Análisis longitudinal para la estandarización de criterios y formación de examinadores

GIELE
Grupo de interés en evaluación de lenguas en España

UAB Idiomes
Servei de Llengües

OBJETIVOS

- **Presentación** Examen Multinivell Inglés B1-B2-C1
- **Consideraciones previas** Cómo estandarizar muestras
- **Desarrollo** Cómo estandarizar evaluadores y garantizar la calidad de las evaluaciones emitidas
- **Análisis** La experiencia de nuestro centro

UAB

Presentación

Examen Multinivel Inglés B1-B2-C1

El examen evalúa el grado de competencia del examinando en las cuatro destrezas comunicativas y consta de las pruebas siguientes:

- Prueba de comprensión escrita y uso de la lengua
- Prueba de comprensión oral
- Prueba de expresión escrita
- Prueba de expresión oral

UAB

Examen Multinivel Inglès B1-B2-C1

En la expresión escrita se evalúan los textos según estos criterios:

- Adecuación de la tarea.
- Cohesión, coherencia y organización.
- Repertorio y corrección léxicos.
- Repertorio y corrección gramaticales.

UAB

Examen Multinivel Inglés B1-B2-C1

En la expresión oral se evalúa la interacción y la producción oral según estos criterios:

- Fluidez, coherencia e interacción.
- Repertorio y corrección léxicos.
- Repertorio y corrección gramaticales.
- Pronunciación y entonación.

UAB

CÓMO ESTANDARIZAR CRITERIOS

- Existen numerosos métodos
- Los estándares dependen de varios factores críticos
  - Características del grupo de evaluadores que se utiliza en el proceso
  - Cantidad de formación que reciben los evaluadores
  - Detalles del método que se implementa

UAB

CÓMO ESTANDARIZAR CRITERIOS

- Los métodos para establecer estándares en evaluaciones con ítems de opción múltiple están bien desarrollados.
- Los métodos para establecer estándares para ítems de respuesta construida no están tan bien desarrollados.

UAB

CÓMO ESTANDARIZAR CRITERIOS

Se recomiendan seguir 9 pasos:

1. Seleccionar el método
2. Formar equipo expertos y diseño
3. Preparar la descripción de las categorías
4. Entrenar a los evaluadores en el método
5. Recopilar datos

UAB

CÓMO ESTANDARIZAR CRITERIOS

Se recomiendan seguir 9 pasos:

6. Dar *feedback* y proponer discusión
7. Establecer la evaluación estandar
8. Evaluación de los evaluadores
9. Evidencias de validez y documentación

UAB

1r paso: Seleccionar el método

En función de:

- Tipología de ítems
- Tiempo y recursos que son necesarios
- Conocimiento previo
- Disponibilidad de evidencias de validez

UAB

1r paso: Seleccionar el método

Tipologías de métodos en función de:

- Ítems de elección múltiples y rúbricas
- Candidatos
- Tareas de los candidatos
- Perfiles de puntuación
- Compromiso

UAB

2º Paso: Formar equipo de expertos y diseño

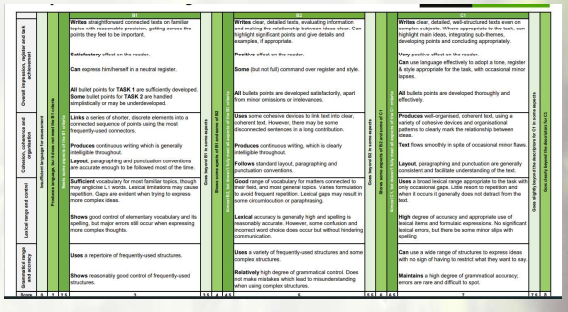
– ¿Quiénes son los expertos? Grupo representativo

– Posibilidad de hacer dos equipos paralelos.

UAB

3r Paso: Preparar una descripción de los criterios

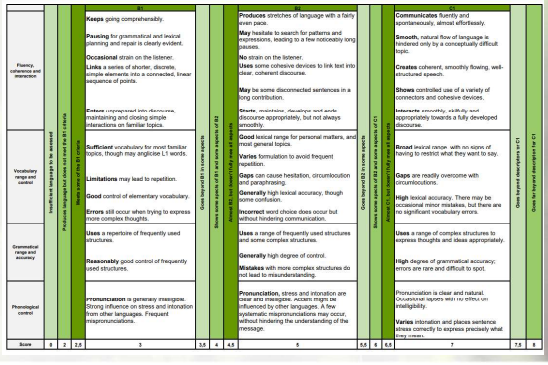
– Criterios EE



UAB

3r Paso: Preparar una descripción de los criterios

– Criterios EO Analítica



UAB

3r Paso: Preparar una descripción de los criterios

– Criterios EO Holística

				B1			B2			C1		
Overall Impression	Not sufficient language to be assessed	produces language but does not meet the B1 criteria	Meets some of the B1 criteria	Can give straightforward descriptions, give reasons and explanations for opinions, with reasonable fluency, presenting them as a connected linear sequence of points. Keeps going comprehensibly, alhough pausing for planning and repair is clearly evident. Uses a range of frequent language (lexis and grammar) with reasonable accuracy to deal with most everyday situations. Pronunciation is generally intelligible. Strong influence on stress and intonation from other languages.	A few aspects of B2, but more aspects of B1	half spects of B1 and half aspects of B2	A few aspects of B1, but more aspects of B2	Gives clear, detailed descriptions and presentations on a wide range of subjects, expanding and supporting ideas with relevant examples. Produces stretches of language at a fairly even pace. There are few noticeably long pauses. Good range of language to speak relatively accurately on most general topics. The influence from other languages and a few mispronunciations might be noticeable, but do not hinder understanding.	A few aspects of C1, but more aspects of B2	half spects of B2 and half aspects of C1	A few aspects of B2, but more aspects of C1	Gives clear, detailed descriptions and presentations on complex subjects, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion. Communicates fluently, coherently and spontaneousiy, almost effortlessly. Good command of a broad range of language with no signs of having to restrict what they want to say. No significant lexical errors, and/or they are difficult to spot. Clear and natural with occasional influence from other lanouages. Uses intonation to express precisely what they mean.		
Score	0	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8

UAB

4º Paso: Entrenar a los evaluadores en el método

- Familiarizarse con las categorías y definir los objetivos de cada paso, explicar y demostrar los avances que se darán en cada uno de ellos
- Familiarizarse con las rúbricas, las plantillas de evaluación, con el contenido a evaluar
- Simular las condiciones más similares al examen

UAB

5º Paso: Recopilar datos

– *Google form*

– Aunque se recoja en papel/lápiz, informatizar cuanto antes para poder llevar un control y un análisis seguro

UAB

6º Paso: Dar comentarios y facilitar la discusión

- Fomentar la discusión alrededor de las valoraciones emitidas por cada experto.
- Con la finalidad de llegar a un acuerdo.

UAB

7º Paso: Establecer la evaluación estándar

Depende del método

- Holística
- Analítica

UAB

8º Paso: Evaluación de los evaluadores

- Cómo han entendido las categorías y mejoras
- Calidad de la formación recibida
- Tiempo dedicado a cada parte
- Concordancia con los resultados esperados
- Su explicación de por qué el comportamiento desviado
- Confianza en sus resultados

UAB

9º Paso: Evidencias de validez y documentación

Informe sobre el análisis de las características psicométricas de fiabilidad y validez de las calificaciones de los criterios.

UAB

CÓMO EVALUAR LOS ESTÁNDARES

- **Procedimiento**
  - Claridad (previa a la implementación)
  - Practicabilidad (facilidad de implementación)
  - Implementación de procedimientos
  - Comentarios de los evaluadores
  - Documentación

UAB

CÓMO EVALUAR LA ESTÁNDARES

- **Interna**
  - Consistencia con el método
  - Consistencia inter-evaluadores
  - Consistencia intra-evaluadores

UAB

CÓMO EVALUAR LOS ESTÁNDARES

- **Externa**
  - Comparar las conclusiones con las de otros métodos
  - Otras fuentes de información

UAB

Análisis: nuestra experiencia

ESTABLECIMENTO DE ESTÁNDARES Y FORMACIÓN- LOS 9 PASOS A SEGUIR -

PASOS	Establecimiento de Estándares	Formación
1	x	x
2	x	x
3	x	
4	x	x
5	x	x
6	x	x
7	x	
8	x	x
9	x	x

UAB

Evidencias de validez y fiabilidad

- Procedimiento
- Interna
  - Consistencia con el método: An. Generalizabilidad
  - Consistencia inter-evaluadores: SD , CCI, gráficas
  - Consistencia intra-evaluadores: correlación

UAB

EVIDENCIAS DE VALIDACIÓN

- Examen multinivel B1-B2-C1: Expresión Escrita y Oral
- Dos centros acreditadores (10 y 5 evaluadores)
- Estandarización de muestras y de evaluadores:

  2018, 2019, 2020, 2023, 2024

UAB

EVIDENCIAS DE VALIDACIÓN

- 10 evaluadores
- Muestras de todos los niveles
  - 9 muestras de Expresión Escrita
  - 8 muestras de Expresión Oral

UAB

E. DE VALIDEZ INTERNA EE

Consistencia con el método

- P X C X R
  - La variabilidad debida a los criterios es inapreciable
  - La variabilidad viene explicada por la combinación Examinandos (P) y Evaluador (R)
- P x R
  - 55 % de la variabilidad es explicada por los Examinandos (P) y solo un 3 % por los Evaluadores (R)

UAB

E. DE VALIDEZ INTERNA EE

Consistencia Inter-Evaluador

- DESVIACIÓN DE LAS PUNTUACIONES
  - COMPARACIÓN DE LA DESVIACIÓN ESTANDAR (SD) ENTRE EVALUADORES
- ACUERDO ENTRE LA PUNTUACIÓN ESTANDARIZADA Y LA PUNTUACIÓN DEL EVALUADOR
  - COEFICIENTE DE CORRELACIÓN INTRACLASE ABSOLUTO (CCI-ABS)

UAB

E. DE VALIDEZ INTERNA EE

Consistencia Inter-Evaluador

- DESVIACIÓN (SD) DE LAS PUNTUACIONES

	SD	DIFERENCIA PO
PO	1,338	
EV1	1,042	0,30
EV2	1,445	0,11
EV3	1,421	0,08
EV4	1,054	0,28
EV5	1,356	0,02
EV6	1,822	0,48
EV7	1,547	0,21
EV8	1,047	0,29
EV9	1,152	0,19
E V10	1,241	0,10

UAB

E. DE VALIDEZ INTERNA EE

Consistencia Inter-Evaluador

- ACUERDO ENTRE LA PUNTUACIÓN ESTANDARIZADA Y LA PUNTUACIÓN DEL EVALUADOR

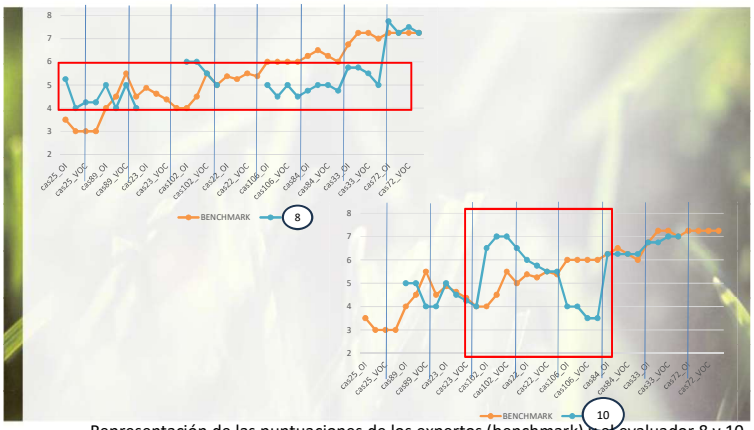
	CCI-ABS	CCI 95% INF	CCI 95% SUP
EV1	0,933	0,868	0,966
EV2	0,745	0,506	0,869
EV3	0,985	0,957	0,995
EV4	0,922	0,820	0,966
EV5	0,915	0,787	0,961
EV6	0,974	0,925	0,990
EV7	0,953	0,860	0,983
EV8	0,713	0,390	0,866
EV9	0,775	0,504	0,897
E V10	0,526	-0,044	0,783

UAB

INTER EVALUADOR - EE

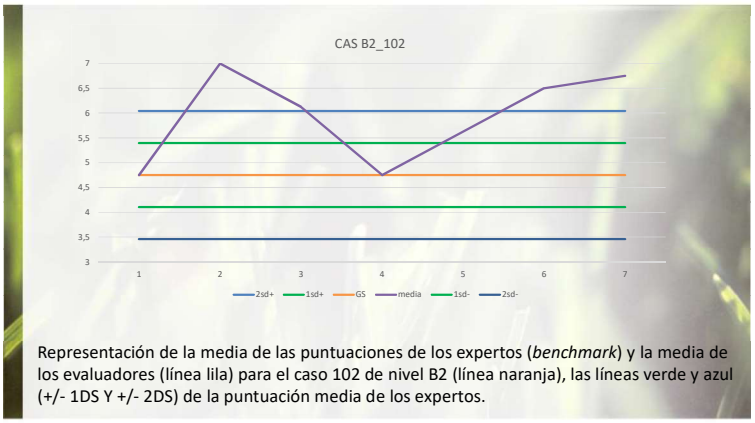
Representación de las puntuaciones de los expertos (benchmark) y el evaluador 1

INTER EVALUADOR - EE



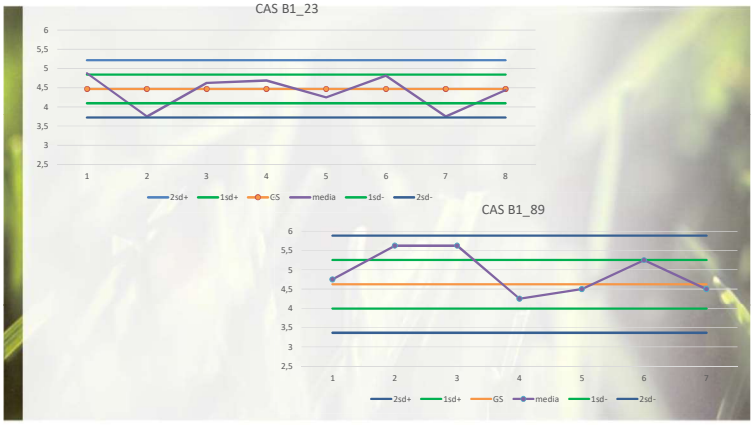
Representación de las puntuaciones de los expertos (benchmark) y el evaluador 8 y 10

INTER EVALUADOR - EE



Representación de la media de las puntuaciones de los expertos (*benchmark*) y la media de los evaluadores (línea lila) para el caso 102 de nivel B2 (línea naranja), las líneas verde y azul (+/- 1DS Y +/- 2DS) de la puntuación media de los expertos.

INTER EVALUADOR - EE



E. DE VALIDEZ INTERNA EE

Consistencia Intra-Evaluador

- INFLUENCIA DEL PASO DEL TIEMPO
  - CORRELACIÓN PEARSON / COEFICIENTE DE CORRELACION INTRACLASE ABSOLUTO
- ACUERDO ENTRE DOS SISTEMAS DE EVALUACIÓN (ANÁLITICA Y HOLÍSTICA)
  - CORRELACIÓN PEARSON/ COEFICIENTE DE CORRELACION INTRACLASE ABSOLUTO

UAB

INTRA EVALUADOR - EE

EJEMPLO NIVEL B1 – PASO DEL TIEMPO	PUNTUACIÓN 2018
CORRELACIÓN	PUNTUACIÓN 2024
GS	1,0000
EV1	0,9696
EV2	0,5918
EV5	0,8266
EV9	0,7062
EV10	0,9995

UAB

E. DE VALIDEZ INTERNA EO

Consistencia con el método

- P X C X R
  - La variabilidad debida a los criterios es inapreciable
  - La variabilidad viene explicada por la combinación Examinandos (P) y Evaluador (R)
- P x R (residual elevado)
  - 38 % de la variabiliadad es explicada por los Examinandos (P) y un 8 % por los Evaluadores (R)

UAB

E. DE VALIDEZ INTERNA EO

Consistencia Inter-Evaluador

- DESVIACIÓN (SD) DE LAS PUNTUACIÓN

	SD	DIFERENCIA PO
PO	1,377	
EV1	1,159	0,22
EV2	1,273	0,10
EV3	1,351	0,03
EV4	1,214	0,16
EV5	1,630	0,25
EV6	1,161	0,22
EV7	1,166	0,21
EV8	1,005	0,37
EV9	1,890	0,51
EV10	1,146	0,23

UAB

INTER EVALUADOR - EO

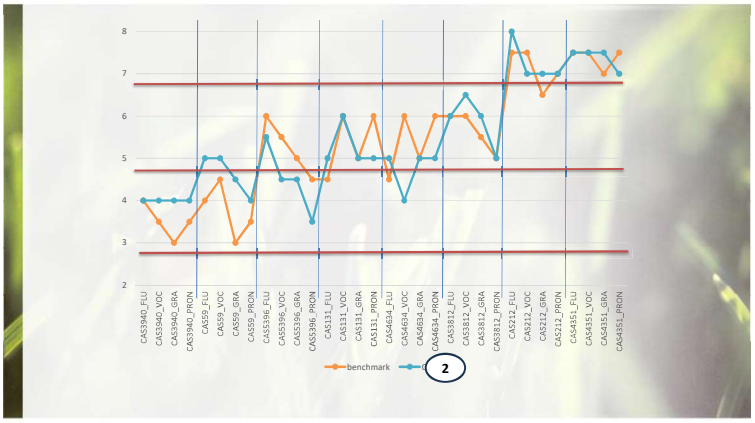
- Consistencia Inter-Evaluador

ACUERDO ENTRE LA PUNTUACIÓN ESTANDARIZADA Y LA PUNTUACIÓN DEL EVALUADOR

	CCI-ABS	CCI 95 % INF	CCI 95 % SUP
EV1	0,845	0,682	0,924
EV2	0,924	0,844	0,963
EV3	0,666	0,256	0,868
EV4	0,967	0,904	0,988
EV5	0,941	0,878	0,971
EV6	0,736	0,389	0,886
EV7	0,658	0,21	0,852
EV8	0,928	0,818	0,972
EV9	0,850	0,653	0,953
EV10	0,930	0,759	0,98

UAB

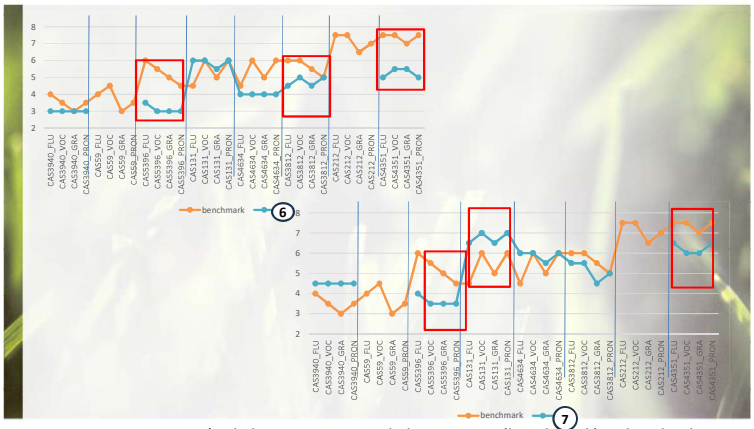
INTER EVALUADOR - EO



Representación de las puntuaciones de los expertos (benchmark) y el evaluador 2

UAB

INTER EVALUADOR - EO



Representación de las puntuaciones de los expertos (benchmark) y el evaluador 6 y 7

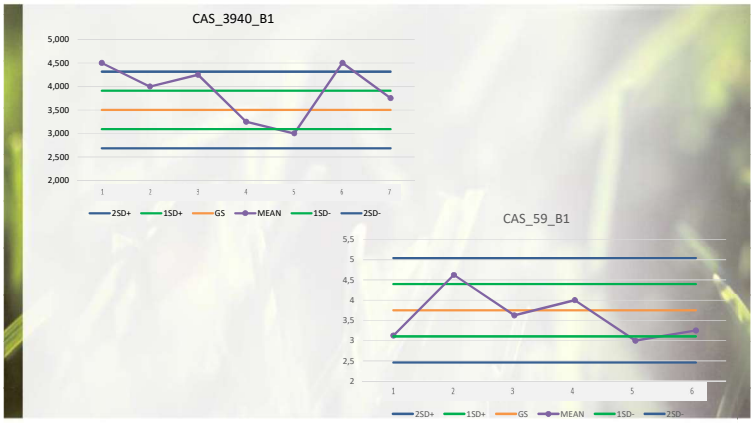
UAB

INTER EVALUADOR - EO

Representación de la media de las puntuaciones de los expertos (*benchmark*) y la media de los evaluadores (línea lila) para el caso 3812 de nivel B2 (línea naranja), las líneas verde y azul (+/- 1DS Y +/- 2DS) de la puntuación media de los expertos.

UAB

INTER EVALUADOR - EO



UAB

E. DE VALIDEZ INTERNA EO

Consistencia Intra-Evaluador

- INFLUENCIA DEL PASO DEL TIEMPO
  - CORRELACIÓN PEARSON / COEFICIENTE DE CORRELACION INTRACLASE ABSOLUTO
- ACUERDO ENTRE DOS SISTEMAS DE EVALUACIÓN (ANÁLITICA Y HOLÍSTICA)
  - CORRELACIÓN PEARSON/ COEFICIENTE DE CORRELACION INTRACLASE ABSOLUTO

UAB

INTRA EVALUADOR – EO

EJEMPLO NIVEL B1 – PASO DEL TIEMPO

CORRELACIÓN	PUNTUACIÓN 2018
	PUNTUACIÓN 2019
GS	1
EV1	0,6310
EV2	0,9995
EV3	0,8559
EV5	0,9890

UAB

¿QUÉ HACEMOS CON ESTOS RESULTADOS?

- Cómo los aprovechamos en nuestro día a día
  - Efectos en nuestra organizaicón de las correcciones
  - Necesidad de formación anual
- Qué hemos aprendido para un futuro
  - Realizar un estudio inter e intra correctores para poder dar un *feedback* más robusto

UAB

REVISIONES

- Reducir el número de 3ª correcciones
- Eliminar 4ª correcciones

UAB

ESTUDIO INTER INTRA

- Realizar un estudio inter-evaluador y intra-evaluador.
- Con una muestra de entre 20 y 25 exámenes para corregir.
- Además, si las muestras seleccionadas también son representativas de las pruebas de expresión escrita, se podrá analizar la interacción entre el tipo de tarea y el corrector.
- Al tratarse de un examen multinivel, seleccionaremos 2 por debajo del punto de corte, 2 por encima del punto de corte (esto hacen 4 casos fronteras) y 4 claramente del nivel.

UAB

FORMACIÓN

- Importancia de realizar formaciones cada año
- Modificar la recogida de datos para poder sacar el máximo partido a los datos recogidos en las formaciones

UAB

Bibliografia

Brennan, R. L. (2006). Educational Measurement (4th ed.). (Ed.). Westport, CT: Praeger, 2006

- Cook, L. & Pitoniak, M. (2027) Educational Measurement (5th ed.) (Ed.). Oxford University Press / Hardcover / 9780197654965 (ISBN-10: 0197654967) (**manuscrito no publicado, Julio 2027**)

UAB

Muchas gracias por su atención.

Laura Riera Grau	Rebeca P. García-Rueda
Laura.Riera@uab.cat	*Rebeca.Garcia@uab.cat*

GIELE
Grupo de interés en evaluación de lenguas en España

UAB Idiomes
Servei de Llengües