Introducción a la medición en las ciencias sociales

Sesión 2 - Validez, Sesgo y Confiabilidad

Área de Investigación

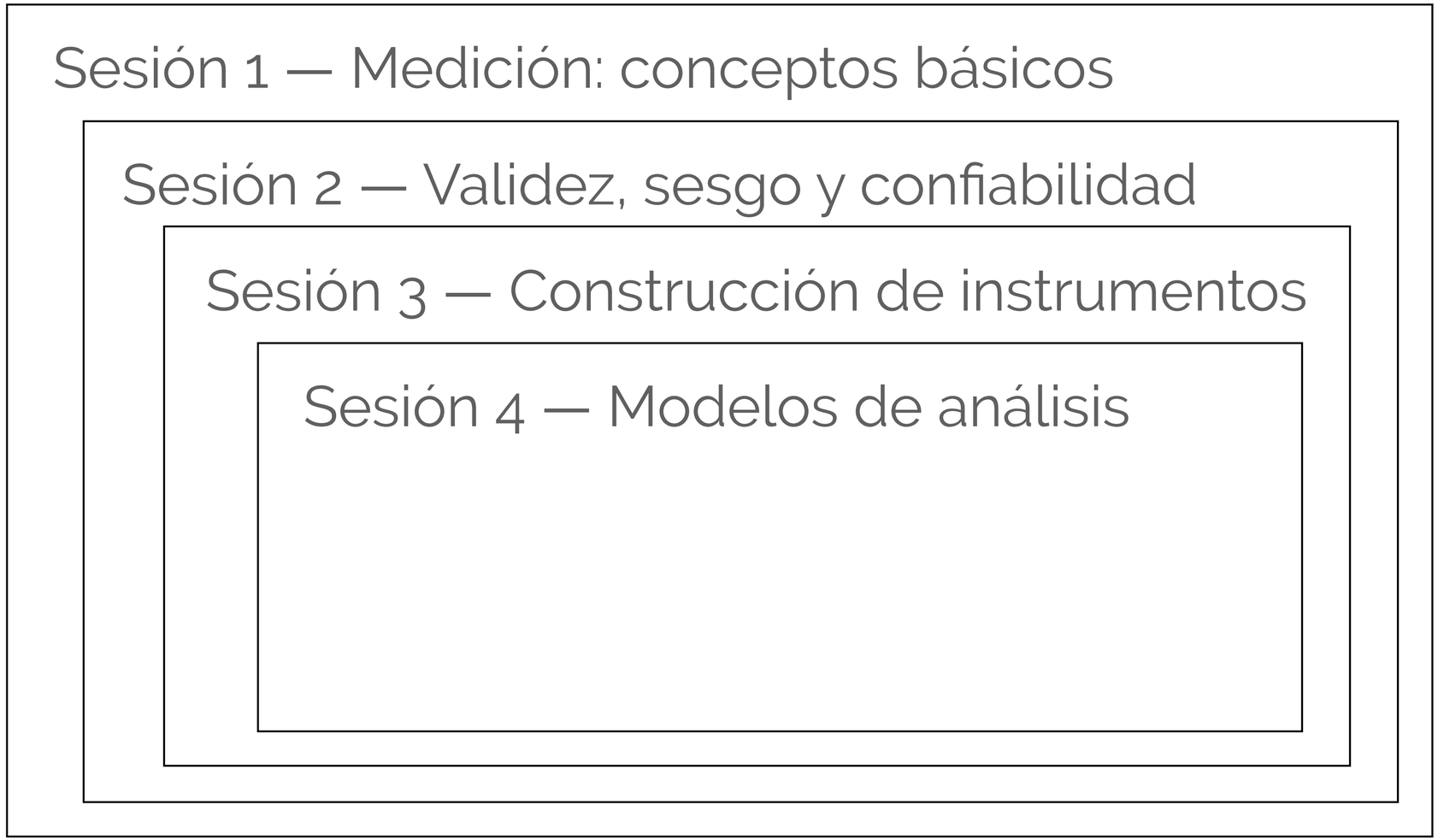
Centro UC
Medición - MIDE

Tabla de contenidos

El plan para el día de hoy

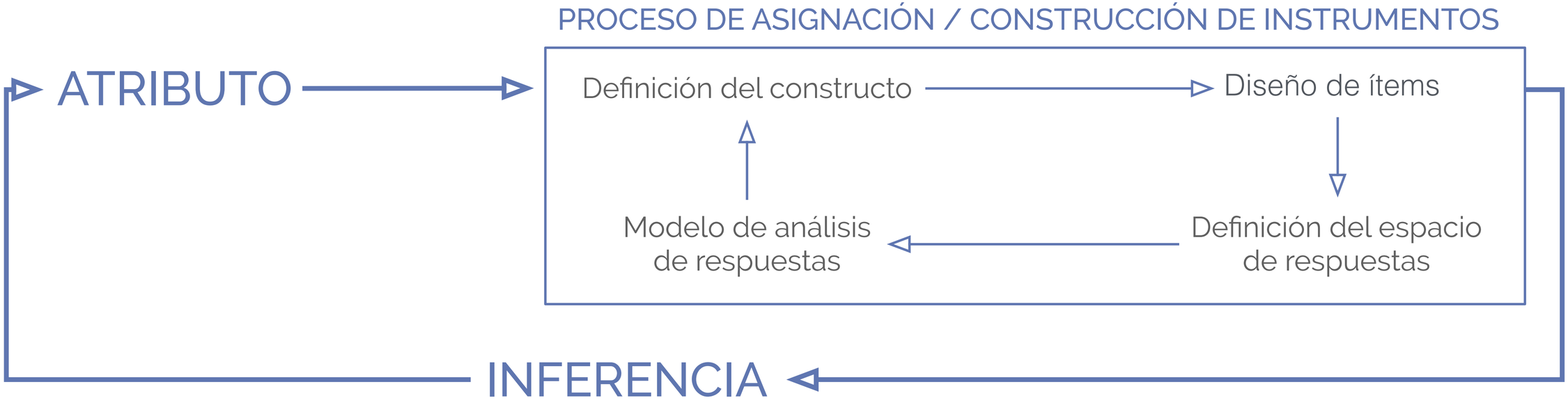
- Plan del curso + Recordar algunos conceptos
- Validez
- Sesgo
- Confiabilidad

Plan de este curso: la lógica de las cuatro sesiones



Medición en ciencias sociales

Un bosquejo general



¿Donde está la validez?

Validez

Validez: un constructo variado y fluido

Definiciones de Validez

- Las conceptualizaciones de aquello que es válido han cambiado en el tiempo y es esperable que continúen haciéndolo.
- Definición clásica de Validez: “Por validez se entiende el grado en el que una prueba o examinación mide lo que se propone medir” (Ruch, 1924).
- Definición actual de Validez: “Validez se refiere al grado en que evidencia y teoría respaldan las interpretaciones de puntajes de pruebas para sus usos propuestos” (AERA, APA & NCME, 2014)
- ¿Qué pasó entre la definición clásica de Ruch y los Estándares?

Validez: un constructo variado y fluido

La evolución

- Definición clásica de Validez: “Por validez se entiende el grado en el que una prueba o examinación mide lo que se propone medir” (Ruch, 1924).
- El estudio de la validez se lleva a cabo desde dos perspectivas: la perspectiva lógica y la empírica.
- La primera versión de los estándares publicada en 1954 por la APA (recomendaciones técnicas para tests psicológicos y técnicas de diagnóstico. En ella se distinguen cuatro tipos de validez: de contenido, concurrente y predictiva y de constructo.
- El famoso paper de Cronbach & Meehl (1955) acerca de la validez de constructo marca un hito, proponiendo explícitamente darle preponderancia por sobre las demás.
- A partir de este paper, la idea a de validez de constructo adquiere cada vez más fuerza y pasa de ser una más entre distintos tipos de validez a convertirse en el concepto unificador de validez.

Validez: un constructo variado y fluido

La evolución

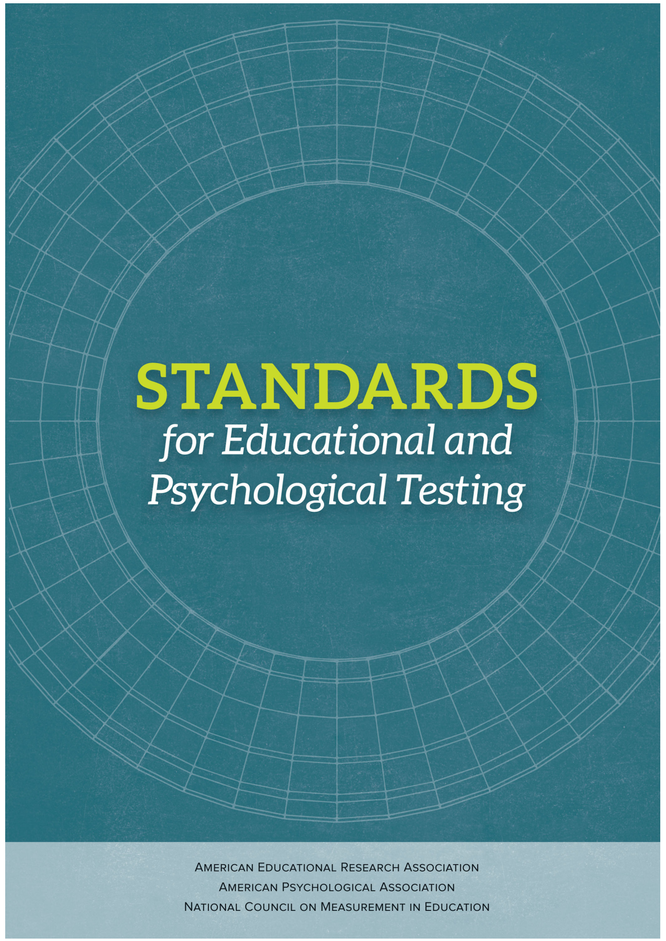
- Con el trabajo de Samuel Messick se consolida la noción de validez en torno a la validez de constructo, es decir, evidencia teórica y empírica combinadas como parte de un proceso de investigación. Se agregan componentes éticos y consideraciones sobre las consecuencias de la medición.
- Los Estándares de 1985 están fuertemente influenciados por el trabajo de Messick. Validez se introduce como un concepto unitario, entendida como una propiedad de las inferencias que se hacen a partir de una prueba. La clasificación en tres tipos (contenido, criterio y constructo) refiere ahora a tres tipos de evidencia de una sola validez.
- En los Estándares del 1999 se distinguen cinco fuentes de validez, las que deben ser integradas en un argumento de validez. La idea de argumento de validez proviene de la teoría de Michael Kane (1992).

Tres fundamentos en los estándares

Recordando

La última versión de los *Estándares* (2014) incluye tres fundamentos de la medición:

- Validez
- Precisión/confiabilidad
- Ecuanimidad



AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Primer fundamento: Validez

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

Validez se refiere al grado en que evidencia y teoría respaldan las interpretaciones de puntajes de pruebas para sus usos propuestos.

El proceso de validación implica la acumulación de evidencia relevante para dar sustento científico sólido a la interpretación de puntajes que se quiere hacer.

Validez es, por ende, la consideración más fundamental en el desarrollo de pruebas y la evaluación de pruebas... Es la interpretación de los puntajes de las pruebas para los usos propuestos la que se evalúa, no la prueba en sí misma...

Afirmaciones respecto a la validez deben referirse a interpretaciones particulares para usos específicos. Es incorrecto usar incondicionalmente la frase "la validez de la prueba".

AERA, APA Y NCME (2014)

Validez: especificación del constructo

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

El proceso de validación comienza con la declaración explícita de la interpretación del puntaje de una prueba que se quiere hacer ... incluyendo la especificación del constructo que esta pretende medir.

Casi nunca es posible atribuir un significado único al puntaje/ patrón de respuestas de una prueba, por lo que desarrolladores y usuarios de pruebas tienen la obligación de especificar la interpretación del constructo que se hará a partir de estos puntajes/patrones de respuesta.

Para dar sustento al desarrollo de una prueba, la interpretación propuesta del constructo debe elaborarse describiendo su extensión y alcance, definiendo los aspectos del constructo que serán representados. Esta descripción detallada sirve de marco conceptual de la prueba, definiendo conocimiento, destrezas, habilidades, rasgos, intereses, procesos, competencias o características que serán medidas

Idealmente, el marco conceptual debe incluir cómo el constructo representado se relaciona y distingue de otros constructo y variables.

AERA, APA Y NCME (2014)

El proceso de validación

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

La validación puede ser entendida como un proceso de construcción y evaluación de argumentos a favor y en contra de la interpretación que desea hacerse de los puntajes y su relevancia para los usos propuestos.

La decisión acerca de qué tipos de evidencia son importantes para el argumento de validez en cada caso, puede aclararse desarrollando una serie de afirmaciones que apoyen la interpretación que se pretende hacer para el uso particular de la prueba.

Una manera de identificar los argumentos que implica la interpretación de los puntaje de un test es estableciendo hipótesis que podrían ir en contra de la interpretación que se pretende hacer. Esto puede incluir hallazgos relacionados con consecuencias no intencionadas.

AERA, APA Y NCME (2014)

El proceso de validación

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

Ejemplo: se usa una prueba de matemáticas para determinar si los estudiantes van al curso avanzado de matemáticas.

¿Qué evidencia sería relevante obtener?

- que ciertas habilidades son prerrequisito para el curso avanzado
- que el contenido del test es consistente con estos prerrequisitos
- que los puntajes puedan ser generalizados entre distintos sets de ítems
- que los puntajes no estén influidos por otras variables no relevantes, como por ejemplo comprensión lectora
- que el éxito en el curso avanzado pueda ser evaluado válidamente
- que los respondentes que obtienen más altos puntajes en el test sean más exitosos que los que obtienen puntajes más bajos.

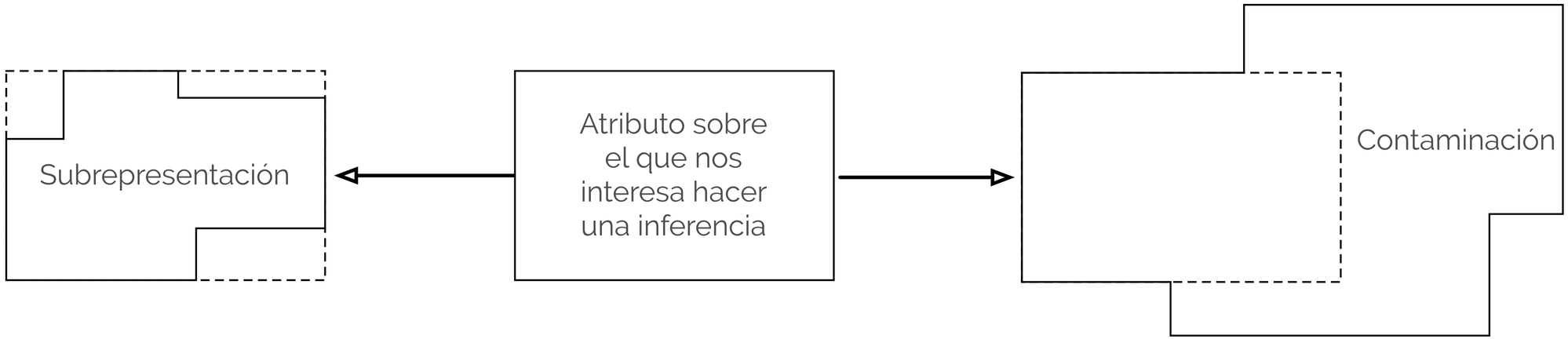
El proceso de validación avanza en la medida en que los argumentos se van articulando y se obtiene evidencia que da cuenta de su solidez

Validez: elementos que la amenazan

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

Subrepresentación (o deficiencia) del constructo) "grado con que una prueba ignora (o falla en medir) aspectos relevantes de un constructo"

Varianza irrelevante al constructo o contaminación del constructo: "grado en que los puntajes de una prueba se ven afectados por procesos ajenos a lo que esta pretende medir"



Validez: argumentación a partir de cinco fuentes de evidencia

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

- Evidencia basada en el contenido de la prueba
- Evidencia basada en los procesos de respuesta
- Evidencia basada en la estructura interna
- Evidencia basada en la relación con otras variables
- Evidencia basada en las consecuencias de la prueba

Validez: evidencia basada en el contenido

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

"Se puede obtener del análisis de las relaciones entre el contenido de la prueba y el constructo que se intenta medir"

[La obtención de] evidencia basada en el contenido puede incluir análisis lógicos o empíricos de la adecuación con que el contenido de una prueba representa el dominio conceptual y de la relevancia del dominio conceptual en la interpretación de los puntajes que se pretende hacer"

AERA, APA Y NCME (2014)

Ejemplos de evidencia basada en el contenido de la prueba:

- Mapas de constructo
- Documentación del proceso de construcción de preguntas y de corrección (por ejemplo: plantillas de preguntas, hipótesis respecto a la dificultad de las preguntas)
- Literatura respecto a la medición del constructo
- Revisiones de expertos

Validez: evidencia basada en el contenido

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

- ¿Tenemos una definición clara del constructo?
- ¿Qué consideramos una definición clara?
- ¿Podemos interpretar las respuestas de las preguntas como evidencia respecto al constructo?
- ¿Tenemos cobertura de todos los aspectos relevantes del constructo a través de las preguntas?

Validez: evidencia basada en procesos de respuesta

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

"Algunas interpretaciones sobre constructos requieren más o menos explicitación acerca de los procesos cognitivos involucrados al responder una pregunta"

"El análisis del proceso de respuesta de los individuos entrega información respecto del ajuste entre el constructo y la naturaleza particular del desempeño o respuesta ejecutado por los respondentes"

"Esta información puede conducir a reconsiderar el formato de una prueba"

AERA, APA Y NCME (2014)

Ejemplos de evidencia basada los procesos de respuesta:

- Entrevistas cognitivas, pensamiento en voz alta
- Estudios de Eye-tracking
- Datos de registro en procesos de respuesta
- Tiempos de respuesta

Validez:
evidencia basada en procesos de respuesta

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

- ¿Están los estudiantes respondiendo por las razones que se espera?
- ¿Qué estrategia están usando los respondentes para resolver un problema?
- ¿Elicita la prueba los procesos cognitivos propios del constructo que se quiere medir?
- ¿Tenemos cobertura de todos los procesos cognitivos relevantes del constructo a través de las preguntas?

Validez: evidencia basada en la estructura interna

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

"Un análisis de la estructura interna de una prueba puede indicar el grado en que la relación entre las preguntas de una prueba y sus demás componentes se adecuan al constructo sobre el cual se basan las interepretaciones"

AERA, APA Y NCME (2014)

Este tipo de evidencia obedece a la pregunta: ¿se comportan las respuestas de la prueba de acuerdo a lo que se espera?

Esto supone la existencia de una teoría a la base del constructo suficientemente clara para poder hacer estas predicciones.

Ejemplos de evidencia basada en la estructura interna:

- Estadísticas de ajuste, dificultad de las preguntas; mapas de ítem-persona
- Contraste entre ranking de dificultad esperado y observado, comparación de modelos
- Curva característica del ítem, funcionamiento diferencial del ítem

Validez: evidencia basada en la relación con otras variables

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

"El análisis de la relación de los resultados de las pruebas con las variables externas a la prueba proveen otra fuente importante de evidencia de validez"

AERA, APA Y NCME (2014)

La evidencia basada en la relación con otras variables incluye:

- Evidencia convergente y discriminante (correlación positiva con ciertas variables y negativa o inexistente con otras)
- Examinación de relaciones entre la prueba y un criterio externo por medio de diseños concurrentes o predictivos
- Posibilidad de generalizar la validez

Validez: evidencia basada en las consecuencias de la evaluación

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

"Involucra la recolección de evidencia para evaluar qué tan apropiadas son las interpretaciones propuestas para los usos intencionados."

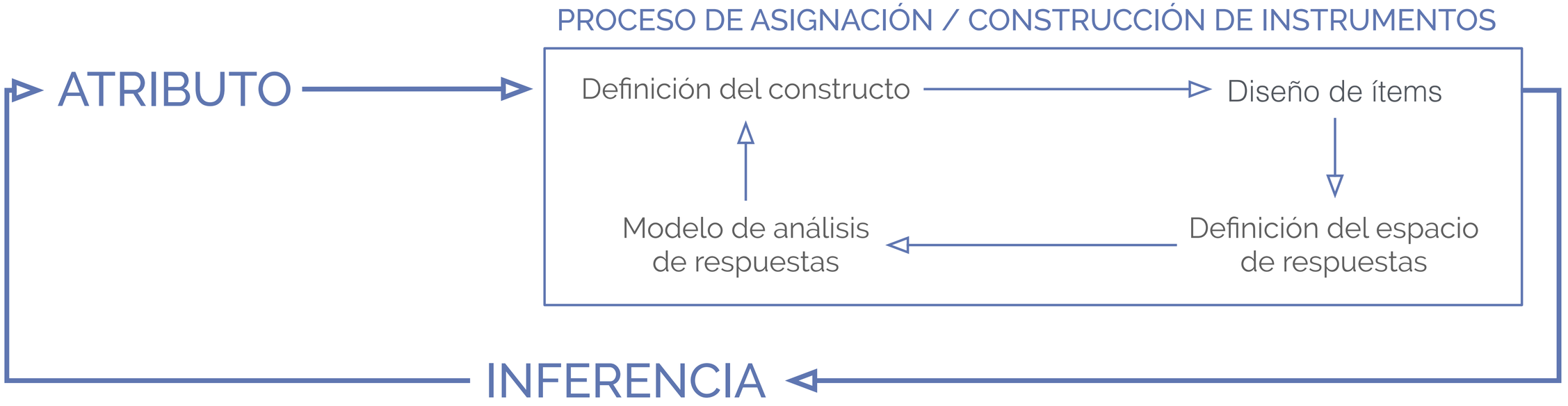
AERA, APA Y NCME (2014)

La evidencia basada en las consecuencias de la evaluación:

- Busca distinguir entre consecuencias relacionadas directamente con la interpretación y el uso de aquellas más indirectas o que podrían ocurrir a más largo plazo
- Incluye la examinación del uso en la práctica de los instrumentos, consecuencias inesperadas y especulación basadas en teorías de la acción,

Medición en ciencias sociales

Un bosquejo general



¿Dónde está la validez?


Proceso de validación - Proceso de validación - Proceso de validación - Proceso de validación - Proceso de validación - Proceso de validación - Proceso de validación - Proceso de validación
Definición del attributo
Creación de preguntas
Definición del espacio de respuesta
Modelo de medición
Evidencia basada en relación con otras variables
Evidencia basada en la estructura interna
Evidencia de contenido
Evidencia de procesos de respuesta

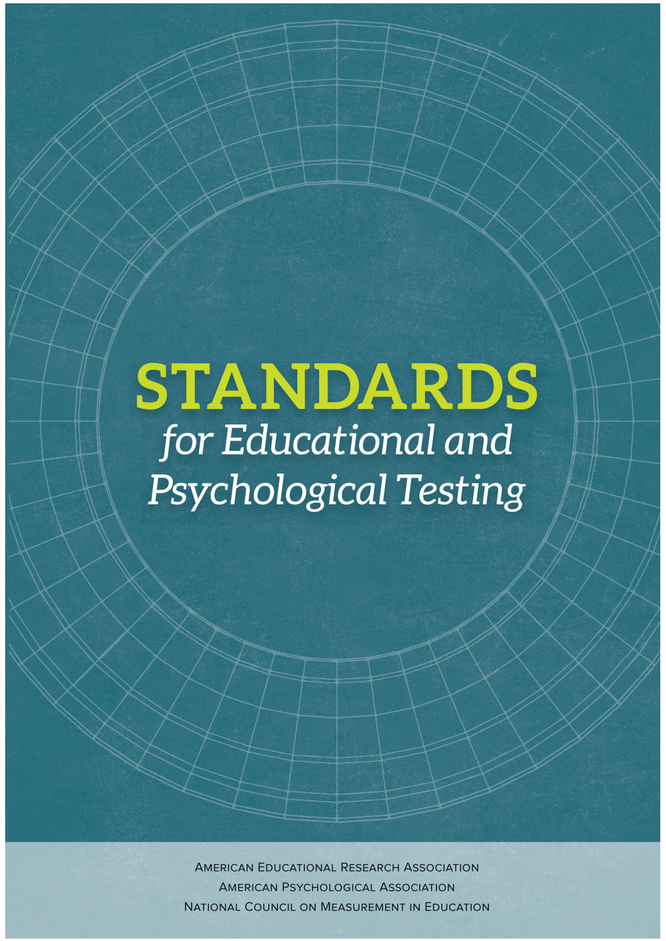
Sesgo

Tres fundamentos en los estándares

Recordando

La última versión de los *Estándares* (2014) incluye tres fundamentos de la medición:

- Validez
- Precisión/confiabilidad
- Ecuanimidad



AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Tercer fundamento: Ecuanimidad

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

La última versión de los estándares (publicada en el 2014) incluyó la ecuanimidad como un tercer fundamento de la medición, además de los fundamentos tradicionales de validez y confiabilidad.

Una prueba que es ecuánime de acuerdo a la definición de los estándares refleja el mismo constructo o constructos para todas quienes responden la prueba, y sus puntajes tienen el mismo significado para todos los individuos de la población de interés; una prueba ecuánime no da ventajas o desventajas a algunos individuos debido a características irrelevantes al constructo de interés.

AERA, APA Y NCME (2014)

Ecuanimidad como ausencia de sesgo

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

El sesgo ocurre cuando una prueba favorece o perjudica a grupos de examinados. Por ejemplo: *género, idioma, origen étnico, o estatus socioeconómico.*

Los *Estándares para el uso de pruebas educacionales y psicológicas*, hablan de sesgo como una amenaza fundamental para la ecuanimidad en las pruebas.

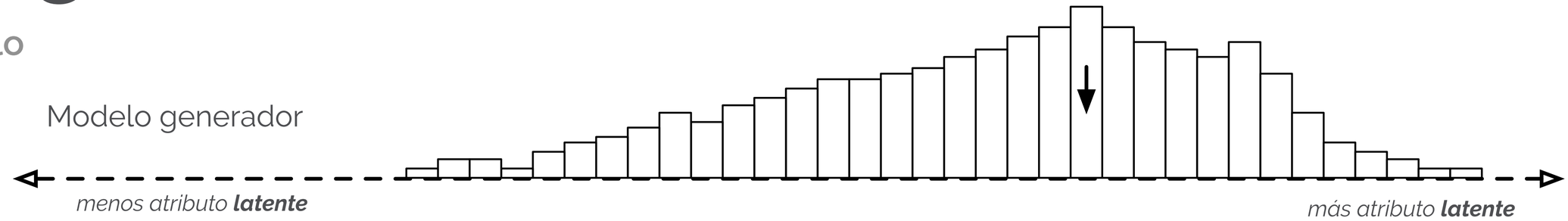
Añaden dos conceptos importantes que han surgido en la literatura, para minimizar el sesgo.

- Accesibilidad: Todos quienes toman la prueba deben tener una oportunidad sin obstáculos de demostrar su habilidad en el constructo que se esté midiendo.
- Diseño universal: Es un enfoque para el diseño de pruebas que busca maximizar la accesibilidad para todos los examinados destinados.

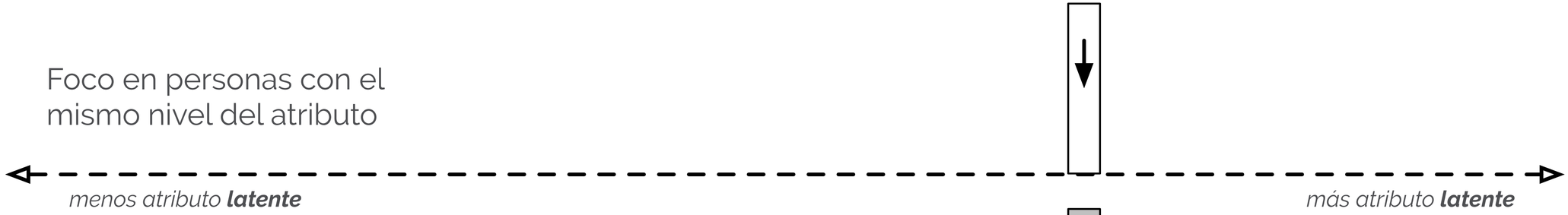
Sesgo

Ejemplo

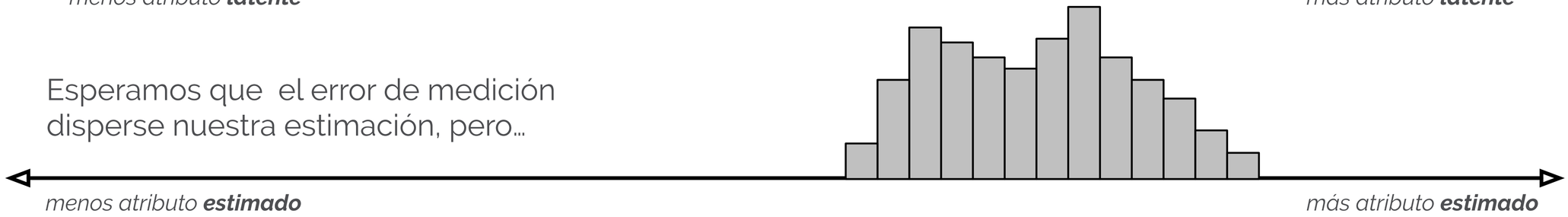
Modelo generador



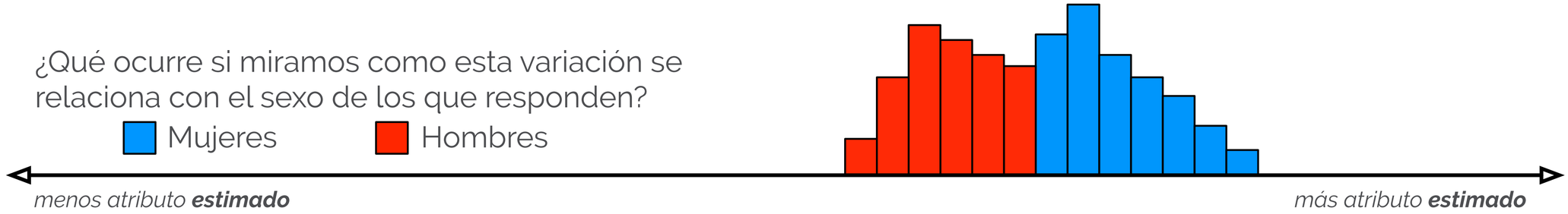
Foco en personas con el mismo nivel del atributo



Esperamos que el error de medición disperse nuestra estimación, pero…



¿Qué ocurre si miramos como esta variación se relaciona con el sexo de los que responden?



Ecuanimidad/Sesgo

Enfatizando

Ecuanimidad: Una evaluación o prueba ecuánime mide para todos los grupos de evaluados un atributo de interés, permitiendo interpretar los resultados con igual validez para todos ellos.

Sesgo: Los grupos no difieren en cuanto al estado real del atributo. Sin embargo, la prueba sugiere que sí lo hacen. En otras palabras, existe sesgo cuando para un mismo nivel del atributo que se está midiendo dos grupos o personas obtienen puntuaciones distintas.

Visión general de tipo de sesgo y sus posibles causas

Tipos de sesgo	Fuentes
Constructo	Incompleta definición del constructo a través de los grupos.
	Adecuación diferencial del contenido de los ítems
	Una pobre muestra del constructo
Método	Conveniencia diferencial social
	Falta de comparabilidad de las muestras
	Diferencias físicas en las condiciones de la toma del test
	Familiarización diferencial con el proceso de respuesta
	Efecto de quién toma el test
	Problemas de comunicación entre los examinados y quien toma el test
Ítem	Inadecuada formulación del ítem
	Uno o unos pocos items pueden recurrir a los rasgos o habilidades adicionales
	Diferencias accidentales en la adecuación de los contenidos de los ítems

Van de Vijver, F.J.R. Poortinga, Y.H. (1997) Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment.

Detección de Sesgo

Una serie de técnicas se han desarrollado para la detección de funcionamiento diferencial del ítem (DIF).

La idea fundamental para cualquier cálculo de DIF consiste en estudiar si al mismo nivel de rasgo, la probabilidad de acierto al ítem es similar o no dependiendo del grupo de pertenencia

Será motivo de estudio para una próxima oportunidad!

Invarianza

- Las mediciones son (por definición) sensibles a la variación en el atributo de interés. Pero, en principio, deben ser invariantes a todo lo demás.

- Invarianza: ¿Qué transformaciones podrían cambiar nuestras inferencias sobre un atributo o el atributo en sí mismo?

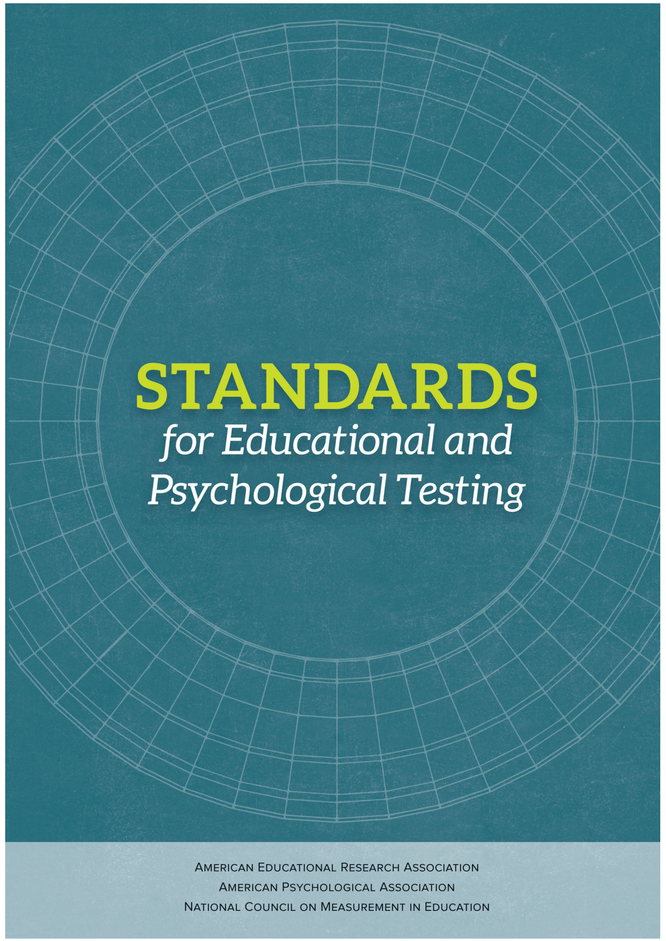
Confiabilidad/ Precisión

Tres fundamentos en los estándares

Recordando

La última versión de los *Estándares* (2014) incluye tres fundamentos de la medición:

- Validez
- Precisión/confiabilidad
- Ecuanimidad



AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Segundo fundamento: Confiabilidad/Precisión

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

De acuerdo a los estándares, el error o incertidumbre de una medición es considerado en términos de la confiabilidad de un instrumento.

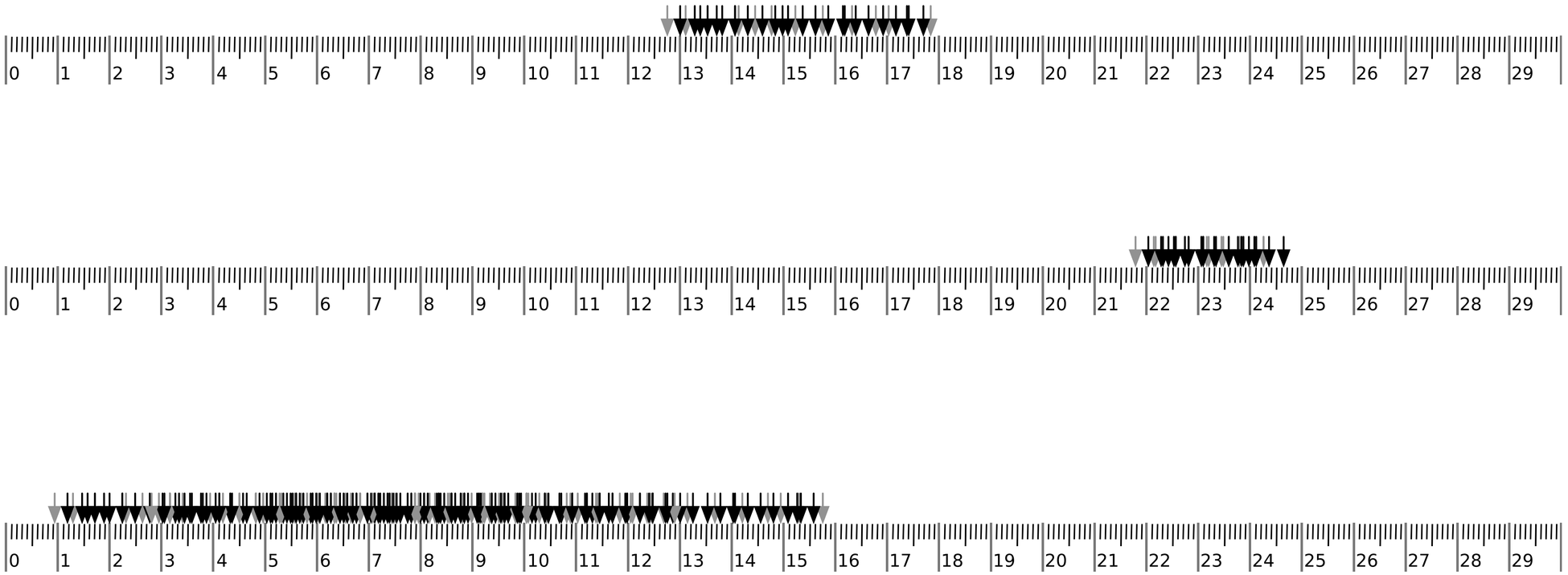
CONFIABILIDAD COMO FUNDAMENTO:

La consistencia de los puntajes entre realizaciones de un procedimiento de medición, independientemente de cómo se calcule o reporte.

Este fundamento apunta, en otras palabras, a la variación que se observaría si repitiéramos una medición en múltiples ocaciones.

Error de medición

Un ejemplo del error como variación en mediciones repetidas



El error de medición

De acuerdo a los Estándares para el uso de pruebas educacionales y psicológicas

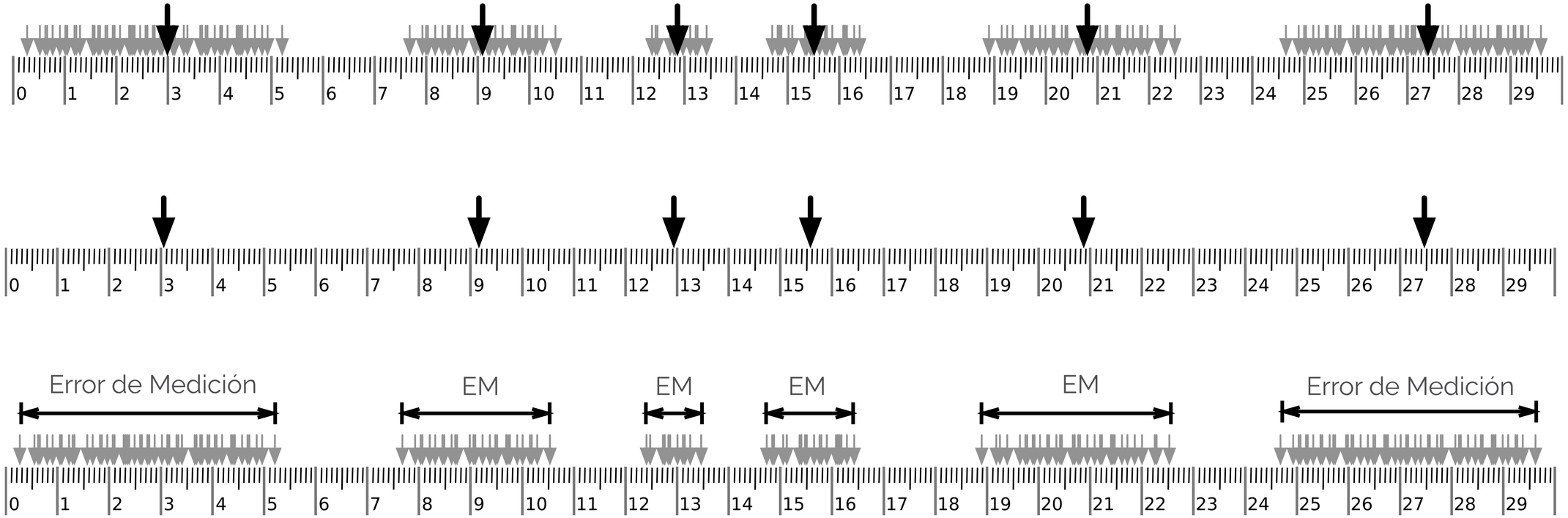
El error de medición reduce la utilidad de los puntajes de pruebas.

Limita qué tanto puede generalizarse un resultado mas allá de de las características particulares de una aplicación específica de una prueba.

Reduce la confianza que se puede tener en los resultados de una medición y por ende en la ***confiabilidad/precisión*** *de los puntajes.*

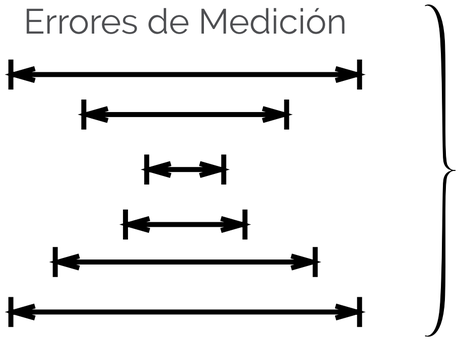
AERA, APA Y NCME (2014)

Variabilidad de una medición



Variabilidad de una medición

Un ejemplo del variabilidad de un instrumento

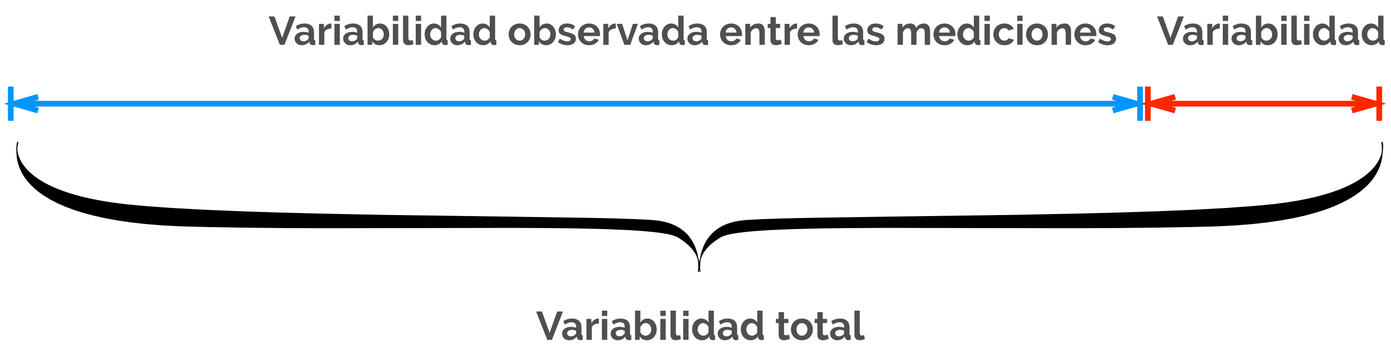


Resumimos la variabilidad producto del error de medición en todas las mediciones para calcular la

Variabilidad asociada al error

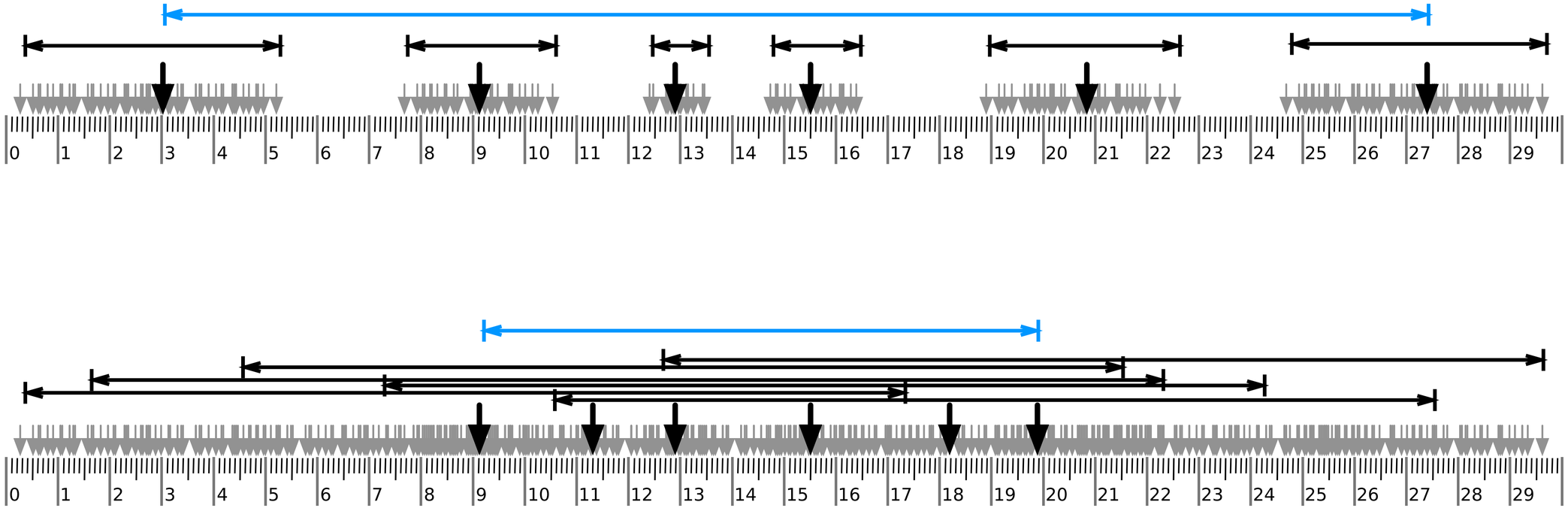
Variabilidad de una medición

Una forma de entender la confiabilidad es en términos del tamaño relativo de estos dos elementos, la variabilidad asociadas a nuestra estimaciones y la variabilidad asociada a al error de medición.



Idealmente queremos que la variabilidad en las mediciones sea mucho mas grande que la variabilidad asociada al error de medición. En otras palabras, queremos que la variabilidad atribuible a nuestras mediciones de cuenta de la mayor parte de la variabilidad total.

Variabilidad de una medición



Confiabilidad/Precisión

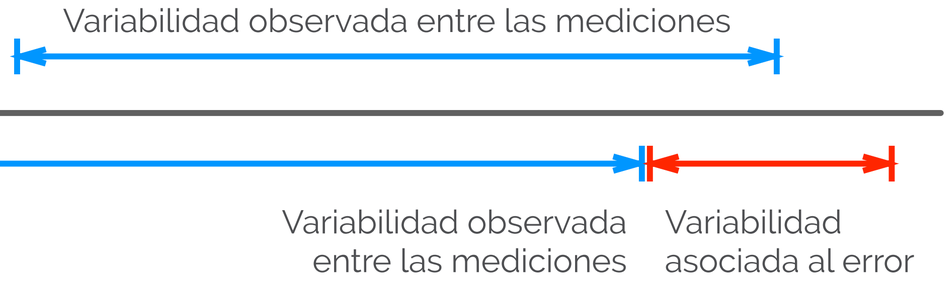
Entendida desde la idea de descomposición de variabilidad

De forma mas concreta, una de las formas de calcular la confiabilidad es en términos de la razón entre la variabilidad de nuestras mediciones y la variabilidad total.

$$\frac{\textbf{Variabilidad observada entre las mediciones}}{\textbf{Variabilidad Total}}$$

(Variabilidad observada entre las mediciones + Variabilidad asociada al error)

$\approx$



Confiabilidad/Precisión

La “confiabilidad/precisión” puede ser estimada de muchas formas distintas en medición en las ciencias sociales dependiendo entre otras cosas de:

(a) el diseño de recolección de respuestas

(b) del modelo de análisis estadístico que se esté utilizando para analizar las respuestas.

Distintos métodos y diseños de recolección de datos intentan solucionar el problema de la falta de múltiples mediciones repetidas e independientes de diferentes formas.

Confiabilidad/Precisión

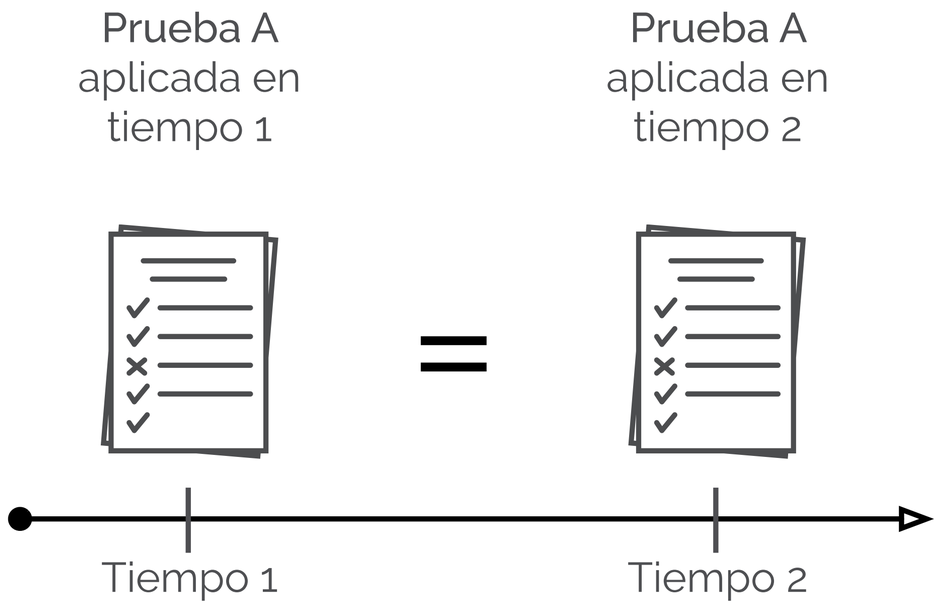
1. Coeficientes de confiabilidad Test-Retest
2. Coeficientes de formas paralelas
3. Coeficientes de confiabilidad de división por mitades
4. Coeficiente de consistencia interna
5. Consistencia entre jueces

Confiabilidad/Precisión:
Coeficiente de confiabilidad Test-Retest

Se usa para determinar la consistencia entre dos aplicaciones medidas de un momento a otro, de un mismo instrumento a un mismo sujeto.

Se calcula correlacionando los puntajes del sujeto en ambas mediciones.

¿Cuál es su dificultad?: Se basa en el supuesto de que la característica es estable en el tiempo y por lo tanto esperamos encontrar mediciones similares en dos puntos del tiempo



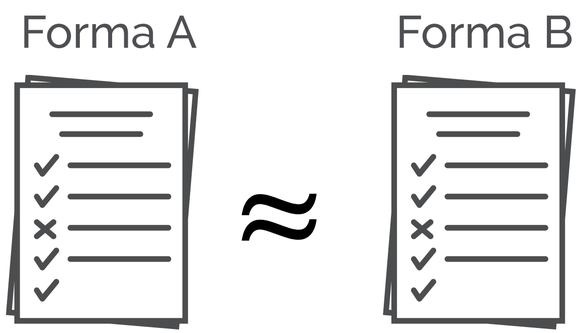
Confiabilidad/Precisión:
Coeficiente de formas paralelas

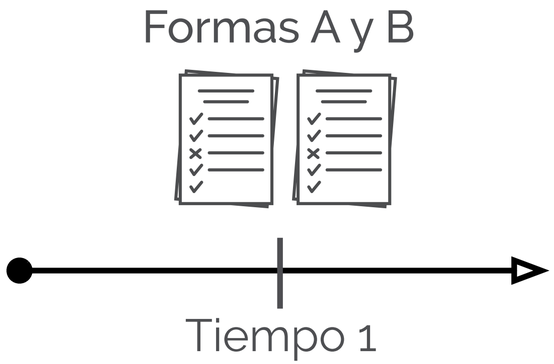
Se usa para determinar la consistencia entre los resultados de dos pruebas construidas para ser equivalentes o “paralelas”, aplicadas a los mismos sujetos.

Las formas paralelas de un test están compuestas por distintos ítems que se construyen siguiendo las mismas especificaciones.

Formas A y B son distintas, pero creemos que son equivalentes en sus propiedades como instrumentos de medición.

Ya que creemos que son equivalentes, el aplicar ambas formas al mismo tiempo contaría como hacer dos mediciones simultáneamente.





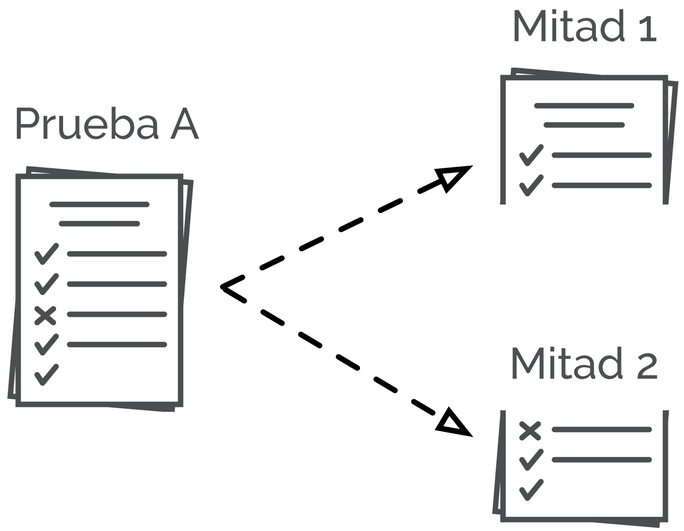
Confiabilidad/Precisión:
Coeficiente de confiabilidad de división por mitades

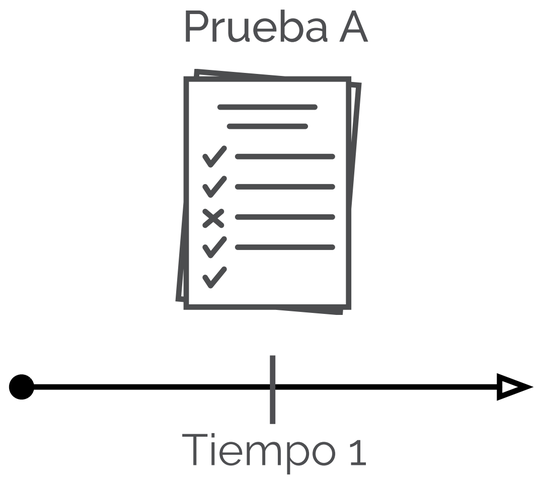
Se divide al azar todos los items que pretenden medir el mismo constructo en las dos formas.

Se administra el único instrumento a un conjunto de sujetos.

Se construye una sola prueba, pero se asume que si dividimos la prueba en dos, ambas mitades pueden ser consideradas como formas paralelas.

Se aplica la única prueba que fue diseñada, y posteriormente, al momento de hacer análisis de las respuestas se dividirán las preguntas.





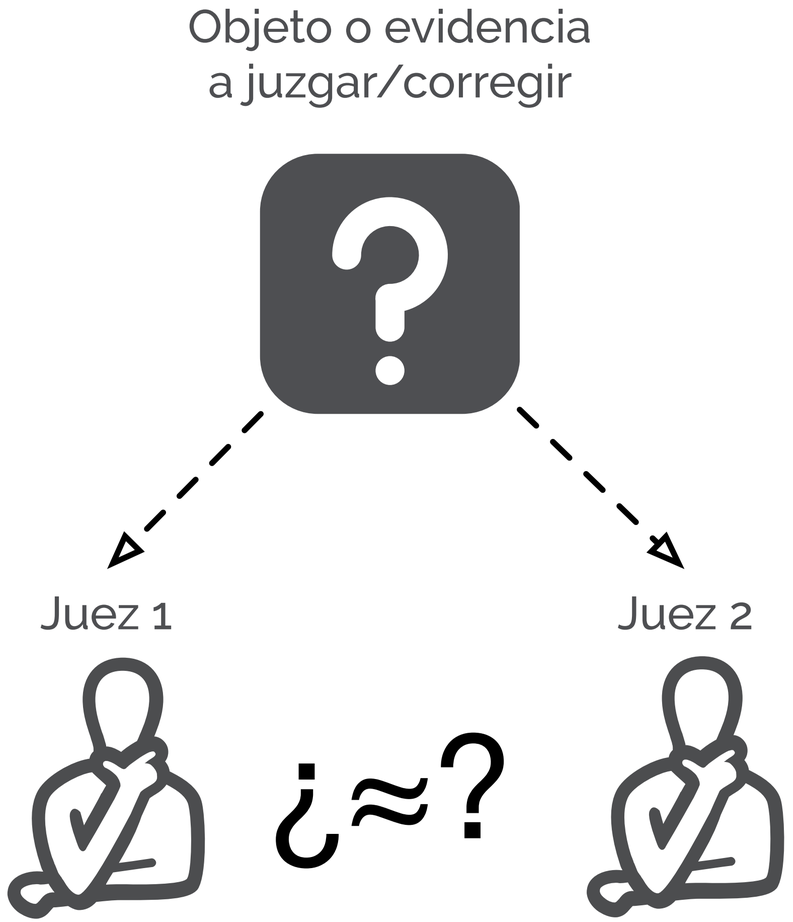
Confiabilidad/Precisión: Coeficiente de consistencia interna

- No es fácil construir formas paralelas de un test.
- No es posible aplicar la prueba (o una forma alternativa) en dos momentos distintos.
- ¿Cómo obtener entonces una estimación del coeficiente de confiabilidad del test cuando este se aplica sólo una vez?
  - Usando coeficientes de consistencia interna como el alfa de Cronbach
- Estos coeficientes usan la información sobre la proporción de variabilidad que está representada por el modelo.

Confiabilidad/Precisión: Consistencia entre jueces

Cuando los test están integrados por ítems de construcción, la asignación de puntajes suele exigir el recurso de jueces encargados de traducir a un número o categoría la respuesta dada por el sujeto a un ítem.

Como la tarea de los jueces puede introducir un componente subjetivo importante en el proceso de asignación de puntajes es conveniente calcular algún coeficiente o índice de acuerdo entre los jueces.



Confiabilidad/Precisión: Consistencia entre jueces

- Fuentes de error de medición ocurre cuando hay inconsistencia entre jueces.
  - Hay evaluadores que no se ajustan plenamente la formación y por lo tanto nunca se aplican las guías de puntuación de una manera correcta.
  - Existen diferencias en la gravedad del evaluador. Es decir, algunos evaluadores tienden a puntuar más alto o bajo que otros.
  - Hay evaluadores que su tendencia es utilizar puntuaciones extremas.
  - Hay evaluadores que sean inconsistentes con ellos mismos por una variedad de razones.

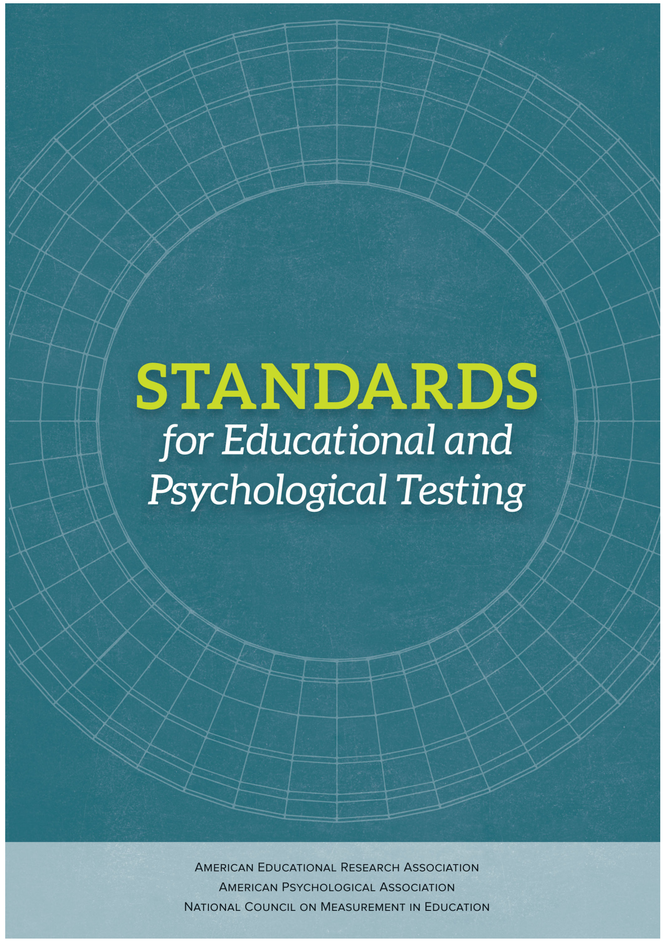
Confiabilidad/Precisión: Consistencia entre jueces

- Importantes pasos a tomar para reducir la inconsistencia entre jueces
  - un programa de formación de evaluadores
  - un sistema de monitoreo que ayude a los administradores y evaluadores saber que están en buen camino
  - Hay tres esenciales formas de monitorear el trabajo de los evaluadores
    - comparar el puntaje otorgado con uno de referencia (asignado por expertos)
    - re- evaluar (por expertos) algunas de sus calificaciones
    - comparar los registros de calificaciones de las calificaciones de todos los evaluadores

Tres fundamentos en los estándares

La última versión de los *Estándares* (2014) incluye tres fundamentos de la medición:

- Validez
- Precisión/confiabilidad
- Ecuanimidad



AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Créditos

Clases preparadas por el área de investigación MIDE – agosto 2016

Diego Carrasco — María Inés Godoy — Daniela Jiménez — David Torres Irribarra

y agradecimientos a Mauricio Rivera

Todos los símbolos provienen de thenounproject.com

 Creado por Quinn Keaveney

 Creado por Jonathan Gibson

 Creado por designify.me

 Creado por Gilbert Bages

 Creado por Agniraj Chatterji

 Creado por TMD

 Creado por Christopher Smith

 Creado por Hannah Strobel

 Creado por Jaime Carrion

 Creado por Gonzalo Bravo

 Creado por Takao Umehara

Introducción a la medición en las ciencias sociales

Sesión 2 - Validez, Sesgo y Confiabilidad

Área de Investigación

Centro UC
Medición - MIDE