

# Certificación de la Competencia Comunicativa

*Perspectiva psicométrica*

Paula Elosua

*VII Jornadas GIELE*

2024-11-16

1

Elosua, 2024



## Familiarizarse con los Principios **Psicométricos** aplicados a la Construcción de Pruebas de Certificación de la Competencia Comunicativa



### Contenido

#### Capítulo 0

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Capítulo 8

Capítulo 9

#### Prefacio: medir y evaluar

Pruebas de Certificación

Planificación

Tabla de especificaciones

Construcción de ítems

Revisión - Pretest

Standard Setting

Evaluación mediatizada por jueces

Equiparación

Epílogo

4

Elosua, 2024



### ¿Medir?



The outfit (Graham Moore, 2022)

5

Elosua, 2024



## Para quién lo haces...

## Lenguas

### Competencia comunicativa

#### Evaluación

“..actividad compleja que integra conocimiento, juicio clínico, información adicional fiable, y medidas psicométricas...proceso de resolución de problemas que hace uso de información relevante sobre personas, grupos o instituciones con el fin de tomar decisiones y ofrecer recomendaciones.” (APA, 2020)

“El término «evaluación» se utiliza en este capítulo con el sentido concreto de valoración del grado de dominio lingüístico que tiene el usuario.”(Consejo de Europa, 2002)

#### Evaluación de la competencia comunicativa

## Evaluación: clasificación

Criterio	Tipos
Propósito	Formativa / Sumativa / Diagnóstica
Interpretación	Normativa / Criterial
Contenido	Teórica / Práctica
Impacto	Bajo Riesgo /Alto Riesgo

## Certificación de la competencia comunicativa

<b>Certificación de la competencia comunicativa</b>	
<b>Sumativa</b>	Nivel final -alcanzado-
<b>Criterial</b>	Compara con estándares - Marco Común Europeo de Referencia (MCER).
<b>Práctica</b>	Medición próxima a situaciones reales ( <b>auténtica</b> )
<b>Alto Riesgo</b>	Consecuencias

# Certificación de la competencia comunicativa

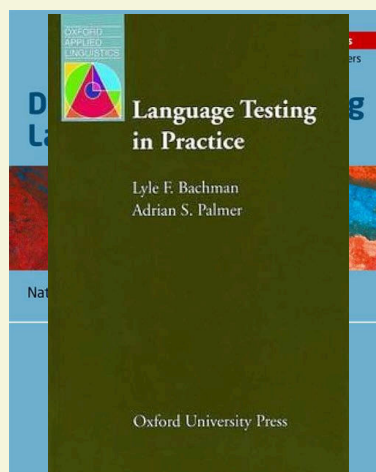
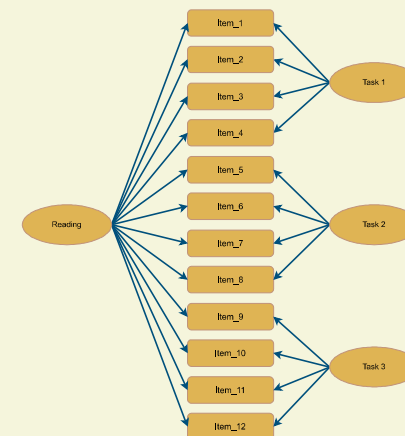
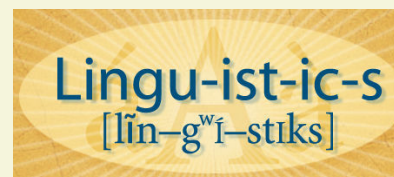
## Evaluación

- Alto-riesgo
- Criterial
- Sumativa
- Práctica

## Exigencias mínimas

- Fiable (consistencia)
- Valida (objetivos)
- No sesgada (no discriminatoria)

VALIDEZ Y EQUIDAD EN LA EVALUACION



## Contenido

Capítulo 0

**Capítulo 1**

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Capítulo 8

Capítulo 9

Prefacio: medir y evaluar

**Pruebas de Certificación**

Planificación

Tabla de especificaciones

Construcción de ítems

Revisión - Pretest

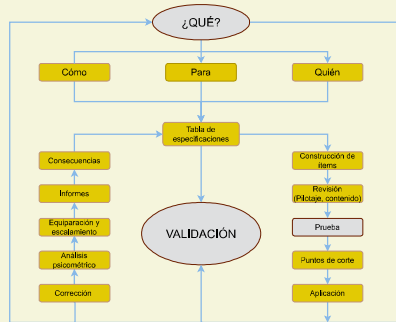
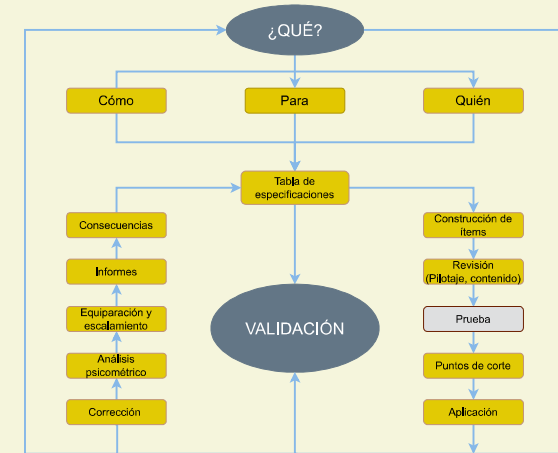
Standard Setting

Evaluación mediatizada por jueces

Equiparación

Epílogo

# ¿Cómo abordamos el diseño de las pruebas?

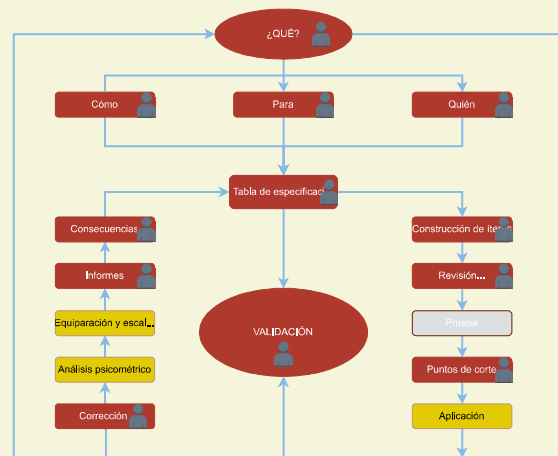


“Test and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee populations.” (AERA et al., 2014, p. 85)

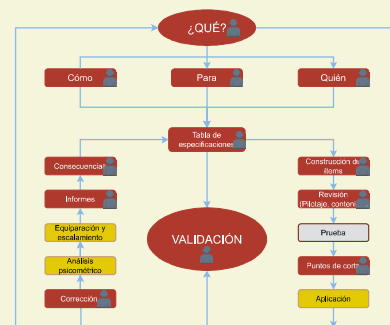
▶ 0:00 / 0:24



# ¿Dónde están los especialistas en lenguas?



# SUBJECT MATTER EXPERTS + PSICOMETRÍA



- Planificación (diseño)
- Tabla de especificaciones
- Construcción de ítems
- Pilotaje, contenido (*DIF*)
- Puntos de corte: *Standards setting*
- Corrección
- Estudios técnicos
- Informes

## Contenido

- Capítulo 0
- Capítulo 1
- Capítulo 2**
- Capítulo 3
- Capítulo 4
- Capítulo 5
- Capítulo 6
- Capítulo 7
- Capítulo 8
- Capítulo 9

Prefacio: medir y evaluar  
 Pruebas de Certificación  
**Planificación**  
 Tabla de especificaciones  
 Construcción de ítems  
 Revisión - Pretest  
 Standard Setting  
 Evaluación mediatizada por jueces  
 Equiparación  
 Epílogo

## Diseño

# ¿Qué? ¿Cómo? ¿Quién? ¿A quién? ¿Para?

“The overall plan provides a systematic framework for all major activities associated with test development, makes explicit the most important a priori decisions, outlines a realistic timeline and emphasizes test security and quality control procedures from the onset... The overall plan requires an explicit delineation of the *validity evidence* that is needed to support each of the intended score interpretations and uses.” (Lane et al., 2016; p. 5)

# Diseño

- Justificación
  - Desarrollo teórico
  - Necesidad práctica
  - Avances (tecnología, metodología...)
- Definición del constructo
  - Taxonomía sistemática del dominio
    - Identificación de niveles (factores)
- Uso de las puntuaciones
  - Certificación
  - Evaluación en el aula
- Destinatarios: Población objeto
- Posibles fuentes de sesgo
  - Variables a considerar en el DIF
  - Acomodación de la prueba
- Características internas (test)
  - Rango de puntuaciones (100-150, 0-300...)
  - Formato
  - Demandas cognitivas

# Pongamos que hablo de...

## mediación

5. Actividades de mediación. 5.1 Objetivos. Traducir oralmente en un nuevo texto coherente, parafraseándolos o resumiéndolos, información, opiniones y argumentos contenidos en textos escritos u orales largos y minuciosos de diverso carácter y procedentes de diversas fuentes. Hacer una interpretación consecutiva en charlas, reuniones, encuentros o seminarios relacionados con la propia especialidad, con fluidez y flexibilidad, transmitiendo la información importante en sus propios términos. Mediar con fluidez y eficacia entre hablantes de la lengua meta o de distintas lenguas sobre temas tanto relacionados con los campos de interés personal o de la propia especialización como de fuera de dichos campos (p. e. en reuniones, seminarios, mesas redondas, o en situaciones potencialmente conflictivas), teniendo en cuenta las diferencias y las implicaciones sociolingüísticas y socioculturales de las mismas y reaccionando en consecuencia, transmitiendo, de manera clara y concisa, información significativa, y formulando las preguntas y haciendo los comentarios pertinentes con el fin de recalcar los detalles necesarios o comprobar supuestas inferencias y significados implícitos. (p. e. ensayos, o conferencias)... (RD 1041/2017, BOE 23 de Diciembre)

# Contenido

- Capítulo 0
- Capítulo 1
- Capítulo 2
- Capítulo 3**
- Capítulo 4
- Capítulo 5
- Capítulo 6
- Capítulo 7
- Capítulo 8
- Capítulo 9

Prefacio: medir y evaluar  
Pruebas de Certificación  
Planificación  
**Tabla de especificaciones**  
Construcción de ítems  
Revisión - Pretest  
Standard Setting  
Evaluación mediatizada por jueces  
Equiparación  
Epílogo

# El cómo


- Número de tareas
- Número de ítems
  - Tipo
  - Orden
  - Formato de respuesta
- Formato presentación
- Instrucciones de entrega de los materiales
- Seguridad
- Requerimientos cognitivos

"Content specification and format specification articulate the scope of the domain including the KSAs to be measured, the cognitive processes required, the item formats and the type of response." (AERA et al., 2014, p.76) \*KSA = Knowledge, skills and abilities

▶ 0:00 / 0:15

# Taxonomía de Bloom

Crear	Nivel de complejidad alto
Evaluar	
Analizar	
Aplicar	
Comprender	
Recordar	Nivel de complejidad bajo



- **Recordar:** reconocer listas, describir, identificar, recuperar, denominar, localizar, encontrar.
- **Comprender:** interpretar, resumir, inferir, parafrasear, clasificar, comparar, explicar, ejemplificar.
- **Aplicar:** implementar, desempeñar, usar, ejecutar.
- **Analizar:** comparar, organizar, deconstruir, atribuir, delinear, encontrar, estructurar, integrar.
- **Evaluar:** revisar, formular hipótesis, criticar, experimentar, juzgar, probar, detectar, monitorear.
- **Crear:** diseñar, construir, planear, producir, idear, trazar, elaborar

Categorías utilizadas por el Marco de referencia para las actividades comunicativas

	Interacción	Expresión
Orales	Conversación	Descripción de su especialidad
	Debate informal	
	Cooperación centrada en el objeto	
Escritas		Informe de su especialidad

# Contenido

- Capítulo 0
- Capítulo 1
- Capítulo 2
- Capítulo 3
- Capítulo 4**
- Capítulo 5
- Capítulo 6
- Capítulo 7
- Capítulo 8
- Capítulo 9

- Prefacio: medir y evaluar
- Pruebas de Certificación
- Planificación
- Tabla de especificaciones
- Construcción de ítems**
- Revisión - Pretest
- Standard Setting
- Evaluación mediatizada por jueces
- Equiparación
- Epílogo

# Ítems

- Unidad básica de información
- Impacto directo sobre la calidad de la evaluación
- Pregunta, actividad, performance.....
- Algo más que una “pregunta tipo test” (a, b, c, d)

## Mínimos a considerar

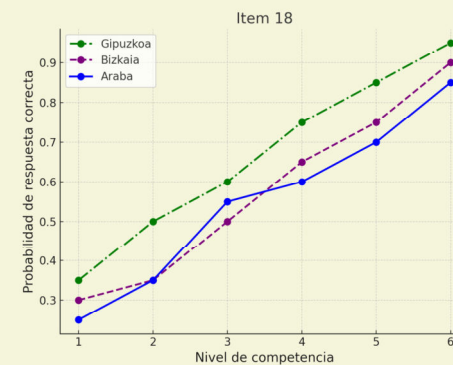
- Representación
- Relevancia
- No DIF (Funcionamiento diferencial del ítem; **D**ifferential **I**tem **F**unctioning)

33

Elosua, 2024



## Algo sobre DIF



Un ítem presenta DIF cuando la probabilidad de respuesta correcta para personas con el mismo nivel de competencia difiere entre grupos (lingüísticos, culturales, género..)

34

Elosua, 2024



## Ítems: Tipos

- Tarea demandada
  - Selección
  - Producción
- Corrección
  - Objetiva
  - Mediatizada por jueces
- Demanda de rendimiento
  - Ejecución típica
  - Ejecución máxima

35

Elosua, 2024



## Ítems de selección

### Ventajas

- Objetividad en la corrección
- Evaluación amplia de contenidos en poco tiempo
- Corrección mecanizada
- Elevada fiabilidad

### Inconvenientes

- Difíciles de construir
- Inhiben la creatividad en la respuesta
- Riesgo de copia

36

Elosua, 2024





# Ítems de construcción

## Ventajas

- Mejoran la representación del constructo
- Demandas cognitivas complejas
- Validez aparente (**auténticos**)
- Fáciles de construir

## Inconvenientes

- Corrección más compleja
  - Sesgo del evaluador
  - Rúbricas
- Exigen más tiempo (candidato, evaluador)
- Limitados con respecto al contenido a evaluar

# Contenido

Capítulo 0

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

**Capítulo 5**

Capítulo 6

Capítulo 7

Capítulo 8

Capítulo 9

Prefacio: medir y evaluar

Pruebas de Certificación

Planificación

Tabla de especificaciones

Construcción de ítems

**Revisión - Pretest**

Standard Setting

Evaluación mediatizada por jueces

Equiparación

Epílogo

# Avoid surprises

- Revisión
  - Contenido
  - Entrevistas cognitivas
- Prueba piloto

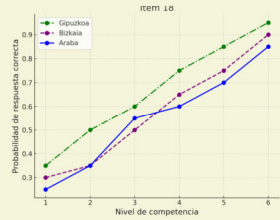
# Revisión de contenido



- Comprensión lectora
- Comprensión auditiva
- Expresión escrita
- Expresión oral

- Grupo de expertos multidisciplinar
  - Constructo
  - Construcción de instrumentos
  - Población destinataria
  - Nivel evaluado
  - ...

# Contenido



- Forma
- Evidencias de validez
  - Relevancia (Jueces)
  - Adecuación (Jueces)
  - Posible sesgo (DIF) (Jueces)

# Entrevistas cognitivas



- Modelos de procesamiento de información (Tourangeau)
  - Comprensión: intencionalidad y significado
  - Recuperación: estrategias de evocación
  - Decisión: reflexión
  - Respuesta: consciente o automática
- Protocolos - entrevistas semi-estructuradas

# Prueba piloto



- Comprensión lectora
- Comprensión auditiva
- Expresión escrita
- Expresión oral

- Ensayo general
- Test en miniatura

# Validación

- Condiciones de aplicación
  - Tiempo
  - Comprensión tareas e ítems
  - ...
- Índices psicométricos clásicos
- Control de posibles errores

“6. To discover weaknesses or needed improvements in the mechanics of test taking, in the directions to examiner and examinee, in the provisions for the responses, in the sample or fore-exercises, in the typographical format, and so forth. ... The experimental tryout may be simple or it may be extremely elaborate and expensive.”(Conrad, 1951, p. 251)

# Ejercicio

## Contenido

Capítulo 0

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

**Capítulo 6**

Capítulo 7

Capítulo 8

Capítulo 9

Prefacio: medir y evaluar

Pruebas de Certificación

Planificación

Tabla de especificaciones

Construcción de ítems

Revisión - Pretest

**Standard Setting**

Evaluación mediatizada por jueces

Equiparación

Epílogo

## Standard setting: Apto - No apto



He aprobado Me han suspendido

- Comprensión escrita
- Comprensión oral

- Objetiva
- Estandarizada
- Fiable
- Válida

## Validación

- Definición de competencias por nivel
- Procedimientos:
  - Angoff (Jueces)
  - Bookmark (Jueces + datos (IRT))
  - ...

“Standard setting should be seen as a critical aspect of the test development process best carried out in concert with all other aspects of the development process. Far from being a purely methodological process, standard setting ideally involves [policy makers](#), [test developers](#) and [measurement specialists](#) early on to ensure that the test results will be useful and defensible.” (Bejar, 2008)

## Competencia mínima (Angoff)

Jueces	1	2	3	4	Media
A	80	75	75	70	75,00
B	70	85	80	80	78,75
C	90	90	85	80	86,25
D	90	75	70	75	77,50
Media	82,50	81,25	77,50	76,25	<b>79,37</b>

- Porcentaje de personas en el nivel mínimo de competencia que respondería correctamente al ítem
- Promedio de proporciones

## Competencia mínima (Angoff)

Jueces	1	2	3	4	Media
A	80	75	75	70	75,00
B	70	85	80	80	78,75
C	90	90	85	80	86,25
D	90	75	70	75	77,50
Media	82,50	81,25	77,50	76,25	<b>79,37</b>

Promedio: **79,37%**

Aplicando el porcentaje del 79,37% a la prueba (4 ítems):

$$\text{Puntodecorte} = 4 \times \frac{79,37}{100} = 3,17$$

Nota de corte (redondeo): **3**

# Ejercicio

## Contenido

Capítulo 0

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

**Capítulo 7**

Capítulo 8

Capítulo 9

Prefacio: medir y evaluar

Pruebas de Certificación

Planificación

Tabla de especificaciones

Construcción de ítems

Revisión - Pretest

Standard Setting

**Evaluación mediatizada por jueces**

Equiparación

Epílogo

# Evaluación mediatizada por jueces



- Expresión escrita
- Expresión oral

- Objetiva
- Estandarizada
- Fiable
- Válida

# Validación

- Validación de rúbricas
- Diseño de asignación de tareas (*raters design*)
- Diseño de asignación de puntuaciones (*rating design*)
  - Número de jueces
  - Método de asignación
  - Resolución de discrepancias

“Scoring is a key step in the development and further use of any assessment tool. While in many testing contexts scoring is almost undisputed, in rater-mediated assessment of writing, it is still a matter of discussion.” (Elosua, 2022)

# Rúbricas

DELE C1. Expresión e Interacción escritas - Escala analítica	
<b>Adecuación al género discursivo</b>	
3	El texto es adecuado para la situación planteada, tiene en cuenta al destinatario y se adecua al estilo y la estructura del género, utilizando el grado de formalidad, el tono, el nivel de detalle y el léxico apropiados para el mismo. En las argumentaciones presenta argumentos, ejemplos, apreciaciones críticas y conclusiones con claridad y fluidez. En textos creativos (artículos, etc.), el estilo es convincente, personal y natural. En cartas e informes, sabe emplear un registro adecuado al contexto planteado e incluye de forma ordenada los detalles necesarios para lograr su objetivo comunicativo.
2	El texto es adecuado para la situación planteada, tiene en cuenta al destinatario y en general se adecua al estilo y la estructura del género, pero presenta alguna inconsistencia en el grado de formalidad, el tono, el nivel de detalle o el léxico, o tiene alguna alteración en la organización que, en cualquier caso, no afectan de forma significativa al registro o la estructura. Las argumentaciones son sólidas y están bien organizadas, aunque en algún caso se eche en falta algún ejemplo o argumento para darle al texto mayor claridad y fluidez. En textos creativos, el estilo es consistente, aunque le falte algo de naturalidad. En cartas e informes, sabe emplear un registro adecuado al contexto planteado, pero puede faltarle algún detalle o puede haber alguna alteración en el orden.
1	El candidato omite o malinterpreta uno de los puntos de orientación. El texto responde a la situación planteada y tiene en cuenta al destinatario, pero presenta inconsistencias de registro, estilo o estructura. El grado de formalidad, el tono, el nivel de detalle o el léxico no son siempre los más apropiados para el contexto. Las argumentaciones carecen de algún elemento importante en las mismas (argumentos secundarios, ejemplos, apreciaciones críticas o conclusiones), lo que afecta a la claridad de las mismas. En textos creativos, el estilo no es consistente ni natural. En cartas e informes, se pueden ver problemas de registro, de información o de orden que afectan al logro del objetivo comunicativo planteado.
0	El candidato omite o malinterpreta dos de los puntos de orientación dados. El texto no se corresponde en registro, estilo o estructura a la situación planteada ni sigue las pautas dadas. El grado de formalidad, el tono, el nivel de detalle o el léxico no son adecuados para el contexto. Las argumentaciones no mantienen una estructura organizada y que permita seguir los razonamientos del candidato. En textos creativos, el estilo es inapropiado y no se siguen las indicaciones dadas. En cartas e informes, no se incluye la información necesaria o no hay un orden claro. El candidato no sigue ninguno de los puntos de orientación dados o se ataja de ellos, ni se ajusta a la extensión indicada. Tarea 1: menos de 150 palabras. Tarea 2: menos de 125 palabras.

- Interpretación de los criterios
- Interpretación de las escalas
  - Número de puntos
    - 3, 4, 5...?
  - Significado de los puntos

# Raters design

	Juez_1	Juez_2	Juez_3
1			
2			
3			
4			
5			
6			
7			

	Juez_1	Juez_2	Juez_3
1			
2			
3			
4			
5			
6			
7			

	Juez_1	Juez_2	Juez_3
1			
2			
3			
4			
5			
6			
7			

- Diseño de distribución de tareas
  - Número de jueces por tarea
  - Fiabilidad
  - Validez aparente

# Rating design

- Método de asignación de puntuaciones
  - Consenso
  - Media
  - Un único juez?
- Método de resolución de discrepancias
  - Tertium quid (2ª ronda)
  - Paridad (2ª ronda)
  - Experto (2ª ronda)
- Validez aparente
- Criterios económicos
- Practicabilidad

60

Elosua, 2024

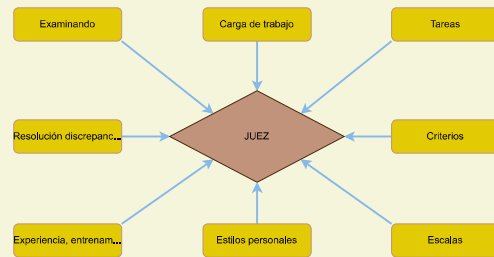


# Características personales

- Estilo
  - Dureza
  - Benevolencia
  - Tendencia al centro
  - ...
- Carga de trabajo
- Experiencia previa
- Entrenamiento
- Monitorización (?)
- Impacto del examinando (presencial)

61

Elosua, 2024

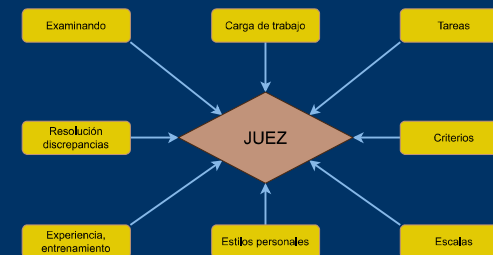


62

Elosua, 2024



# Juez: instrumento de medida

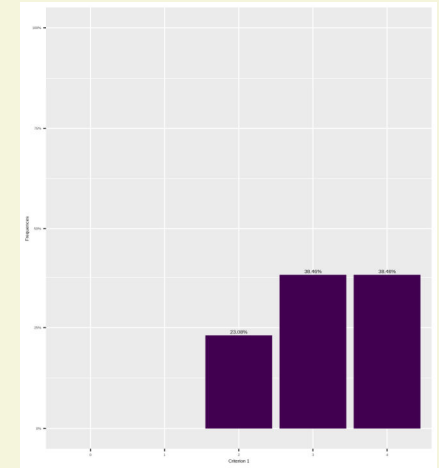
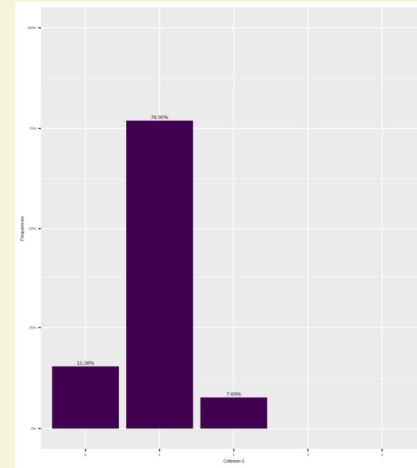


- Control de sesgo (bias)
  - Individual
  - Pares
  - Global

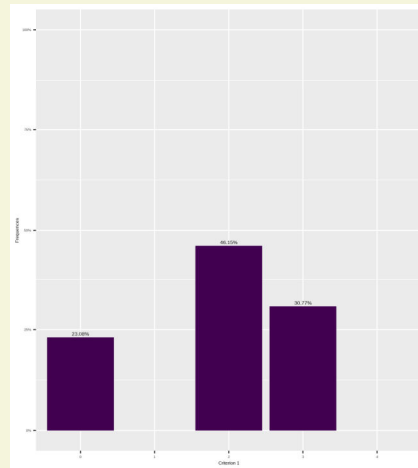
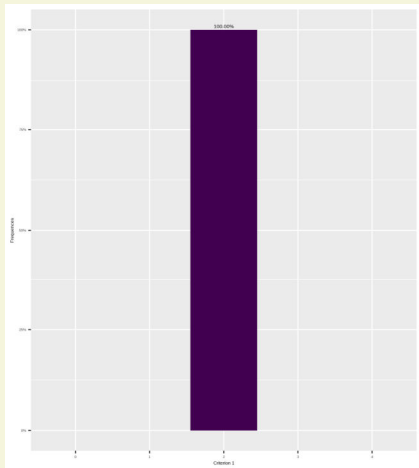


# Algunos ejemplos

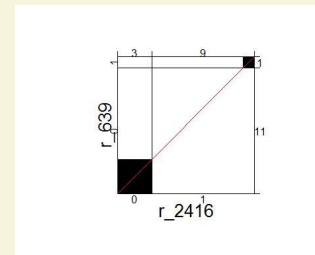
# Individual: Media aritmética



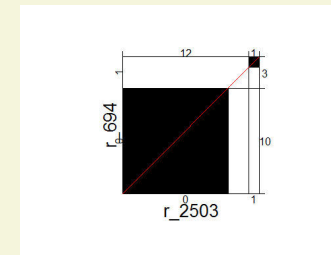
# Individual: DT



# Pares



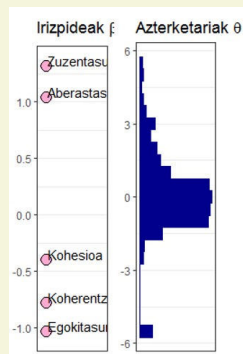
• Kappa = 0,06



• Kappa = 0,43

# Global

- Many-facets
- Generalizabilidad



Elosua, 2024

# Ejercicio

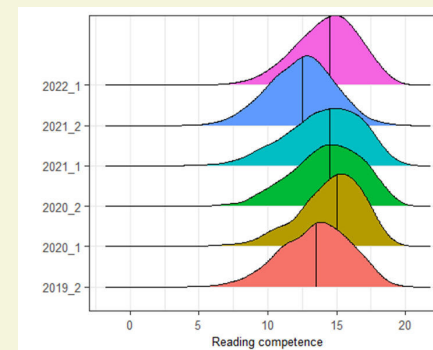
## Contenido

- Capítulo 0
- Capítulo 1
- Capítulo 2
- Capítulo 3
- Capítulo 4
- Capítulo 5
- Capítulo 6
- Capítulo 7
- Capítulo 8**
- Capítulo 9

Prefacio: medir y evaluar  
Pruebas de Certificación  
Planificación  
Tabla de especificaciones  
Construcción de ítems  
Revisión - Pretest  
Standard Setting  
Evaluación mediatizada por jueces  
**Equiparación**  
Epílogo

Elosua, 2024

## Equiparación de puntuaciones: Interrogante

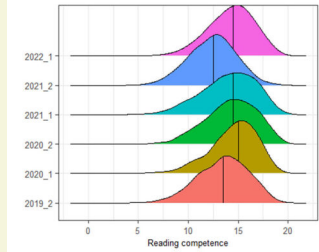


¿La prueba de esta edición es más difícil o la competencia de los candidatos es menor?

Elosua, 2024



## Equiparación de puntuaciones



- Procedimiento estadístico
- Fundamental en la validez y equidad de las calificaciones
- Calificaciones comparables y consistentes entre versiones y muestras

## Conceptos Claves

- **Pruebas**
  - **Referencia:** Sirve como “modelo”
  - **Nueva Forma:** Aquella que se desea equiparar
- **Equiparación de Puntuaciones:** Ajuste estadístico que se realiza para garantizar la comparabilidad entre puntuaciones de diferentes versiones (ediciones) de un examen, a pesar de las diferencias en el contenido y entre los grupos de candidatos.

## Diseños de Equiparación

	Test 1	Test 2
1	■	
2	■	
3	■	
4	■	
5		■
6		■
7		■
8		■

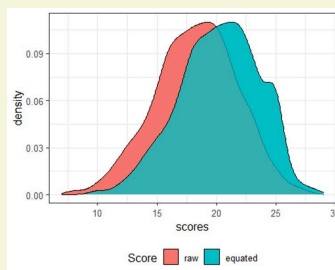
Test 1	Test 2
■	
■	
■	
■	■
■	■
	■
	■
	■

- **Items de anclaje:** ítems comunes entre versiones. **NEAT:** Non equivalent anchor test.
- **Grupos equivalentes:** Grupos comparables. **EG:** Equivalent group.

## Proceso de Equiparación

1. Definir el diseño
2. Recoger datos
3. Implementar del método
4. Estimar el error (variabilidad)
5. Validar

## Ejemplo Práctico



Formas	N	Media	DT	Mínimo	Máximo
Nueva	2010	18.29	3.45	7.00	28.00
Referencia	1942	20.34	3.57	7.00	30.00
EQUIPARADA	2010	20.23	3.35	9.00	29.00

Elosua, 2024



## Contenido

Capítulo 0

Capítulo 1

Capítulo 2

Capítulo 3

Capítulo 4

Capítulo 5

Capítulo 6

Capítulo 7

Capítulo 8

**Capítulo 9**

Prefacio: medir y evaluar

Pruebas de Certificación

Planificación

Tabla de especificaciones

Construcción de ítems

Revisión - Pretest

Standard Setting

Evaluación mediatizada por jueces

Equiparación

**Epílogo**

Elosua, 2024



“Developments in language testing research in the past 20 years (i.e., in the 1980s and 1990s) have brought language testers into closer contact and affinity with their fellow **applied linguists**, as well as with specialists in **measurement and technology**. The blossoming of language testing research that we have witnessed in this period has provided us with a rich variety of research approaches and tools, while at the same time broadening the research questions that are being investigated.

...

The professionalization of our field and **validation research** will both, I believe, continue to be vital to language testing. However, these cannot proceed independently of each other. The primary impetus for professionalization is the need to establish standards for **ethical conduct**, which itself must be grounded in **valid test use**. Thus, there can

be no consideration of ethics without validity.” (Bachman, 2000, p.2)

▶ 0:00 / 1:03



## Pruebas de Certificación de la Competencia Comunicativa

- VALIDEZ y uso CORRECTO
- **Binomio**: Lingüística y psicometría
- **SME** (Subject Matter Expert) y **Know how**





Gràcies - Eskerrik asko - Gracias - Gracias

Paula Elosua

[Paula.elosua@ehu.eus](mailto:Paula.elosua@ehu.eus)