UNIVERSITAT DE BARCELONA

CONSTRUCCIÓN Y ADAPTACIÓN DE TEST Y CUESTIONARIOS

Joan Guàrdia Olmos
Cristina Cañete Massé
Maribel Peró Cebollero

Diciembre, 2021

MÁXIMA DE PARTIDA

La producción de instrumentos es inmensa, probablemente existe alguno que se ajusta a sus necesidades teóricas y/o aplicadas. Analice su estandarización y baremos a su realidad y adáptelo si es preciso.

Si no existe esa propuesta ……

!INVÉNTELA!

BASES DE LA MEDICIÓN

Weber

$$K = \Delta E / E$$

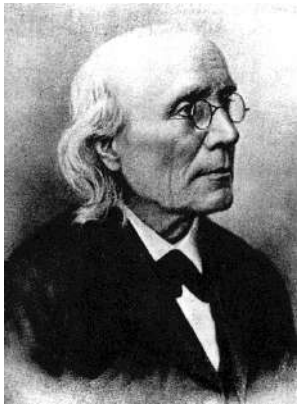
donde:

K = constante de Weber

ΔE = incremento mínimo en la magnitud del estímulo necesario para percibir un cambio en la sensación

E = magnitud del estímulo

FECHNER Y LA ESTRUCTURA DE LA MEDIDA



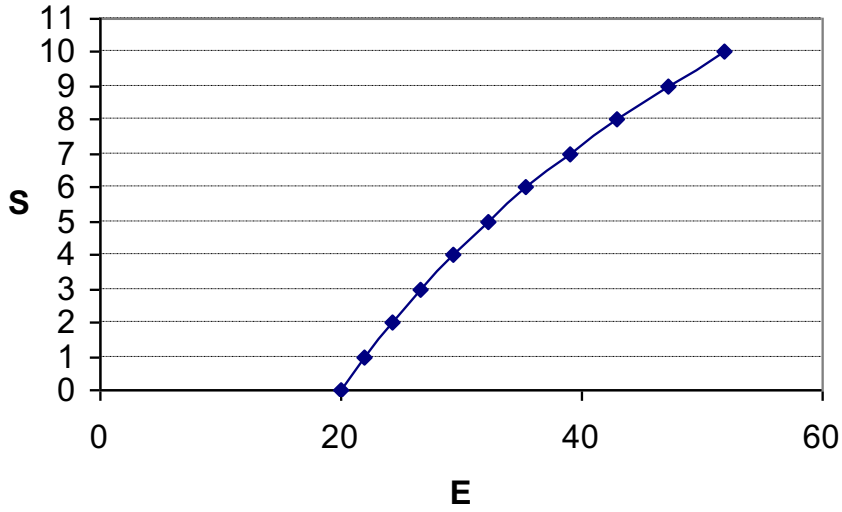
Asume la ley de Weber y la igualdad de las diferencias apenas perceptibles. (d.a.p.)

Ley Logarítmica

Si se incrementa la estimulación en proporción geométrica las sensaciones lo harán en progresión aritmética.

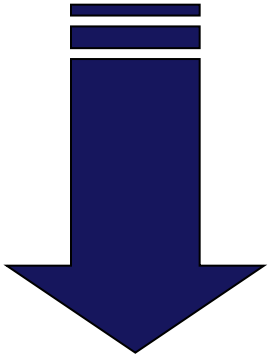
$$S = K * Log(E)$$

E	S
20	0
22	1
24.2	2
26.62	3
29.282	4
32.2102	5
35.4312	6
38.9744	7
42.8718	8
47.1589	9
51.8747	10



Psicometría

Conjunto de métodos, técnicas y teorías implicadas en la medida de las variables psicológicas

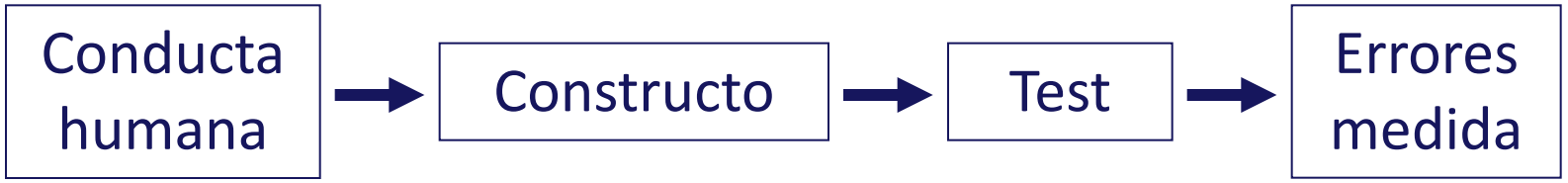


Propiedades métricas

Teoría de los Tests

Instrumento de medida o test:

Sistematización de un conjunto de indicadores con la finalidad de evaluar diversos niveles de un constructo (variable teórica inobservable que solo se puede medir mediante indicadores observables que lo representen).



Historia de los Test

1. Test sensoriales y motores – laboratorio antropométrico de Galton (fundado en 1882).
2. Escalas de inteligencia – Alfred Binet (1817).
3. Test colectivos:
   - 3.1. Test α y β del ejército (EEUU – 1ª Guerra Mundial).
   - 3.2. Test de personalidad – *Self report inventory* (Woodworth, 1917).
   - 3.3. Test factoriales.
   - 3.4. Test de modelos de rasgo latente o teoría de respuesta al ítem.

Fases en la construcción de un test

1. Análisis del rasgo – definición operacional.
2. Elaboración ítems.
3. Elección de los ítems.
4. Estudio de la fiabilidad.
5. Estudio de la validez.
6. Baremación.
7. Normas de aplicación.

Tipos de test psicométricos

- En función del material utilizado:
  - De lápiz y papel
  - Manipulativos
  - De medidas fisiológicas
- En función de las instrucciones:
  - Verbales
  - No verbales
- En función del formato:
  - De invención de respuesta
  - De elección de respuesta
  - De emparejamiento

- En función del tiempo asignado:
  - De velocidad
  - De potencia
  - Mixtos
- En función de la forma de administración:
  - Individuales
  - Colectivos

TEORIA CLÁSICA DE LOS TEST

Teoría Clásica de los Test (TCT)

Modelo lineal de la puntuación verdadera propuesto por *Spearman*:

$$\mathbf{X = V + e}$$

Donde:

X – puntuación en el test

V – puntuación verdadera, E(X)

e - error

Directrices para la construcción de los ítems del test

- ***Definir de forma específica el objetivo de la evaluación.***

- ***Especificar el contexto en que se utilizaran los ítems: población objetivo, circunstancias ambientales en que se aplicarán.***

- ***Dominio y contexto de interés***

Elaboración de los ítems del test

- ***Cerrado o de elección***
  - Elección múltiple
  - Elección binaria
  - Escalas graduadas
  - Escalas del diferencial semántico
  - Escalas de adjetivos
- ***Abierto o de elaboración de respuesta***
  - Respuesta extensa
  - Respuesta restringida
- ***Ordenación de diversos estímulos***
- ***Emparejamiento de diversos estímulos***
- ***Comparación y valoración de diversos estímulos***

Recomendaciones para la confección de ítems de respuesta múltiple

- Cada ítem dirigido a evaluar un único problema.
- Plausibilidad de las alternativas incorrectas.
- Ubicación al azar de la alternativa correcta.
- Evitar ítems que se puedan contestar de forma lógica o con sentido común.
- No repetir palabras o expresiones para cada alternativa.
- Misma longitud de las alternativas.
- No utilizar como alternativa: todas las anteriores o ninguna de las anteriores.
- Evitar las negaciones dobles (en el enunciado y en la alternativa).

Errores más habituales en la confección de ítems

- No misma plausibilidad de las alternativas incorrectas.
- Enunciados en forma negativa.
- No misma longitud de las alternativas.
- Reiteraciones innecesarias de texto.
- Uso de la alternativa: todas las anteriores o ninguna de les anteriores.

Análisis de ítems: pasos

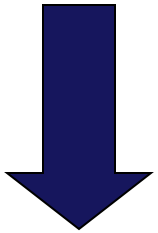
1. Decidir las principales propiedades de las puntuaciones del test.
2. Identificar los análisis de ítems más relevantes en función de estas propiedades.
3. Administrar los ítems a una muestra representativa de la población a la que está dirigida el test.
4. Estimar los análisis identificados en el segundo paso, por ítem.
5. Establecer un plan para la selección de los ítems. Identificar y revisar los que funcionen mal.
6. Seleccionar el bloque final de ítems.

Análisis de ítems: indicadores clásicos

- Índice de dificultad
- Índice de discriminación
  - Basado en las diferencias
  - Basado en las correlaciones
- Análisis de los distractores
- Fiabilidad del ítem
- Validez del ítem
- Dimensionalidad

Análisis de ítems: Índice de discriminación

Evaluar si los ítems intentan poner de manifiesto las diferencias individuales entre los sujetos en la variable que se mide



Un ítem tiene poder discriminante si las personas que tienen un nivel alto en la variable aciertan más el ítem u obtienen puntuaciones más elevadas que las personas que tienen un nivel bajo en la variable

Índice de Fiabilidad del ítem

- Se utiliza para estimar la fiabilidad con que cada ítem mide la característica o variable que mide el test.

- Depende del índice de discriminación y de la desviación típica del ítem:

$$IF_i = S_i \cdot ID_i$$

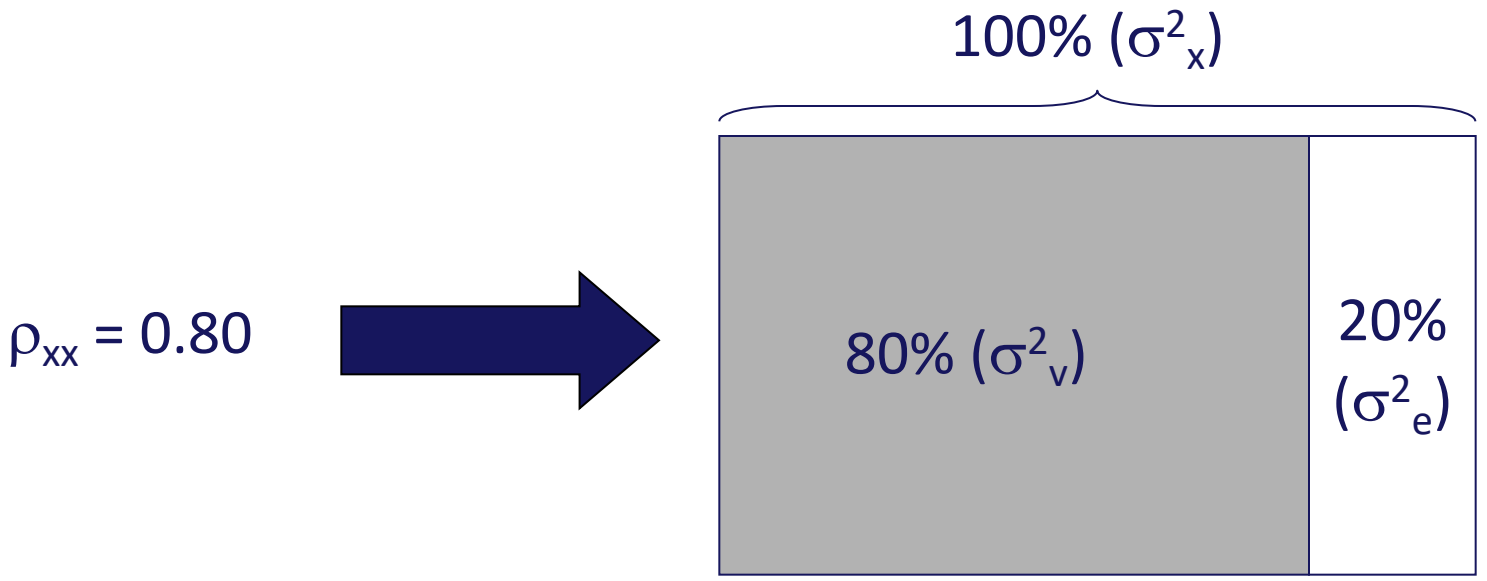
Índice de Validez del ítem

- Correlación del ítem con un criterio externo.

Fiabilidad

Precisión en la medida ≠ Error de medida

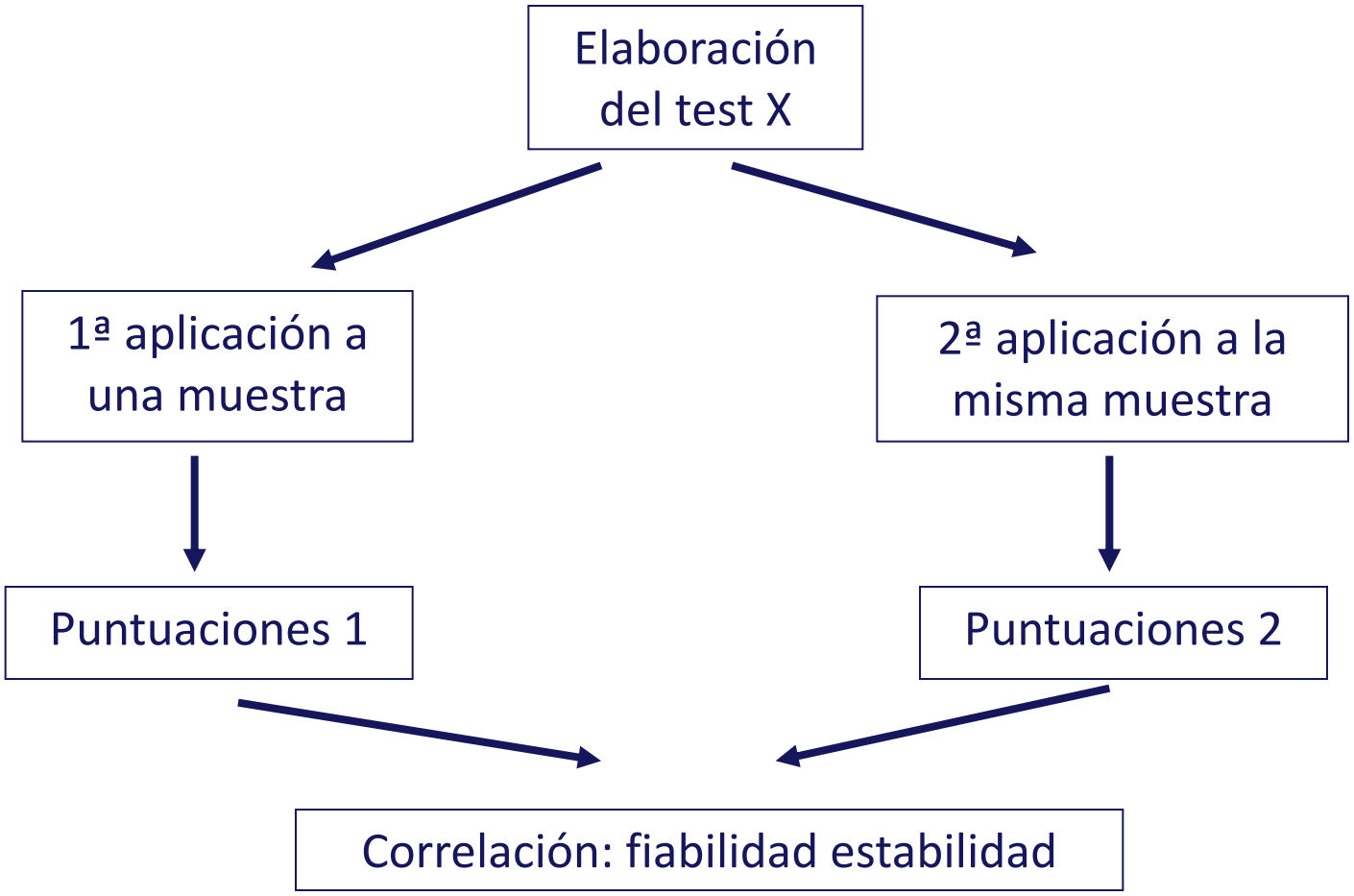
X = V + e



Procedimientos para estimar la fiabilidad

- Estabilidad de la medida → Test-retest
- Equivalencia de las medidas → Formas paralelas
- Consistencia interna → Intercorrelaciones entre los ítems:
  - Dos mitades – Corrección Spearman-Brown
  - Coeficiente alfa de Cronbach (análisis de las covarianzas entre los ítems)

Fiabilidad: test-retest

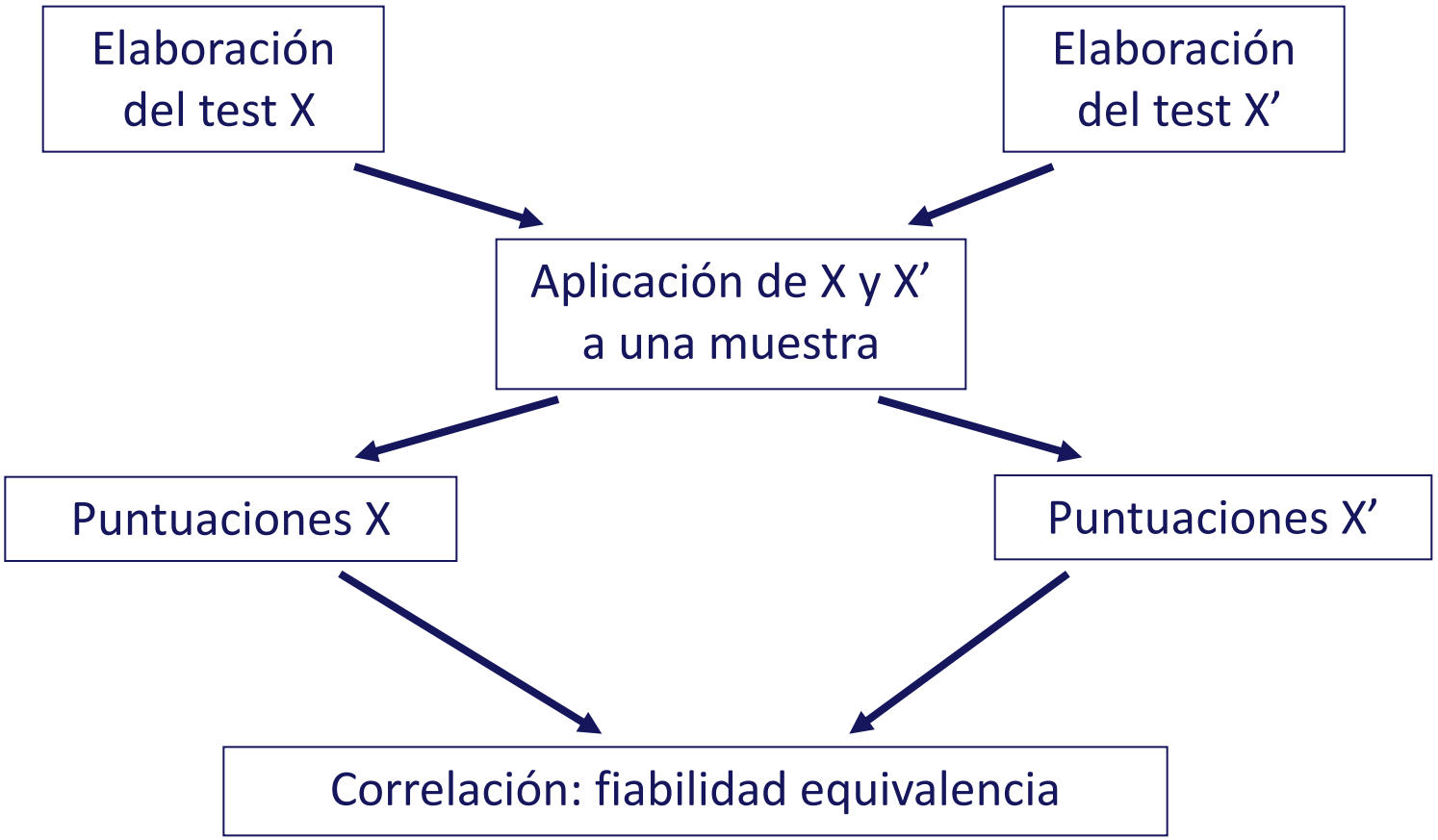


Criterios valoración fiabilidad test-retest (estabilidad medida)

Muñiz, J. (2005) Utilización de los tests. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.). *Análisis de los ítems*, (pp. 133-172). Madrid: La Muralla, S.A.

- Inadecuada: $r < 0.55$
- Adecuada pero con algunas restricciones: $0.55 \leq r < 0.65$
- Adecuada: $0.65 \leq r < 0.75$
- Buena: $0.75 \leq r < 0.80$
- Excelente: ≥ 0.80

Fiabilidad: Formas paralelas

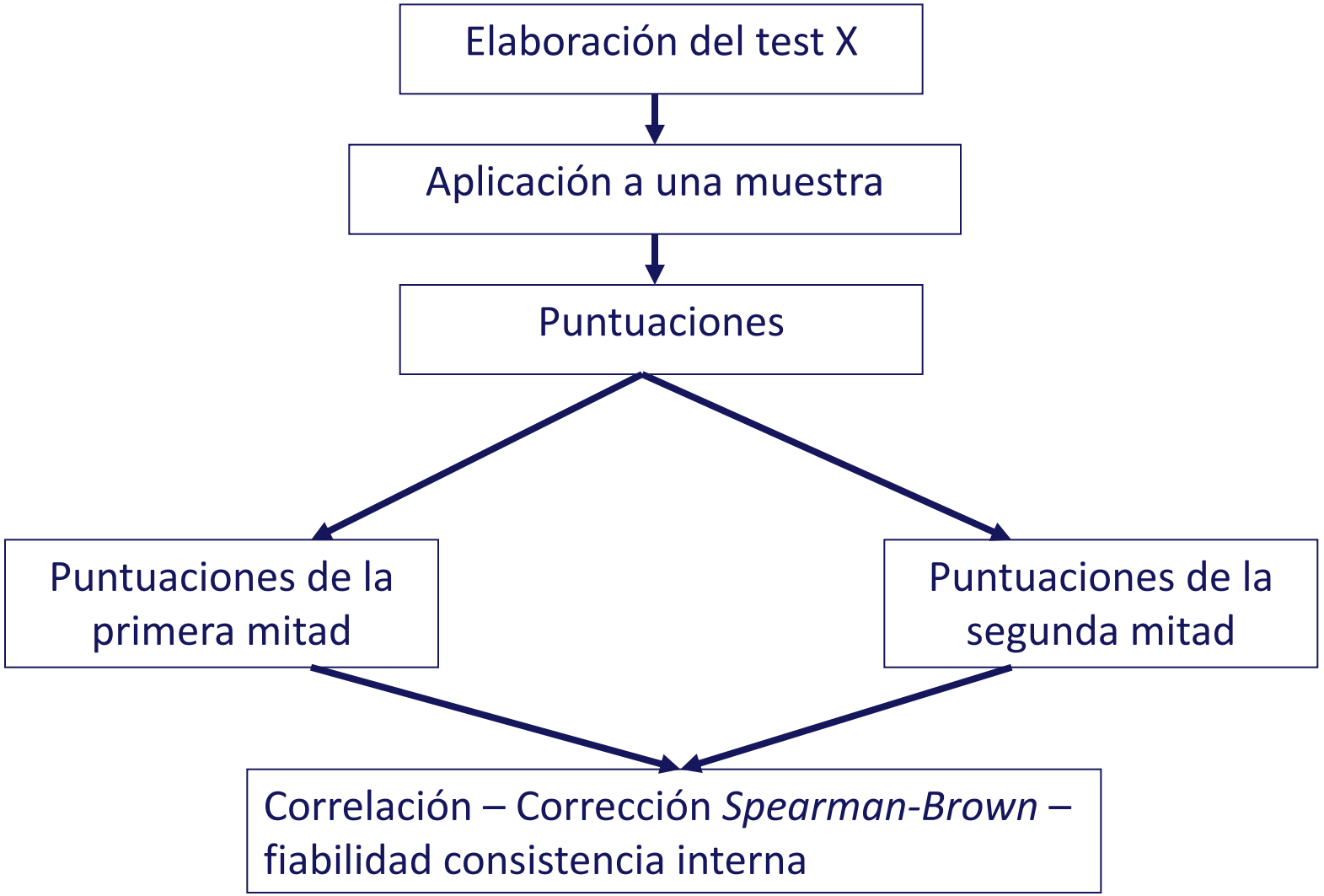


Criterios valoración fiabilidad formas paralelas (equivalencia medida)

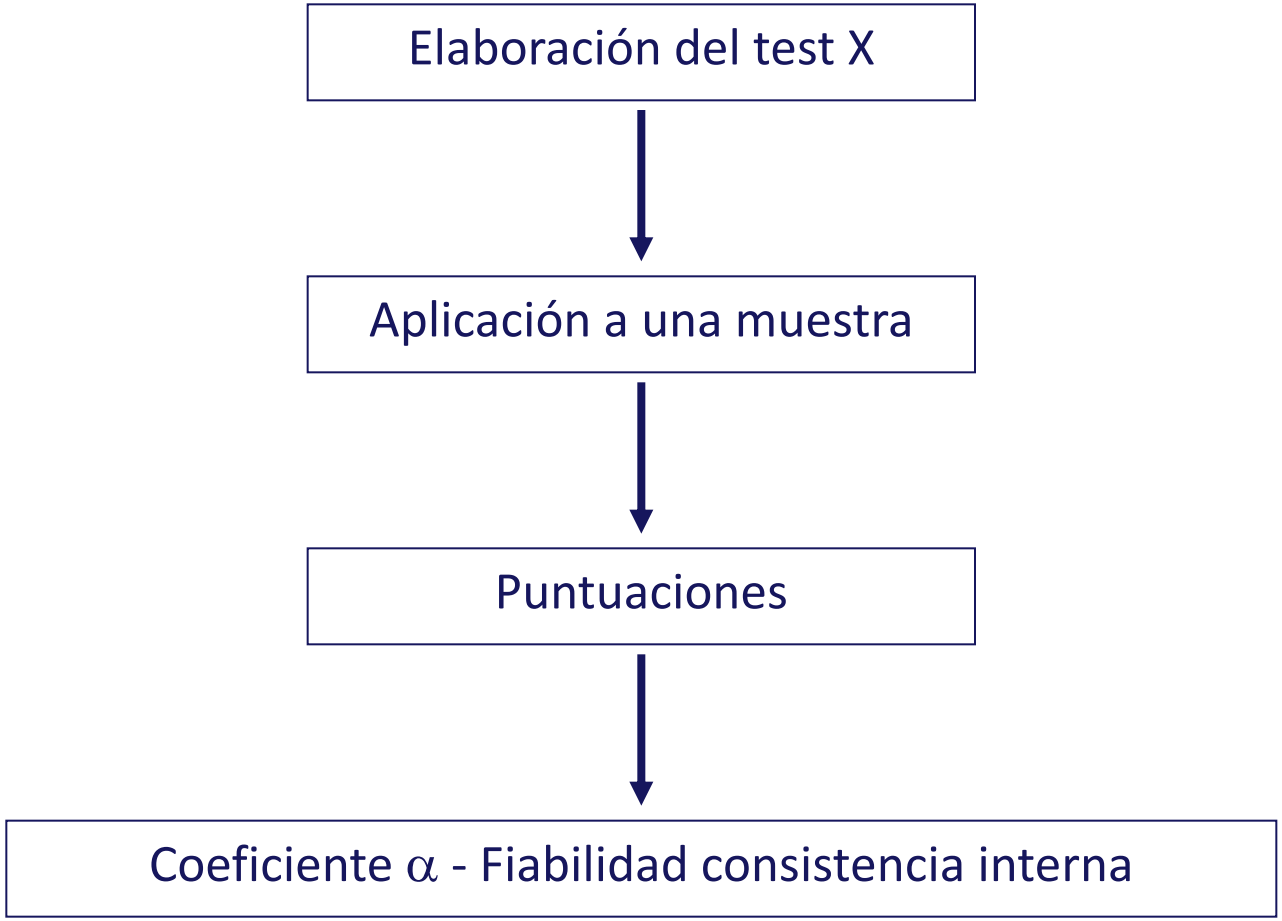
Muñiz, J. (2005) Utilización de los tests. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.). *Análisis de los ítems*, (pp. 133-172). Madrid: La Muralla, S.A.

- Inadecuada: $r < 0.50$
- Adecuada pero con algunas restricciones: $0.50 \leq r < 0.60$
- Adecuada: $0.60 \leq r < 0.70$
- Buena: $0.70 \leq r < 0.80$
- Excelente: ≥ 0.80

Fiabilidad: Dos mitades



Fiabilidad: Alfa de Cronbach



Criterios valoración fiabilidad consistencia interna

Muñiz, J. (2005) Utilización de los tests. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.). Análisis de los ítems, (pp. 133-172). Madrid: La Muralla, S.A.

- Inadecuada: $r < 0.60$
- Adecuada pero con algunas restricciones: $0.60 \leq r < 0.70$
- Adecuada: $0.70 \leq r < 0.80$
- Buena: $0.80 \leq r < 0.85$
- Excelente: ≥ 0.85

Factores que influyen en la fiabilidad del test

- Longitud o número de ítems – más ítems mayor fiabilidad.
- Homogeneidad
- Variabilidad de las respuestas y puntuaciones – al aumentar la variancia aumenta la verdadera y la del error más lentamente.
- Dificultad de los ítems
- Tiempo de respuesta
- Muestreo
- Número alternativas de respuesta
- Nivel del sujeto en la variable medida

Validez

Test Válido – Sirve para medir adecuadamente lo que se quiere medir.

Validez – Concepto unitario que incluye diversas facetas a validar:

Histórico:

- De contenido
- De criterio (concurrente y predictiva)
- De constructo

Estándares 1999:

- De contenido
- Proceso de respuesta
- Estructura interna
- Relaciones con otras variables
- Consecuencial

Validez de Contenido

Tipos:

- Aparente
- Muestral
- Curricular

Pasos:

- Definir y especificar el campo o dominio del test.
- Selección jueces expertos.
- Proceso de apareamiento de ítems con los temas de interés que mide el test.
- Selección de los ítems adecuados.

Validez de criterio

Grado de eficacia con que podremos diagnosticar o predecir la variable *criterio* a partir de las puntuaciones del test.

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \, \sigma_y}$$

X – Puntuaciones del test

Y – Puntuaciones del criterio

1. Identificar y definir el criterio y selección del método para medirlo.
2. Selección muestra representativa de sujetos.
3. Administración test.
4. Medida del criterio – mismo tiempo que el test: concurrente/decisión, después del test: predictiva.
5. Cálculo coeficiente.

Criterios valoración validez predictiva

Muñiz, J. (2005) Utilización de los tests. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.). *Análisis de los ítems*, (pp. 133-172). Madrid: La Muralla, S.A.

- Inadecuada: $r < 0.20$
- Suficiente: $0.20 \leq r < 0.35$
- Buena: $0.35 \leq r < 0.45$
- Muy buena: $0.45 \leq r < 0.55$
- Excelente: ≥ 0.55

Validez de Decisión: Sensibilidad y Especificidad

		Diagnòstic		
		Positiu	Negatiu	Total
Resultat del test	Positiu	Decisió correcta (f_{11})	Fals positiu (f_{12})	$f_{1.}$
	Negatiu	Fals negatiu (f_{21})	Decisió correcta (f_{22})	$f_{2.}$
Total		$f_{.1}$	$f_{.2}$	N

Índices de Validez de Decisión

$$proporción\ clasificaciones\ correctas = \frac{acuerdos\ resultados\ positivos + acuerdos\ resultados\ negativos}{N}$$

$$Sensibilidad = \frac{N^{o}\ personas\ con\ trastorno\ clasificadas\ con\ trastorno}{N^{o}\ total\ de\ personas\ con\ trastorno}$$

$$Especificidad = \frac{N^{o}\ personas\ sanas\ clasificadas\ como\ sanas}{N^{o}\ total\ de\ personas\ sanas}$$

Índice de Validez: Coeficiente Kappa

$$K = \frac{F_c - F_a}{N - F_a}$$

$$F_c = f_{11} + f_{22} \qquad F_a = \frac{f_{1.} \cdot f_{.1} + f_{2.} \cdot f_{.2}}{N}$$

Significación:

$$IC \Rightarrow k \pm z_\alpha \cdot \sigma_e \qquad \sigma_e = \sqrt{\frac{F_a}{N \cdot (N - F_a)}}$$

Valoración: Coeficiente Kappa

Altman, D.G. (1991). *Practical statistics for medical research*. New York: Chapman and Hall.

.00 - .20 – pobre

.21 - .40 – débil

.41 - .60 – moderada

.61 - .80 – buena

.81 - 1 – muy buena

Valores predictivos

- Estudio Retrospectivo

$$VPP = \frac{f_{11}}{(f_{11} + f_{12})}$$

$$VPN = \frac{f_{22}}{(f_{21} + f_{22})}$$

- Estudio prospectivo

$$VPP = \frac{P \cdot S}{P \cdot S + (1 - P) \cdot (1 - E)}$$

$$VPN = \frac{(1 - P) \cdot E}{(1 - P) \cdot E + P \cdot (1 - S)}$$

Exemple: Validesa de Decisió

		Depressió Post-part	
		No	Sí
Escala de Hamilton	Sí	50	125
	No	300	25

$$P_c = \frac{125+300}{500} = 0{,}85$$

$$S = \frac{125}{150} = 0.8\widehat{3}$$

$$E = \frac{300}{350} = 0.8571$$

$$K = \frac{425-280}{500-280} = 0{,}659$$

$$F_a = \frac{175 \cdot 150 + 325 \cdot 350}{500} = 280$$

$$VPP = \frac{0.3 \cdot 0.8\widehat{3}}{0.3 \cdot 0.8\widehat{3} + (1-0.3)\,(1-0.8571)} = 0.7142$$

$$VPN = \frac{(1-0.3)\,0.8571}{(1-0.3)\,0.8571 + 0.3\left(1-0.8\widehat{3}\right)} = 0.9231$$

Validez de Constructo

Trata de inferir el grado en que un sujeto tiene un rasgo o atributo medido por el test.

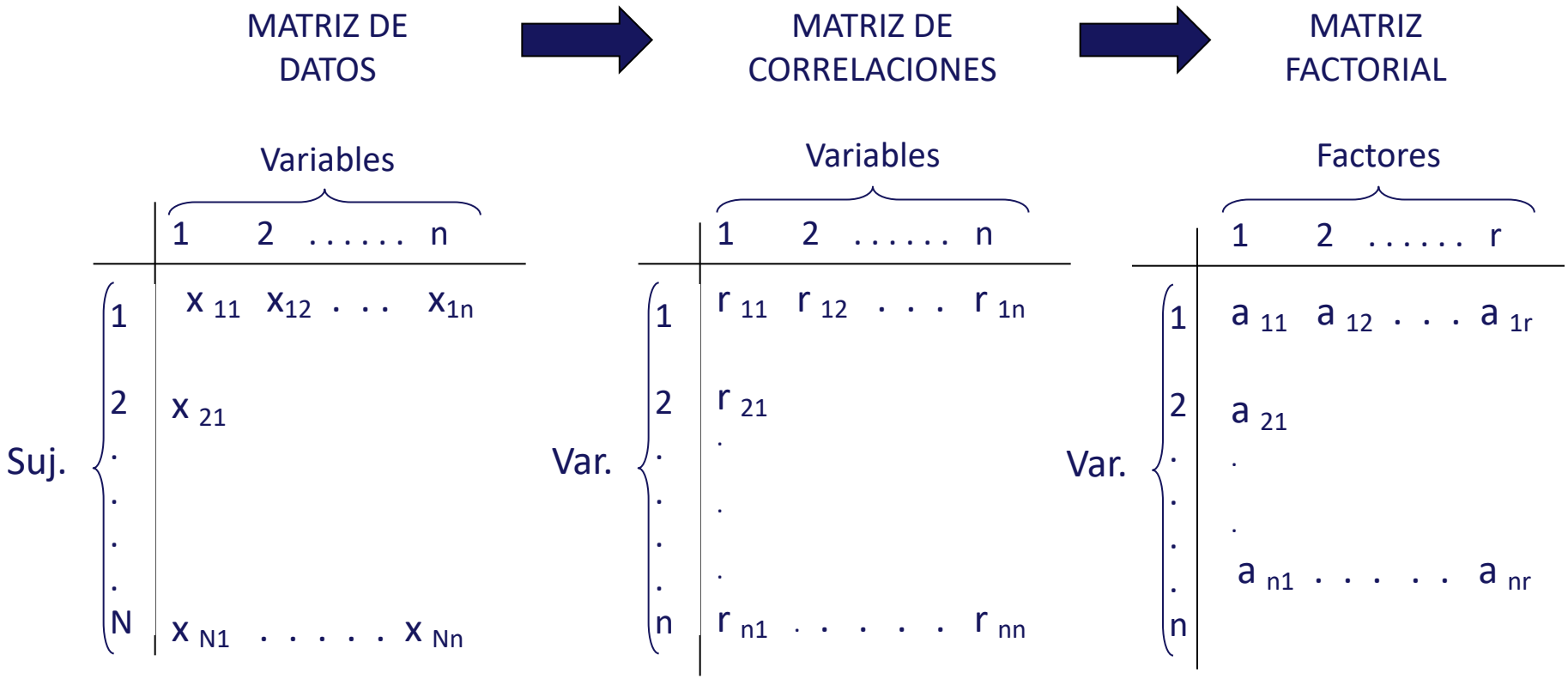
Se basa en evidencias múltiples y contingentes:

- Análisis factorial – identificación (AFE) o ***confirmación (AFC)*** de los rasgos psicológicos que mide el test
- Matriz multirasgo-multimétodo (Campbell y Fiske, 1959)
  - Validez convergente (mismo rasgo – diferente método)
  - Validez discriminante (diferente rasgo – mismo método)

Validez: Análisis Factorial

Técnica de reducción de datos:

n variables → r factores (r < n)



Validez: Análisis Factorial

Supongamos que construimos un test para medir tres factores de manera que:

F1: ítems 1 y 2 F2: ítems 3, 4 y 5 F3: ítems 6 y 7

Estructura matriz factorial:

AFE:

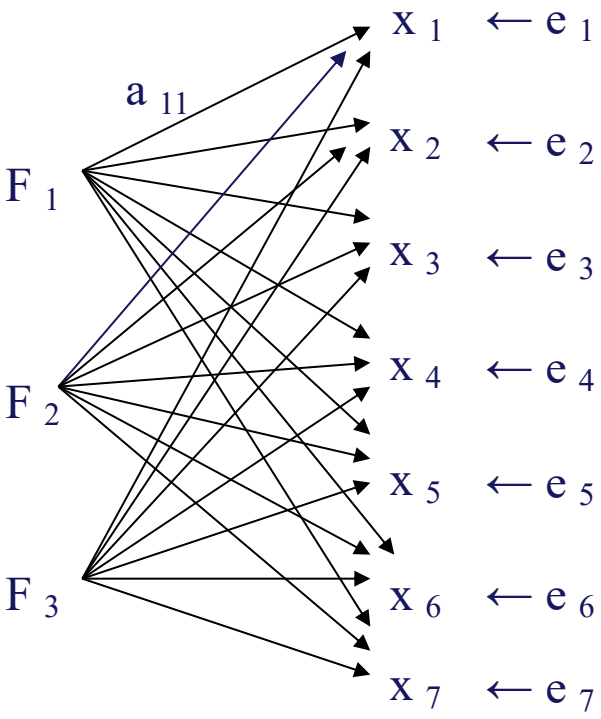
ítems	factores 1	2	3
1	a_{11}	a_{12}	a_{13}
2	a_{21}	a_{22}	a_{23}
3	a_{31}	a_{32}	a_{33}
4	a_{41}	a_{42}	a_{43}
5	a_{51}	a_{52}	a_{53}
6	a_{61}	a_{62}	a_{63}
7	a_{71}	a_{72}	a_{73}

AFC:

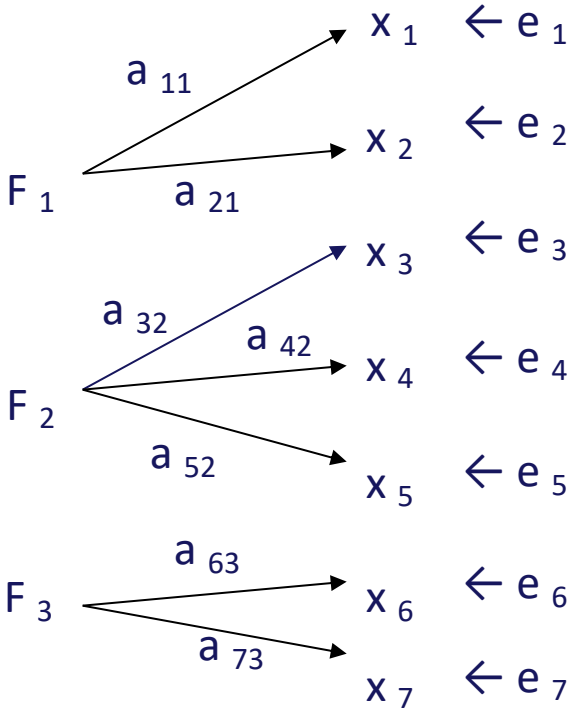
ítems	factores 1	2	3
1	a_{11}	0	0
2	a_{21}	0	0
3	0	a_{32}	0
4	0	a_{42}	0
5	0	a_{52}	0
6	0	0	a_{63}
7	0	0	a_{73}

Validez: Análisis Factorial – representación gráfica

AFE:



AFC:



Validez: matriz multirasgo-multimétodo

	A1	B1	C1	A2	B2	C2	A3	B3	C3
A1	(0.89)								
B1	0.51	(0.88)							
C1	0.38	0.37	(0.76)						
A2	….	….	….	(…..)					
B2	….	….	….	….	(….)				
C2	….	….	….	….	….	(….)			
A3	….	….	….	….	….	….	(….)		
B3	….	….	….	….	….	….	….	(….)	
C3	….	….	….	….	….	….	….	….	(….)

A, B i C – rasgos

1, 2 i 3 - métodos

(…) Valores monorasgo-monométodo (FIABILIDAD)

… Valores heterorasgo-monométodo (V. DISCRIMINANTE)

… Valors monotret-heteromètode (V. CONVERGENTE)

… Valores heterorasgo-heterométodo

Transformación de las Puntuaciones

Puntuación Directa → No aporta información sobre el rendimiento del sujeto

↓

TCT → Interpretación: comparando la puntuación obtenida con las puntuaciones de la población a la que pertenece el sujeto

↓

Inferir → Muestra representativa de la población

↓

Grupo normativo

Fase de un Estudio Normativo

1. Identificación de la población de interés.

2. Selección de una muestra representativa → Grupo normativo.

3. Recogida de los datos → Aplicación del test.

4. Transformación de puntuaciones directas en normativas.

5. Descripción del estudio normativo en el manual del test.

Tipo de Transformaciones

- Normas Cronológicas.
- Percentiles.
- Puntuaciones típicas.
- Puntuaciones típicas normalizadas.
- Puntuaciones típicas derivadas.

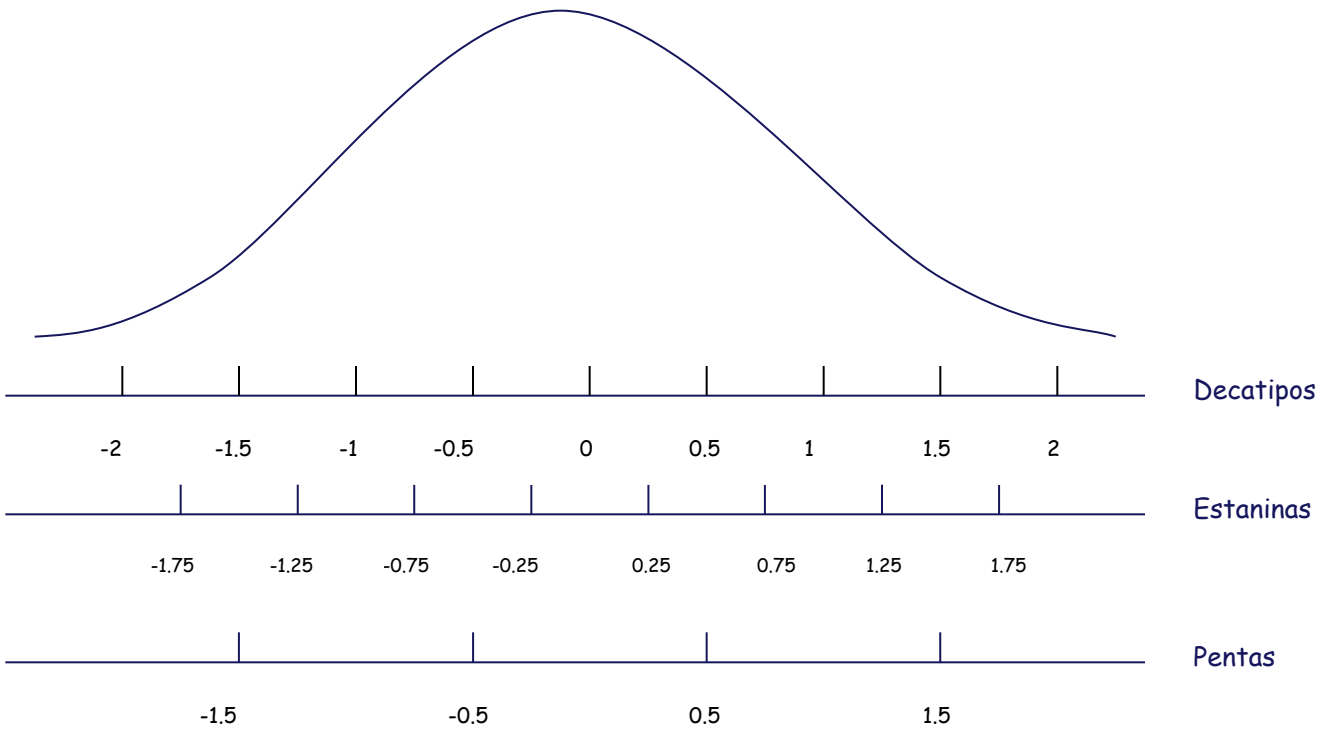
Tipos de Puntuaciones Típicas Derivadas

$$z_d = x_d + s_d \cdot z_x$$

- A cada z le corresponde una puntuación derivada:
  - Puntuaciones T $\rightarrow$ T = 50 +10 z (límites: 24 ÷ 76 - 99%)
  - Puntuaciones D $\rightarrow$ D = 50 + 20 z (límites: 0 ÷102 - 99%)
  - QI $\rightarrow$ QI = 100 + 15 z (límites: 61 ÷ 139 - 99%)

- A un rango de z le corresponde un valor:
  - Sten o decatipos (10) $\rightarrow$ Sten = 5.5 +2 z
  - Estaninas (9) $\rightarrow$ e = 5 + 2 z
  - Pentas (5) $\rightarrow$ p = 3 + z

Puntuaciones Típicas Derivadas



Construcción de un Test

1. Constructo que se quiere medir y teoría subyacente.- ¿Qué queremos medir?
2. Creación ítems - Indicadores observables de la conducta externa del sujeto.
3. Recogida de datos.
4. Análisis ítems.
5. Selección ítems.
6. Obtención del test normativo: Fiabilidad y Validez.
7. Selección del grupo normativo.
8. Transformación de las puntuaciones.

Fuentes más habituales de error en la adaptación de test

1. Contexto de utilització – sociocultural.

2. Construcción de la prueba – traducción inversa (comprobar fiabilidad, validez y estandardización).

3. Aplicación.

4. Interpretación de los resultados – equivalencia entre versiones.

Aspectos éticos y deontológicos: APA

Competencia.

Integridad.

Responsabilidad profesional y científica.

Respeto por los derechos y la dignidad de las personas.

Procurar el bienestar de las personas.

Responsabilidad social.

Utilidad evaluación.

Posibilidad de realització.

Adecuación.

Precisión.

Utilización de los test

Muñiz, J. (2005) Utilización de los test. En J. Muñiz, A.M. Fidalgo, E. García-Cueto, R. Martínez y R. Moreno (Eds.). *Análisis de los ítems*, (pp. 133-172). Madrid: La Muralla, S.A.

- Evaluación de la calidad del test
  - Descripción del test
  - Valoración de las características del test
  - Valoración global del test
- Directrices elaboradas por la Comisión Internacional de Tests (ITC) para el uso adecuado del test (Bartram, 2001) - Objetivo: mejorar el uso de los tests.
- http://www.efpa.be
- http://www.cop.es
- http://www.intestcom.org

Directrices ITC

1. Uso ético de los test
    1. Actuar de forma ética y profesional
    2. Asegurarse que son competentes para el uso de los test
    3. Responsabilizarse del uso que hacen de los test
    4. Asegurarse que los materiales del test están seguros
    5. Asegurarse que los resultados de los test se tratan confidencialmente
2. Utilización adecuada de los test
    1. Determinar la utilidad potencial de los test en una situación evaluativa
    2. Escoger test técnicamente correctos y adecuados a la situación
    3. Prestar atención a los aspectos relacionados con el sesgo del test
    4. Hacer los preparativos necesarios para la aplicación del test
    5. Aplicar los test adecuadamente
    6. Puntuar y analizar los resultados de los test con precisión
    7. Interpretar los resultados adecuadamente
    8. Comunicar los resultados de forma clara y precisa
    9. Revisión de la adecuación del test y de su uso

Inconvenientes de la TCT

- Dependencia de un grupo normativo.
- Existen tantos coeficientes diferentes (fiabilidad, validez, etc.) como aplicaciones del test se hagan a diferentes grupos.
- Dificultad de la comparación de puntuaciones en diferentes test.
- Dificultad en conseguir la fiabilidad de formas paralelas.
- Supuesto débil de la igualdad del error típico de medida constante para un mismo test en diferentes muestras $\rightarrow$ Estimación puntuaciones verdaderas y comparaciones incorrectas.

TEST REFERIDOS AL CRITERIO

Test norma de grupo respecto a test referidos a criterio

- Test de norma de grupo: discriminar entre sujetos maximizando su variabilidad.

- Test referidos al criterio: grado en que los sujetos dominan un campo educativo o profesional, la discriminación intersujeto pasa a un segundo plano.

TEST REFERIDOS AL CRITERIO

SE UTILIZAN PARA EVALUAR EL ESTATUS ABSOLUTO DEL SUJETO CON RESPECTO A ALGÚN DOMINIO DE CONDUCTAS BIEN DEFINIDO.

SE CONSTRUYEN PARA PERMITIR LA INTERPRETACION DE TESTS INDIVIDUALES Y DE GRUPO CON RELACION A UN CONJUNTO DE OBJETIVOS, DESTREZAS Y COMPETENCIAS, BIEN DEFINIDO.

ANALISIS DE LOS ITEMS

VALIDEZ DE CONTENIDO DE LOS ITEMS

SELECCIÓN GRUPOS CRITERIO
MEDIDAS PRE/POST INSTRUCCION
GRUPOS INSTRUIDOS/NO INSTRUIDOS
GRUPOS DE CONTRASTE (CASO/CONTROL)

ESTADÍSTICOS DE LOS ITEMS
ÍNDICE DE DIFICULTAD
INDICES DE DISCRIMINACION (Índices de Sensibilidad a la Instrucción)
ÍNDICES DE HOMOGENEIDAD

DETERMINACION LONGITUD DEL TEST
Fijación del nivel máximo de error (precisión)

FIABILIDAD DE LOS TRC

FIABILIDAD *ESTIMACION PUNTUACION* DOMINIO

FIABILIDAD DE LAS *CLASIFICACIONES*:

ÍNDICES DE ACUERDO CON DOS APLICACIONES

- Índice de Hambleton y Novick
- Índice Kappa de Cohen
- P^* de Crocker y Algina

ÍNDICES DE ACUERDO CON UNA SOLO APLICACION

- Método de Huyhn
- Método de Subkoviak

VALIDEZ DE LOS TRC

VALIDEZ DE CONTENIDO / CRITERIO / CONSTRUCTO

VALIDEZ DE LAS DECISIONES DE CLASIFICACION

INDICES DE ACUERDO (Fiabilidad)

INDICE DE SENSIBILIDAD
Proporción de Sujetos CON el Trastorno detectados como tales con el Test

INDICE DE ESPECIFICIDAD
Proporción de Sujetos SIN el Trastorno detectados como tales con el Test

ESTABLECIMIENTO DE PUNTOS DE CORTE (ESTANDARES)

PUNTOS DE CORTE
APTO/NO APTO; PASA/FALLA; NORMAL/CASO

PROCEDIMIENTOS BASADOS EN JUICIOS SOBRE EL CONTENIDO DEL TEST

- Método de Nedelsky
- Método de Angoff
- Método de Ebel
- Método de Jaeger

PROCEDIMIENTOS BASADOS EN EL RENDIMIENTO DE GRUPOS DE VALIDACION

Método del grupo límite

Método de los grupos de contraste

TEORIA DE RESPUESTA A LOS ÍTEMS (TRI)

Diferencias TCT y TRI

- Unidad de análisis – TCT: puntuación observada en el test versus TRI: ítem.

- La TRI incorpora términos al modelo que describen las características de los ítems.

- Los supuestos son radicalmente diferentes.

Teoría de respuesta a los ítems (TRI)
Supuestos básicos

- Dimensionalidad del espacio latente:

  $\theta = (\theta_1, \theta_2, \ldots, \theta_j, \ldots, \theta_k)$

  θ : Variable latente que se desea estimar: $-\infty \leq \theta \leq \infty$

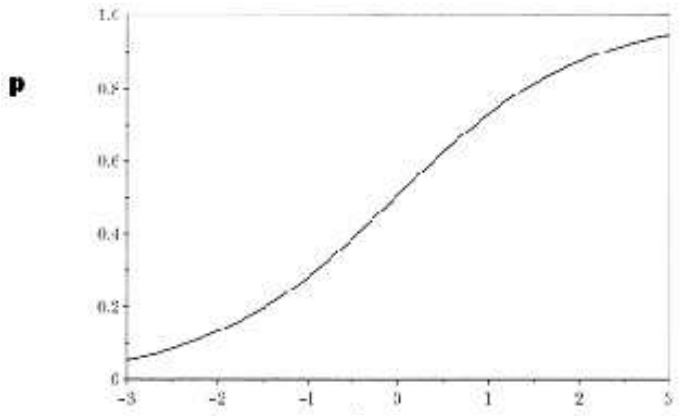
  θj: Variable aleatoria

- Independencia local:

$$f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{i=n} f_i(x_i \mid \theta)$$

Teoría de respuesta a los ítems (TRI)

Curba característica del ítem (C.C.I.): función matemática (generalmente logística) que relaciona la probabilidad de responder correctamente a un ítem con el nivel de habilidad (θ) que tiene el sujeto que ha respondido el ítem en la variable medida por el ítem.



$$y = \frac{e^x}{1 + e^x}$$

e: 2.718

X: cualquier valor o función

TRI: Parámetros del modelo (estimación: ML)

- Nivel de habilidad: θ

  Escala de intervalo

  Rango teórico: $-\infty$ a $+\infty$, lo más habitual es estandarizar la escala con media 0 y desviación estándar 1

- Dificultad del ítem: b

  Mismo rango que θ

  Su valor es aproximadamente el punto en la escala de θ al que corresponde una probabilidad de acertar el ítem de (1 + c)/2

  Cuanto más grande sea b más difícil es el ítem

- Discriminación del ítem: a

  Su valor es proporcional a la pendiente de la CCI en el punto en que $\theta = b$

- Probabilidad de acertar por azar el ítem: c

  Técnicamente: valor de la asíntota inferior de la CCI (probabilidad da acertar el ítem cuando $\theta = -\infty$

TRI: modelos

ML3p $$p_i(\theta) = c_i + (1 - c_i)\frac{e^{D\,a_i\,(\theta - b_i)}}{1 + e^{D\,a_i\,(\theta - b_i)}}$$

$p_i(\theta)$: probabilidad de acertar el ítem i para un valor de θ

b_i: índice de dificultad del ítem i

a_i: índice de discriminación del ítem i

c_i: índice de pseudoazar del ítem i

D: constante (si D = 1.7, la logística se aproxima a la normal)

ML2p: no hay aciertos por azar $$p_i(\theta) = \frac{e^{D\,a_i\,(\theta - b_i)}}{1 + e^{D\,a_i\,(\theta - b_i)}}$$

ML1p: a es igual para todos los ítems $$p_i(\theta) = \frac{e^{D\,(\theta - b_i)}}{1 + e^{D\,(\theta - b_i)}}$$

TRI: puntuaciones verdaderas

La curba característica del test permite transformar las puntuaciones θ a una nueva escala: las puntuaciones verdaderas (de 0 a n):

$$PV_j = \sum_{i=1}^{n} p_i(\theta_j)$$

Donde:

PV_j: puntuación verdadera que corresponde a la persona con un nivel en el rasgo latente de θ_j

n: número de ítems

p_i (θ_j): valor correspondiente a cada CCI para $\theta = \theta_j$

TRI: Función de información

Información (Fisher): recíproco de la precisión con que se puede estimar un parámetro. Según esta definición la información que tiene un ítem determinado al estimar un valor concreto de θ es:

$$I(\theta) = \frac{1}{\sigma^2_{(\hat{\theta}|\theta)}}$$

Error típico de medida: $$\sigma_{(\hat{\theta}|\theta)} = \sqrt{\sigma^2_{(\hat{\theta}|\theta)}}$$

Estimación por intervalo: $$\hat{\theta} \pm z_{\alpha/2} \cdot \sigma_{(\hat{\theta}|\theta)}$$

TRI: Función de información del ítem

ML1p: $I_i(\theta) = D^2 \; p_i(\theta) q_i(\theta)$

Donde:

$I_i(\theta)$ – cantidad de información del ítem i en el nivel θ

D – constante de escalamiento: 1.7

$p_i(\theta)$ – probabilidad de acertar el ítem

$q_i(\theta) = 1 - p_i(\theta)$

ML2p: $I_i(\theta) = D^2 \; a_i^2 \; p_i(\theta) q_i(\theta)$

Donde: a_i – parámetro de discriminación del ítem i

ML3p: $I_i(\theta) = \dfrac{D^2 \; a_i^2 \; q_i(\theta) \left[p_i(\theta) - c_i \right]^2}{p_i(\theta)\left(1 - c_i\right)}$

Donde: c_i – parámetro de pseudoazar del ítem i

TRI: Cantidad de información máxima

ML1p i ML2p: está en el punto donde $\theta = b$

$$IM_i = \frac{D^2\ a_i^2}{4}$$

ML3p: viene dada por $\theta = b_i + \left[1/\left(D\ a_i\right)\right]\left\{\ln\left[1/2 + 1/2\ \sqrt{1 + 8\ c_i}\right]\right\}$

$$IM_i = \frac{D^2\ a_i^2}{8\left(1 - c_i\right)^2}\left[1 - 20\ c_i - 8\ c_i^2 + \left(1 + 8\ c_1\right)^{3/2}\right]$$

TRI: Función de información del test

$$IT(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Influida por:

- Calidad de los ítems – más información como más discriminativos sean los ítems

- Número de ítems que forman el test – más ítems más información

Pasos para construir un test en el marco de la TRI

1. Tener buena comprensión del constructo a medir.
2. Redactar como mínimo el doble de ítems de los definitius.
3. Estudio piloto (muestra de 50 a 100): mala redacción, TCT (distractores, discriminación, etc.).
4. Administrar ítems seleccionados a una muestra de calibración.
5. Selección inicial de los ítems aplicando los indicadores clásicos.
6. Analizar la dimensionalidad de los datos.
7. Estimar los parámetros de los ítems aplicando el modelo de la TRI y valorar el ajuste del modelo.
8. Especificar la función de información del test objetivo.
9. Seleccionar los ítems que cumplen con la función de información objetivo.
10. Calcular la función de información del test según se vayan añadiendo ítems hasta obtener una aproximación razonable a la función objetivo.

Ventajas TRI respecto a la TCT

- Si se cumplen los supuestos del modelo, el modelo seleccionado es el apropiado y se calibra correctamente, entonces se obtendrá el mismo valor de los parámetros de los ítems independientemente de la muestra de calibración utilizada.
- TRI: la estimació de θ no depdende del test utilizado, igual que la invariancia de los ítems.
- TRI: permite estimar la precisión con que cada ítem y cada test mide los diferentes niveles de habilidad (no asume el supuesto de igualdad de errores de medida de la TCT).
- Desde la TRI se pueden construir instrumentos de evaluación personalizados más eficientes (número mínimo de ítems).
- TCT: opción de preferencia si el tamaño de muestra es pequeño para ajustar un modelo de TRI, no se cumplen los supuestos TRI o en adaptaciones de test.

TEST ADAPTATIVOS COMPUTERIZADOS (TAI)

Test Adapatativos Computerizados (TAI)

Objetivo: aplicar solo aquellos ítems que aporten información máxima para el nivel de θ evaluado.

Pasos:

1. Dado el nivel de habilidad estimado para la persona establecer qué ítem del banco será el próximo para presentarle.
2. Administrar ítem.
3. Estimar nuevamente el nivel de habilidad.
4. Repetir pasos 1 a 3 hasta que se cumpla algún criterio.

Funcionamento diferencial de los ítems

Un ítem funciona diferencialmente (DIF) si la probabilidad de responder correctamente al ítem es función no solo de θ sino también de cualquier otra característica irrelevante.

TRI:
- Contraste de las diferencias en b
- Comparación de modelos

Sin necesidad de ajustar un modelo de TRI:
- Mantel Haenszel – No exige un modelo de medida específico.

DIF: contraste de las diferencias en b

Obtención de las CCI para cada grupo: si coinciden no hay DIF y si no coinciden hay DIF

ML1p: hay DIF si hay diferencias en b entre los grupos

H_0: $b_F = b_R$

H_1: $b_F \neq b_R$

$$D = \frac{\Delta b}{S_{\Delta b}}$$

Error típico de la diferencia entre los parámetros b:

$$S_{\Delta b} = \sqrt{S_{bF}^2 + S_{bR}^2}$$

Si $|D| \geq |z_{\alpha/2}|$ hay DIF

DIF: comparación de modeles

Comparación del modelo compacto (todos los parámetros son iguales para los dos grupos) con el modelo aumentado (algun/os parámetro/s difiere/n entre los grupos).

Razón de verosimilitud:

$$LR = -2\ln\left[L(C)/L(A)\right] = \left[-2\ln L(C)\right] - \left[-2\ln L(A)\right]$$

Donde:

L(C) – verosimilitud modelo compacte

L(A) – verosimilitud modelo aumentado

LR sigue la distribución χ^2 con g.l. igual a la diferencia en el número de parámetros estimados

DIF: Matel-Haenszel

Construcción tablas 2 x 2 para los m intervalos en que se divide la puntuación total del test (1…k…m).

Nivel k	Acierto (1)	Error (0)	Total
Referencia	A_k	B_k	N_{Rk}
Focal	C_k	D_k	N_{Fk}
Total	N_{1k}	N_{0k}	N_k

$$\chi^2 = \frac{\left(\left| \sum_{k=1}^{m} A_k - \sum_{k=1}^{m} \frac{N_{Rk}\, N_{1k}}{N_k} \right| - 0.5 \right)^2}{\sum_{k=1}^{m} \frac{N_{Rk}\, N_{Fk}\, N_{1k}\, N_{0k}}{N_k^2 \left(N_k - 1 \right)}}$$

χ^2 *con* 1 *g.l.*

Magnitud del DIF :

$$\hat{\alpha}_{MH} = \frac{\sum_{k=1}^{m} \frac{A_k\, D_k}{N_k}}{\sum_{k=1}^{m} \frac{B_k\, C_k}{N_k}}$$

Medida de las actitudes

- Modelo de Thurstone (variabilidad perceptual de los sujectos).
  - Ley del juicio comparativo
  - Ley del juicio categórico:
    - Método de los intervalos sucesivos
    - Método de los intervalos aparentemente idénticos
    - Método de ordenación por rangos.
- Técnica *Likert*.- Método de escalamiento centrado en los sujetos.
- Escalograma de Guttman.- Modelo determinista de escalamiento (ordenación de estímulos y sujetos en una dimensión).