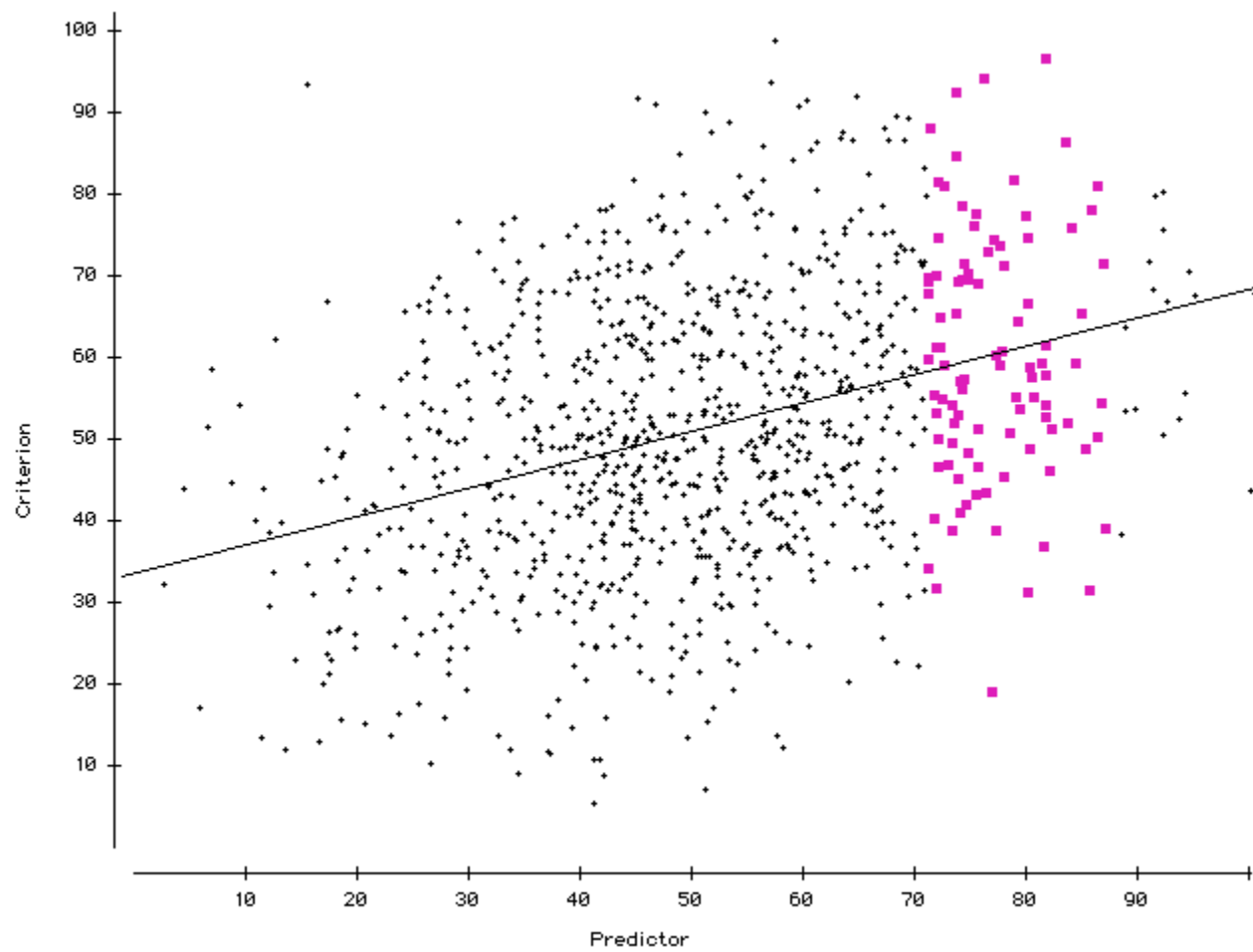
Acceso a la Medicina: El contexto

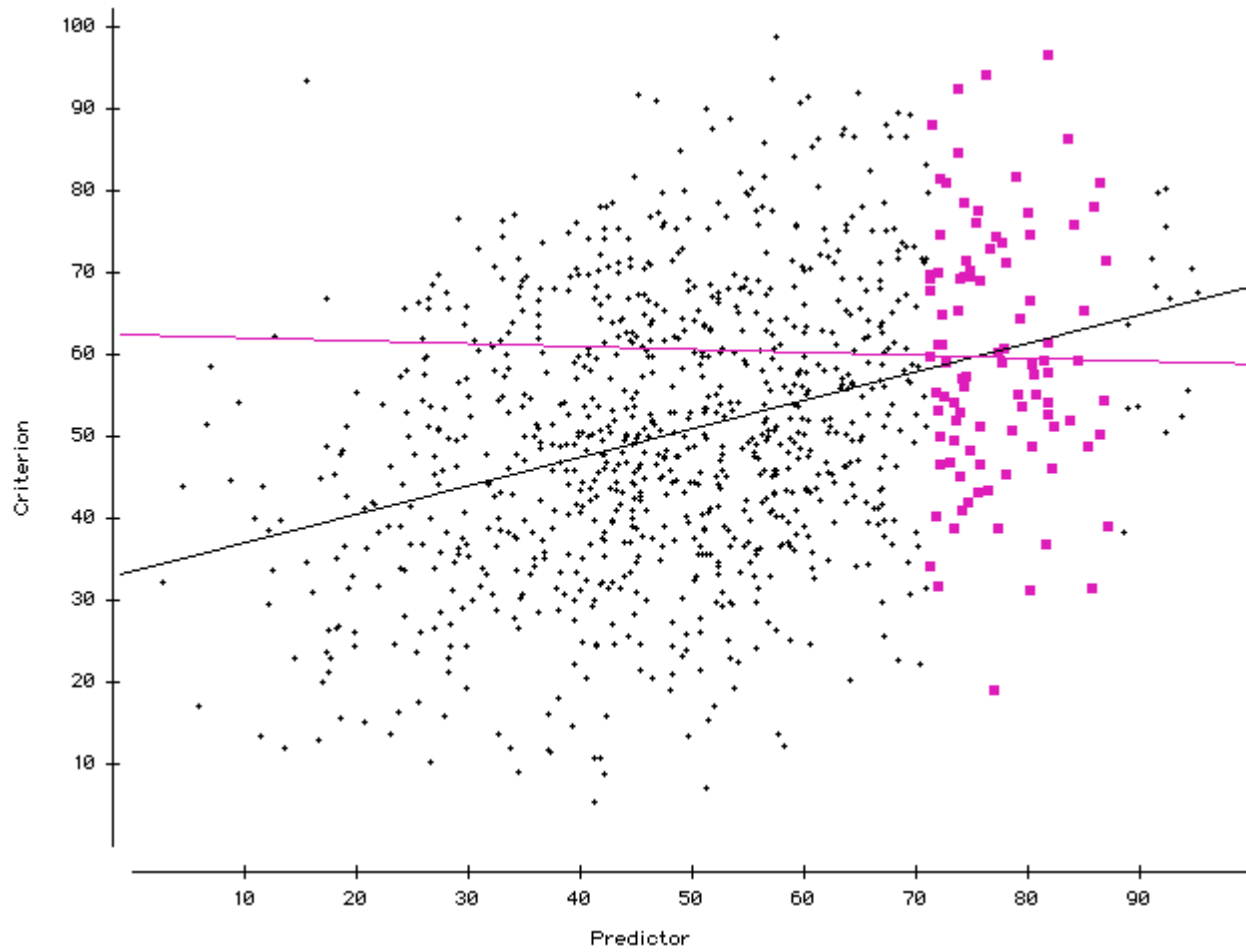
- Kings College Londres
 - Quinta universidad más grande del Reino Unido
 - La escuela de medicina más grande de Europa
 - Ubicado en el sureste de Londres: población étnicamente diversa
 - Admisiones altamente selectivas, basadas en logros en exámenes de alto riesgo
 - Resultado:
 - Estudiantes de escuelas privadas sobrerrepresentados
 - Estudiantes blancos y asiáticos sobrerrepresentados
 - Estudiantes de ascendencia africana sub-representados
 - Estudiantes en desventaja socioeconómica sub-representados

Acceso a la Medicina

- La meta:
 - Ampliar el acceso y el éxito en la educación médica
 - Producir servicios médicos más “culturalmente competentes”
- El desafío:
 - Hazlo sin “bajar el listón”
 - No brinde una "segunda oportunidad" a los estudiantes que ya tienen ventajas







Una solución en tres partes...

- Reclutamiento: Programa para elevar aspiraciones en secundaria
- Selección: solo los estudiantes de escuelas públicas en el área local son elegibles
 - Evaluación a través del rendimiento escolar, las características personales y las tareas de razonamiento científico
- Retención
 - Año básico adicional (dos años para cubrir el plan de estudios tradicional del primer año)
 - Soporte dedicado
 - gastos de manutención y matrícula proporcionados solo para el primer año
 - soporte tutorial dedicado

Tareas de razonamiento científico(SRTs)

- Basado en el trabajo de Shayer & Adey (1981).
- Conjunto de tareas administradas por grupos
- No se evalúa el conocimiento científico, sino la capacidad de incorporar nuevos hechos en el esquema existente.
- Comparado con los estudiantes de medicina existentes.
- Por lo tanto, proporcione formas alternativas de mostrar el talento.

La historia hasta ahora

- Programa en su año 16 (ahora "Programa de título médico extendido")
- "Estado estacionario: 300 estudiantes matriculados (50 en cada año)"
- Desafíos aún sustanciales
 - Alto costo de mantenimiento
 - Métodos de selección modificados con el tiempo
- Pero...
 - Estudiantes no tradicionales indistinguibles de los estudiantes de la ruta tradicional
 - No se ve como una 'opción blanda'

La validez revestida

53

“La validez es un juicio evaluativo integrador del grado en que la evidencia empírica y los fundamentos teóricos respaldan la adecuación y adecuación de las inferencias y acciones basadas en los puntajes de las pruebas u otros modos de evaluación”. (Messick, 1989 p. 13)

Consecuencias sociales:

- “Preocupación correcta, concepto incorrecto” (Popham, 1997)
- No hay tal cosa como “validez consecucional”

- “Como ya se ha enfatizado varias veces, no se trata de que las consecuencias sociales adversas del uso de la prueba invaliden el uso, sino que las consecuencias sociales adversas no deben atribuirse a ninguna fuente de invalidez de la prueba, como la variación irrelevante del constructo”. (Messick, 1989 p. 88)

- “Si las consecuencias sociales adversas son empíricamente atribuibles a las fuentes de invalidez de la prueba, entonces la validez del uso de la prueba está en peligro. Si las consecuencias sociales no se pueden rastrear de esa manera, o si el proceso de validación puede descartar las fuentes de invalidez de la prueba como los determinantes probables, o al menos hacerlas [salto de página] menos plausibles, entonces la validez del uso de la prueba no se anula. (pág. 88-89)

- “Las consecuencias sociales adversas asociadas con la interpretación y el uso válidos de la prueba pueden implicar los atributos evaluados válidamente, sin duda, ya que funcionan bajo las condiciones sociales existentes del entorno aplicado, pero no son en sí mismos indicativos de invalidez”. (pág. 89)

Resumen (1)

- La validez es una propiedad de las inferencias, no de las evaluaciones.
- Dos amenazas a las inferencias válidas
 - Las puntuaciones dependen de cosas que no deberían
 - varianza irrelevante para el constructo
 - Las puntuaciones no dependen de las cosas que deberían
 - Sub-representación de constructo
- La definición de constructo es esencial porque:
 - Con una buena definición de constructo
 - El diseño de la evaluación es un proceso técnico, por lo que
 - los definidores de construcción están a cargo
 - Con una pobre definición de constructo
 - El diseño de la evaluación es un proceso cargado de valores, por lo que
 - Los desarrolladores de evaluación están a cargo

Resumen (2)

- El tema principal: ¿Cuál es la fuente de cualquier impacto diferencial?
 - “Desempaquetando” la varianza irrelevante para el constructo
 - Aleatorio: ¿El resultado depende de factores aleatorios?
 - Sistemática: ¿La evaluación apoya las mismas inferencias para todos los grupos?
 - Varianza relevante de constructo
 - ¿Es la construcción correcta?
 - ¿Produciría una construcción diferente un resultado más equitativo?

Gracias

www.dylanwiliam.net