I) ORIENTACIONES DIDÁCTICAS

- Los distintos procedimientos de hallas la fiabilidad de los test anteriores se denominan: Test referidos a normas: el rendimiento de los sujetos se evalúa en referencia a otros sujetos que forman el grupo normativo. Estos solo proporcionan información del sujeto de su posición relativa con el resto del grupo.
- Test referidos a criterio: su evaluación tiene lugar en función del número de objetivos logrados en el test.

II) DEFINICIÓN Y OBJETIVOS DE LOS TESTS REFERIDOS AL CRITERIO

- Orígenes; Flanagan y Nedelsky: Introdujeron el concepto de estándar relativo y absoluto respecto a las puntuaciones en los tests.
- Ebel: a él se le debe la denominación de Test Referido a Criterio.
- Glaser: establece la diferenciación con los tests normativos.
- Hambleton: causas que generan su aparición:
  - Necesidad de conocer la eficacia de los programas educativos
  - Interés por evaluar el nivel de habilidad alcanzado
  - Clima contrario al uso de tests que caracteriza a la sociedad americana de los 60
- Popham definición mas aceptada: un Test Referido a Criterios se utiliza para evaluar el status absoluto del sujeto con respecto a algún dominio de conductas bien definido.

III) DIFERENCIAS ENTRE LOS TESTS REFERIDOS A LA NORMA Y LOS TESTS REFERIDOS AL CRITERIO

	TRC Tests referido a criterio	TRN Test referido a norma
CONSTRUCCIÓN DE UN TEST	• Se delimita el dominio de contenido o conductas y el uso del test • Atención a las especificaciones del contenido y a la elaboración y análisis cualitativo de los ítems • Validez de contenido: fundamental: relevancia y representatividad de los ítems respecto al dominio específico •	• Los ítems se derivan se teorías de rasgos
CRITERIOS DE SELECCIÓN DE ÍTEMS	• se basa en los objetivos: ▪ Estimación de la puntuación dominio de los sujetos Test referido al dominio ▪ Establecimiento de estándares mediante puntos de corte Test de maestría: clasificación de sujetos	• Maximizar las diferencias individuales (tests de dificultad media y alto índice de discriminación del test)
objetivo	• Dependiendo del objetivo, la estimación de la fiabilidad de las puntuaciones se realiza de forma diferente.	• No son apropiados porque no permiten describir la precisión de las puntuaciones individuales ni la consistencia de las decisiones,.
EVALUACIÓN: en el significado y interpretación de las puntuaciones d los test.	• La puntuación representa el estimador muestral del rendimiento en el dominio y su significado es en términos absolutos • Para la estimación de la puntuación en el dominio se puede utilizar la proporción de respuestas correctas	• La puntuación obtenida es un indicador de su puntuación verdadera en un rasgo latente y sólo tiene significado en relación al grupo normativo

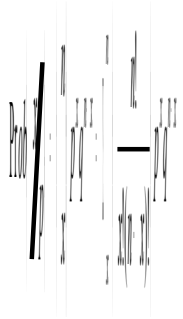
IV) LONGITUD DEL TEST

- Es importante **determinar de la longitud del test** (número de ítems) ya que de ello depende la **utilidad de las puntuaciones** obtenidas
  - Si el número de ítems es **pequeño**: la **interpretación es limitada:** clasificación poco fiable.
  - Si el número de ítems **grande** se pueden asegurar valores de probabilidad de clasificación incorrecta mínimos.
- **Formas de reducir el número de errores sin tener que aumentar la longitud del test:**
  1. **Modelo bayesianos**
  2. **Método basados en tests computarizados**
  3. **Modelo Millman**

Modelo de Millman:

- Basado en el **modelo binomial**: considera la proporción esperada de ítems que un sujeto pueda responder correctamente para ser considerado apto y el error máximo tolerable
- **Supuestos**:

1. Test compuesto por una **muestra aleatoria de ítems dicotómicos**
2. **La probabilidad de una respuesta correcta es constante** para todos los ítems del test
3. Las **respuestas son independientes** entre sí
4. Los **errores se ajustan al modelo binomial**



$\Pr ob\left(x/p\right)$: probabilidad de que un sujeto con una **puntuación p,** conteste correctamente **x ítems** de un test que tiene **n ítems**

- Calcular la longitud de un test supuesta una determinada proporción de aciertos.

$$n = \frac{p_c(1-p_c)}{e^2}$$

- n : número de ítems del test
- p_c : proporción de aciertos correctos para considerarse apto
- e :error máximo admisible

V) FIABILIDAD EN LAS CLASIFICACIONES EN LOS TESTS REFERIDOS A CRITERIO

- Desde el enfoque de los puntos de corte (tests de maestría) **test fiable** si, tras la aplicación a los mismos sujetos en distintas ocasiones o de dos formas paralelas, se clasifican siempre en la misma categoría.
  - Métodos basados en dos aplicaciones del test: índice de Hambleton y Novick; el coeficiente Kappa de Cohen y el índice de Croker y Algina
  - Métodos basados en una sola aplicación del test: método de Huynh; el método de Subkoviak y el coeficiente de Livingston.

A) ÍNDICES DE ACUERDO QUE REQUIEREN DOS APLICACIONES DEL TEST

1) Índice de Hambleton y Novick

- Utilización de la proporción de sujetos que consistentemente son clasificados dentro del grupo **de maestría o no maestría.**
- La proporción de sujetos consistentemente clasificados en ambos tests se expresa:

$$p_c = \sum_{i=1}^{n} p_i = \frac{n_{11}}{N} + \frac{n_{22}}{N} + ... + \frac{n_{mm}}{N}$$

  - : proporción de sujetos clasificados en ambas formas
  - : número de sujetos
  - : número de sujetos en cada casilla en los que ambos tests coinciden al clasificarlos

- El valor máximo de es igual a 1, que se obtiene cuando los sujetos son clasificados de la misma forma con los 2 tests
- El valor mínimo: es igual a la **proporción** de clasificaciones consistentes que se espera por azar , valor que viene dado en función de las frecuencias marginales de la tabla

$$p_a = \sum_{j=1}^{m} \frac{N_j N_j}{N^2}$$

2) Coeficiente Kappa de Cohen

- Este coeficiente elimina del valor de la **proporción** de sujetos clasificados consistentemente el valor de la proporción de clasificación consistente esperada por azar

$$k = \frac{p_c - p_a}{1 - p_a}$$

- Este **coeficiente proporciona una medida de la consistencia de clasificación** de los sujetos independientemente del posible valor esperado por azar.
- **Este valor oscila**:
  - Entre 1: fiabilidad perfecta.
  - 0: atribuida al azar.
- Este coeficiente puede expresarse en función de las **frecuencias absolutas**:

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

  - : frecuencia observada de clasificaciones coincidentes
  - : frecuencia de coincidentes esperadas por azar
  - : número total de sujetos

- **Error típico de medida de K**

$$S_e = \sqrt{\frac{F_a}{N(N - F_a)}}$$

- Después se calcula el intervalo confidencial:

$$k \pm Z_x * S_e$$

3) Índice de Croker y Algina

- El índice γ es una **alternativa al coeficiente Kappa**
- Se basa en que la probabilidad mínima de una decisión consistente es **0.50**
- Tiene lugar si las puntuaciones del test son estadísticamente independientes y el punto de corte está en la mediana de la distribución conjunta de las puntuaciones obtenidas en las dos aplicaciones.

$$\gamma = \frac{p_c - 0.50}{1 - 0.50} = 2p_c - 1$$

- γ = 1 cuando las decisiones son **totalmente consistentes**.
- γ = 0 cuando las decisiones no son más consistentes que las que resultarían de usar tests estadísticamente **independientes**, cuyas puntuaciones presentan la misma distribución y un punto de corte igual a la mediana de la distribución común

B) ÍNDICES DE ACUERDO QUE REQUIEREN UNA SOLA APLICACIÓN DEL TEST

1) Método de Huynh

- Un solo test y una sola aplicación: procedimiento matemático sofisticado para estimar la consistencia de clasificación.
- Método para pronosticar las puntuaciones de en un test "B" conocidas las puntuaciones de una muestra de sujetos en una aplicación (test "A")
- Este método presupone que la distribución de **puntuaciones** es aproximadamente **normal** y es adecuado cuando el **número de ítems es superior a 8** y la **razón entre la media de las puntuaciones** de los sujetos en el test y el número de ítems oscila entre **0,15 – 0,85**
- **Pasos:**

a. Calcular la **media** $\bar{X}$, la **varianza** S_x^2, el **coeficiente de correlación** r_{KR21} y especificar el valor **del punto de corte** (c)

b. Calcular la **puntuación típica** z_c correspondiente al valor del punto de corte, con una corrección de 0,5 y se acude a las tablas de curva normal para buscar el valor p_c que deja por debajo la z_c obtenida

$$Z_x = \frac{(C - 0.5 - \bar{X})}{S_x}$$

c. A partir de las **tablas de Gupta** (incluidas al final del libro) se obtiene la probabilidad P_{zz} de que dos variables distribuidas normalmente con 1 correlación r_{KR21} sean menores que z_c

d. Se calculan los valores p_z y p_{zz}

$$p_c = 1 - 2(p_{zz} - p_z) \quad ; \quad k = \frac{p_{zz} - p_z^2}{p_z - p_z^2}$$

2) Método de Subkoviak

Procedimiento con una sola aplicación cuando **no es posible establecer una forma paralela** de un test, por lo que **simula** las puntuaciones de una segunda forma paralela al test. Buena estimación valores p_c y p_a :

$$K = \frac{p_c - p_a}{1 - p_a}$$

3) Coeficiente de Livingston

- Se desarrolla dentro del contexto del TCT
- Todos los métodos que se han estudiado para el calculo de la fiabilidad consideran por igual tanto los errores que cometamos cuando clasificamos a un sujeto perteneciente al grupo de maestría en el grupo de no – maestría o al revés. Este coeficiente si tiene en cuenta este tipo de errores: **considera que son más importantes los errores de clasificación de los sujetos más distanciados del p.c. que los que están más cerca (mas fácil cometer errores)**

$$K_{xv}^2 = \frac{\alpha S_x^2 + (\bar{X} - C)^2}{S_x^2 + (\bar{X} - C)^2}$$

- α = coeficiente alfa
- S_x^2 = varianza del test
- C = punto de corte
-

	▪ $\bar{X}$ = media del test ▯ A medida que el p.c. se distancia del valor de la medida del test, aumenta el valor de K^2_{xv} . ▯ Cuando $\bar{X}$ (media del test) = X_c (punto de corte): K^2_{xv} = 0 ▯ Cuando X_c =1; K^2_{xv} =1

Preguntas de examen

06/PN

▯ El coeficiente kappa:
- a puede ser mayor que la unidad
- b es un estimador de consistencia de las clasificaciones
- c representa las clasificaciones realizadas al azar.

05/PN

▯ El índice P* de Croker y Algina (1986) considera que la probabilidad mínima de una decisión consistente es:
- a 0,25
- b 0,50
- c 0,75

▯ PROBLEMA: Dos test que miden un mismo trastorno de personalidad han clasificado a los sujetos de la siguiente forma (0 significa no trastorno y 1 trastorno) (NC 95%)

		TEST B		
		1	0	
TEST A	0	3	12	
	1	9	1	

- a el coeficiente kappa de Cohen es estadísticamente significativo
- b el I. C. para kappa es 0,53 – 1
- c la frecuencia esperada por azar es 10,40

		TEST B		
		1	0	
TEST A	0	3	12	15
	1	9	1	10
		12	13	25

$$k=\frac{F_c-F_a}{N-F_a} = \frac{21-12.4}{25-12.4} =0.68$$

$N=25$

$F_c=12+9=21$

$$Fa=\frac{15*12}{25}+\frac{10*13}{25}=12.4$$ la opción b ya no se cumple

$k\pm Z_x*S_e = 0.68\pm1.96*0.2=$ 1.072 y 0.288 tampoco se cumple la c

$$S_e=\sqrt{\frac{F_a}{N(N-F_a)}} = \sqrt{\frac{12.4}{25(25-12.4)}}= 0.2$$

$Z_x=1.96$

04/PN

▯ Los tests referidos a criterio:
- a combinan las puntuaciones del test y del criterio
- b sólo tienen validez predictiva o relativa al criterio
- c no requieren de un grupo normativo.

▯ Los tests referidos a criterio:
- a son útiles para calcular los percentiles de los sujetos en la variable medida
- b se utilizan para establecer estándares de rendimiento en dominios de interés
- c utilizan los mismos métodos que los tests formativos para estimar la fiabilidad.

Otras preguntas:

▯ Los tests referidos al criterio (TRC) enfatizan :
- a el rasgo o constructo subyacente
- b la especificación del dominio de contenidos
- c las diferencias individuales entre los sujetos.

▯ Los TRC:
- a combinan las puntuaciones del test y del criterio
- b sólo tienen validez predictiva o referida al criterio;
- c no requieren la utilización de un grupo normativo.

▯ Los TRC:
- a son útiles para calcular los percentiles de los sujetos en la variable medida

b se utilizan para establecer estándares de rendimiento en dominios de interés
c utilizan los mismos métodos que los tests normativos para estimar la fiabilidad.

- Un método para determinar la fiabilidad de las clasificaciones que requiere dos aplicaciones del test es el propuesto por:
  a Livingston
  b Cohen
  c Huynh.
- En los TRC resulta crucial la determinación de:
  a los baremos interpretativos
  b la longitud del test
  c las formas paralelas del test.
- En los TRC uno de los coeficientes más utilizados para el estudio de la fiabilidad es:
  a Rulon
  b Kappa
  c Spearman - Brown.
- En los TRC, el procedimiento de Millman se basa en el modelo:
  a Bayesiano
  b Logístico
  c binomial.

- En los TRC el cálculo de la fiabilidad hace referencia ante todo a la:
  a homogeneidad interna
  b fiabilidad de las clasificaciones
  c estabilidad temporal.
- Cuando en los TRC clasificamos erróneamente a un sujeto dentro del grupo de maestría, cometemos un error de:
  a Falso – positivo
  b Falso – negativo
  c Verdadero - negativo.
- En los TRC un estimador de la consistencia de la clasificación de los sujetos es el coeficiente de:
  a Kappa
  b Alfa
  c Beta .
- Un método para determinar la fiabilidad de las clasificaciones que requiere una sola aplicación del test es el propuesto por:
  a Subkoviak
  b Cohen
  c Croker y Algina.
- El índice *P** de Croker y Algina (1986) considera que la probabilidad mínima de una decisión consistente es:
  a 0.25
  b 0.50
  c 0.75.