Cuestionario de Evaluación de Tests Revisado

CET–R

Este documento se complementa con el artículo que describe las razones que han motivado la revisión y actualización del modelo original CET (Prieto y Muñiz, 2000), así como el proceso de revisión llevado a cabo. Para hacer referencia al modelo se debe utilizar la siguiente referencia:

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G. y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. ***Papeles del Psicólogo,*** **37, 192-197.**

OBSERVACIONES PARA TENER EN CUENTA AL RESPONDER EL CUESTIONARIO DE EVALUACIÓN DE TESTS REVISADO (CET-R):

1) ***El CET-R es un cuestionario y, como tal, sus preguntas y opciones no deben modificarse.*** En algunas ocasiones, al utilizar el CET original en los procesos de revisión, los revisores han modificado alguna de las opciones existentes al no encontrar la opción de respuesta que estaban buscando. Esto debe evitarse. Debe, por tanto, responder valiéndose de las opciones que el CET-R ofrece. Si en alguna ocasión no encuentra la opción que está buscando, debe elegir la más similar. En las secciones para comentarios abiertos podrá hacer las aclaraciones que estime pertinentes.

2) ***Número de CET-R a rellenar.*** Si se ha de revisar una batería o más de un test (por ejemplo, el test normal y su versión abreviada), se pueden seguir dos estrategias. La que indica el CET-R es rellenar tantos CETs como tests haya que revisar. La que resulta menos costosa y también posible, si tiene sentido, sería utilizar solo un CET- R, dejando constancia donde corresponda de los diferentes resultados obtenidos por las distintas versiones del test.

3) ***Se debe revisar sólo la documentación entregada.*** En los tests comercializados se espera que realice su revisión a partir de la documentación que se le entrega. Ha ocurrido alguna vez que el manual omite cierta información que el/la revisor/a puede considerar relevante para evaluar la calidad de la prueba. En ese caso, lo apropiado es que pida dicha información al/a coordinador/a quien, a su vez, pedirá a la editorial la información requerida y verá si es posible obtenerla y en qué condiciones para después distribuirla a los revisores.

4) ***Revisión detallada de todas las secciones del manual.*** Cabe señalar que, en ocasiones, a pesar de que ciertos análisis no aparecen en un apartado diferenciado (por ejemplo, no aparece un apartado explícito de "análisis de ítems"), sí están incluidos en el manual, aunque pasan desapercibidos. Por ello, es importante que se revise con detalle toda la información del manual, de modo que no se marque la opción "no se aporta información en la documentación" cuando aparezca en otras secciones.

5) ***Se espera que todas las calificaciones queden argumentadas en las preguntas abiertas que resumen cada apartado***. El CET-R pide justificaciones sólo de algunas respuestas a preguntas concretas, pero no de la mayoría. De todos modos, esta justificación es muy importante y debe quedar reflejada de alguna forma en los comentarios generales. Cuando se pide, por ejemplo, "comentarios sobre la validez en general", se espera encontrar información que justifique las puntuaciones dadas a todas las preguntas sobre validez.

6) ***Deben responderse todas las preguntas.*** No debe dejar ninguna pregunta sin contestar, a no ser que en la pregunta explícitamente se indique que sólo se debe contestar si se han realizado cierto tipo de análisis. Si alguna pregunta del CET-R no resulta apropiada y no se responde, la razón por la que dicha pregunta se ha dejado en blanco debe quedar justificada en el resumen de la sección correspondiente, donde se pueden introducir comentarios abiertos.

7) ***En tests adaptados, los estudios de la versión original y de la adaptada no tienen la misma relevancia***. En los tests adaptados de otros idiomas y/o culturas, un asunto relevante es qué peso dar a los estudios realizados con el test original y a los estudios realizados en/tras el proceso de adaptación. Nuestra posición en este asunto es que deben tenerse en cuenta todos los estudios aportados, si bien debe dar más relevancia y peso en la evaluación a los que se aporten en el proceso de adaptación utilizando las poblaciones objetivo.

8) ***Para el cálculo de los promedios, los resultados se obtendrán a partir de la media aritmética de los apartados para los que se tiene información.*** No es necesario ponderar teniendo en cuenta el tamaño muestral de los estudios realizados, puesto que dicho tamaño se tiene en cuenta en otros apartados.

9) ***Términos psicométricos.*** En las revisiones anteriores realizadas con el CET hemos advertido que no siempre se asigna el mismo significado a los términos psicométricos empleados. Algunos términos que inducen a error se comentan a continuación:

a) Cuando se pregunta en el apartado de **Análisis de ítems** por su calidad se pide una valoración de la información psicométrica que el manual ofrece de los ítems y no si, tras su lectura, le parece que están bien o mal redactados.

b) Algo similar ocurre cuando se pregunta por **validez de contenido**. En realidad, se quiere saber qué comprobaciones se aportan sobre si el test evalúa las partes relevantes del constructo de interés.

c) Por lo que se refiere a los análisis de **sensibilidad y especificidad** que permiten evaluar la capacidad diagnóstica del test, en ocasiones los resultados son presentados como diferencias entre grupos y otras como evidencias de la capacidad del test para predecir la pertenencia a un cierto grupo diagnóstico. Esta información referida a capacidad diagnóstica debe incluirse en **evidencias de validez para predecir un criterio**.

En la dirección *http://glosarios.servidor-alicante.com/psicometria* encontrará un breve glosario de términos psicométricos que puede resultar útil.

Tras estas observaciones, en la página siguiente, encontrará el cuestionario propiamente dicho. En el primer apartado se le pedirá que haga una descripción general del test que ha de evaluar. En el segundo apartado valorará las características del test (la calidad de sus materiales, instrucciones, adaptación, desarrollo, sus ítems, etc.) y sus propiedades psicométricas (análisis de ítems, validez, fiabilidad e interpretación de las puntuaciones). En el tercer apartado se le pedirá una valoración global del test. Por último, encontrará la lista de referencias bibliográficas citadas en el CET-R.

1

DESCRIPCIÓN GENERAL DEL TEST

Si el test está compuesto de subtests heterogéneos en su formato y características, rellene un cuestionario para cada subtest. Cuando tenga sentido y sea factible se podrá utilizar un solo CET-R, dejando constancia donde corresponda de los diferentes resultados obtenidos por los distintos subtests.

1.1. **Nombre del test:**

1.2. **Nombre del test en su versión original** ***(si la versión española es una adaptación)***:

1.3. **Autor/es del test original:**

1.4. **Autor/es de la adaptación española:**

1.5. **Editor del test en su versión original:**

1.6. **Editor de la adaptación española:**

1.7. **Fecha de publicación del test original:**

1.8. **Fecha de publicación del test en su adaptación española:**

1.9. **Fecha de la última revisión del test** ***(si el test original es español, o de su adaptación española si se trata de un test adaptado)***:

1.10. **Clasifique el área general de la/s variable/s que pretende medir el test:** ***(Es posible marcar más de una opción)***

Identifique el área de contenido definido en la publicación. Si no hay una definición clara debe señalarlo en el apartado "Otra" e indicar cuál es el área de contenido más adecuada según la información proporcionada en el manual.

- ☐ Inteligencia
- ☐ Aptitudes
- ☐ Habilidades
- ☐ Psicomotricidad
- ☐ Neuropsicología
- ☐ Personalidad
- ☐ Motivación
- ☐ Actitudes
- ☐ Intereses
- ☐ Escalas de desarrollo
- ☐ Rendimiento académico / competencia curricular
- ☐ Escalas clínicas
- ☐ Potencial de aprendizaje
- ☐ Calidad de vida/bienestar
- ☐ Estrés/burnout
- ☐ Estilos cognitivos
- ☐ Otra (Indique cuál:)

1.11. Breve descripción de la/s variable/s que pretende medir el test:

Se trata de hacer una descripción no evaluativa del test, con 200-600 palabras. La descripción debe proporcionar una idea clara del test, lo que pretende medir y las escalas que lo conforman.

1.12. Área de aplicación: *(Es posible marcar más de una opción)*

Identifique el área o áreas de aplicación definidas en la publicación. Si no hay una definición clara debe señalarlo en el apartado "Otros" e indicar cuál es el área de aplicación más adecuada según la información proporcionada en el manual.

- ☐ Psicología clínica
- ☐ Psicología educativa
- ☐ Neuropsicología
- ☐ Psicología forense
- ☐ Psicología del trabajo y las organizaciones
- ☐ Psicología del deporte
- ☐ Servicios sociales
- ☐ Salud general y bienestar
- ☐ Psicología del tráfico
- ☐ Otra (Indique cuál:)

1.13. Formato de los ítems: *(Es posible marcar más de una opción)*

- ☐ Respuesta construida
- ☐ Respuesta dicotómica (sí/no, verdadero/falso, etc.)
- ☐ Elección múltiple
- ☐ Respuesta graduada / Tipo Likert
- ☐ Adjetivos bipolares
- ☐ Otro (Indique cuál:)

1.14. Número de ítems: *(Si el test tiene varias escalas, indique el número de ítems de cada una)*

1.15. Soporte: *(Es posible marcar más de una opción)*

- ☐ Administración oral
- ☐ Papel y lápiz
- ☐ Manipulativo
- ☐ Informatizado
- ☐ Otro (Indique cuál:)

1.16. Cualificación requerida para el uso del test de acuerdo con la documentación aportada:

Algunos países han adoptado sistemas para la clasificación de los tests en distintas categorías, en función de la cualificación requerida por los usuarios. Un sistema muy utilizado es el que divide los tests en tres categorías: Nivel A (tests de rendimiento y conocimientos), Nivel B (tests colectivos de aptitudes e inteligencia) y Nivel C (tests de aplicación individual de inteligencia, personalidad y otros instrumentos complejos)

- ☐ Ninguna
- ☐ Entrenamiento y acreditación específica. ***Indique el nombre de la institución que lleva a cabo la acreditación:***
- ☐ Nivel A
- ☐ Nivel B
- ☐ Nivel C
- ☐ Otra (*Indique cuál:*)

1.17. Descripción de las poblaciones a las que el test es aplicable:

Especifique el rango de edad, nivel educativo, etc. y si el test es aplicable en ciertas poblaciones específicas: minorías étnicas, personas con discapacidad, grupos clínicos, etc.

1.18. Indique si existen diferentes formas del test y sus características (formas paralelas, versiones abreviadas, versiones informatizadas o impresas, versiones para diferentes poblaciones –infantil vs. adultos- etc.) En el caso de que existan versiones informatizadas, describa los requisitos inusuales del hardware y software, si los hubiere, que fueran necesarios para administrar correctamente el test (grabación de sonido, pantallas de resolución inusual, etc.):

1.19. Procedimiento de corrección: *(Es posible marcar más de una opción)*

- ☐ Manual
- ☐ Hoja autocorregible
- ☐ Lectura óptica de la hoja de respuestas
- ☐ Automatizado por ordenador (existe software de corrección o plataformas de corrección on-line)
- ☐ Efectuada por la empresa suministradora (las hojas de respuesta se envían a la empresa para que esta se ocupe de la corrección)
- ☐ Mediante expertos
- ☐ Otro (Indique cuál:)

1.20. Puntuaciones:

Describa el procedimiento para obtener las puntuaciones directas, totales o parciales, corrección de la probabilidad de responder correctamente por azar, inversión de ítems, etc.

1.21. Escalas utilizadas: *(Es posible marcar más de una opción)*

- ☐ Puntuaciones basadas en percentiles:
  - ☐ Centiles
  - ☐ Quintiles
  - ☐ Deciles
- ☐ Puntuaciones estandarizadas:
  - ☐ Puntuaciones típicas
  - ☐ Eneatipos
  - ☐ Decatipos
  - ☐ T (Media 50 y desviación típica 10)
  - ☐ D (Media 50 y desviación típica 20)
  - ☐ CI de desviación (Media 100 y desviación típica 15 (Wechsler) o 16 (Stanford-Binet)
- ☐ Puntuaciones estandarizadas normalizadas (puntuaciones estandarizadas obtenidas bajo el supuesto de que su distribución es normal):
- ☐ Puntuaciones directas solamente
- ☐ Otras (Indique cuáles:)

1.22. Posibilidad de obtener informes automatizados: No

En caso afirmativo **haga una breve valoración** del informe automatizado en la que se hagan constar las características fundamentales, tales como tipo de informe y estructura, claridad, estilo, así como su calidad:

1.23. Tiempo estimado para la aplicación del test (instrucciones, ejemplos y respuestas a los ítems):

En aplicación individual:

En aplicación colectiva:

1.24. Documentación aportada por el editor: *(Es posible marcar más de una opción)*

- ☐ Manual
- ☐ Libros o artículos complementarios
- ☐ Discos u otros dispositivos magnéticos
- ☐ Información técnica complementaria y actualizaciones
- ☐ Otra (Indique cuál:)

1.25. Precio de un juego completo de la prueba (documentación, test, plantillas de corrección; en el caso de tests informatizados no se incluye el coste del hardware): . *Indique la fecha de consulta de precios:*
.

1.26. Precio y número de ejemplares del paquete de cuadernillos (tests de papel y lápiz): . *Indique la fecha de consulta de precios:* .

1.27. Precio y número de ejemplares del paquete de hojas de respuesta (tests de papel y lápiz): . *Indique la fecha de consulta de precios:* .

1.28. Precio de la administración y/o corrección, y/o elaboración de informes por parte del editor: . *Indique la fecha de consulta de precios:* .

2

VALORACIÓN DE LAS CARACTERÍSTICAS DEL TEST

2.1. Calidad de los materiales del test (objetos, material impreso o software):

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Impresión y presentación de calidad, objetos bien diseñados, software atractivo y eficiente, etc.)

2.2. Calidad de la documentación aportada:

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Descripción muy clara y completa de las características técnicas, fundamentada en abundantes datos y referencias)

2.3. Fundamentación teórica:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Descripción muy clara y documentada del constructo que se pretende medir y del procedimiento seguido para medir dicho constructo)

2.4. Adaptación del test *(si el test ha sido traducido y adaptado para su aplicación en España)*:

- **N/A** ☐ Característica no aplicable para este instrumento
- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Se describe con detalle el procedimiento de traducción/adaptación de los ítems a la cultura española, se estudia la equivalencia del constructo entre la versión original y adaptada, etc. Es decir, se siguen las recomendaciones internacionales de traducción/adaptación de tests: directrices de la ITC; ver Muñiz, Elosua y Hambleton, 2013)

2.5. Desarrollo de los ítems del test:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Descripción detallada del proceso de generación de ítems, calidad de la redacción y adecuación de su formato según las directrices aceptadas (Haladyna, Downing y Rodríguez, 2002; Moreno, Martínez y Muñiz, 2006, 2015); aplicación piloto con análisis de ítems y descripción de los cambios realizados durante el proceso)

2.6. Calidad de las instrucciones para que quienes han de responder al test comprendan con facilidad la tarea:

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Claras y precisas. Muy adecuadas para las poblaciones a las que va dirigido el test, incluyendo posibles acomodaciones a poblaciones especiales cuando el test también pueda aplicarse en este tipo de poblaciones)

2.7. Calidad de las instrucciones para la administración, puntuación e interpretación del test:

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Claras y precisas. Tanto para la administración del test, como para su puntuación e interpretación)

2.8. Facilidad para registrar las respuestas:

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (El procedimiento para emitir o registrar las respuestas es muy simple por lo que se evitan los errores en la anotación)

2.9. Bibliografía del manual:

Valore si en la elaboración del test se han tenido en cuenta las teorías más aceptadas sobre el constructo y el grado en que las referencias metodológicas aportadas son adecuadas.

- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Reflejan una revisión adecuada y actualizada sobre el constructo y las referencias metodológicas que aporta son adecuadas)

2.10. Análisis de los ítems:

2.10.1. Datos sobre el análisis de los ítems:

- **N/A** ☐ Característica no aplicable para este instrumento
- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuados
- ★★ ☐ Adecuados, pero con algunas carencias
- ★★★ ☐ Adecuados
- ★★★★ ☐ Buenos
- ★★★★★ ☐ Excelentes (Información detallada sobre diversos estudios acerca de las características psicométricas de los ítems: dificultad o media, variabilidad, discriminación, validez, distractores, etc.)

2.11. Validez

Los estándares de la AERA, NCME y APA (1999, 2014) han producido un cambio fuerte en el concepto de validez: no se valida el test, sino interpretaciones o usos concretos de sus puntuaciones. No hay distintos tipos de validez (de contenido, de constructo, referida al criterio, etc.), sino un tipo único. Se aceptan, eso sí, distintas fuentes de evidencias de validez. La importancia de recoger una u otra evidencia dependerá principalmente del uso que se vaya a hacer del test. De las distintas evidencias, las tres más relevantes son las basadas: (a) en el contenido; (b) en las relaciones con otras variables (con un criterio que se pretende predecir, con otro test que mida el mismo o un constructo relacionado, etc.); y (c) en la estructura interna (como, por ejemplo, evaluando la estructura factorial). Los ítems que aparecen a continuación evalúan el grado en que las evidencias aportadas en cada caso son más o menos adecuadas. Si el manual del test usara la diferenciación clásica de distintos tipos de validez (e. g., validez de constructo o validez referida a un criterio), se deberá incorporar la información al apartado correspondiente en función del tipo de análisis realizado).

2.11.1. Evidencia basada en el contenido:

Este aspecto es especialmente esencial en los tests referidos al criterio y particularmente en los tests de rendimiento académico. Emita su juicio sobre la calidad de la representación del contenido o dominio. Si en la documentación aportada aparecen las evaluaciones de los expertos, tómelas en consideración.

2.11.1.1. Calidad de la representación del contenido o dominio:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (En la documentación se presenta una precisa definición del dominio. Los ítems muestrean adecuadamente todas las facetas del dominio. Se aporta evidencia de la validez de contenido del test definitivo)

2.11.1.2. Consultas a expertos:

Las cifras acerca del tamaño de las muestras empleadas y de los estadísticos que aparecerán más adelante tienen un carácter orientativo.

- **0** ☐ No se aporta información en la documentación
- ★ ☐ No se ha consultado a expertos sobre la representación del contenido
- ★★ ☐ Se ha consultado de manera informal a un pequeño número de expertos
- ★★★ ☐ Se ha consultado a un pequeño número de expertos mediante un procedimiento sistematizado ($N < 10$)
- ★★★★ ☐ Se ha consultado a un número moderado de expertos mediante un procedimiento sistematizado ($10 \leq N \leq 30$)
- ★★★★★ ☐ Se ha consultado a un amplio número de expertos mediante un procedimiento sistematizado ($N > 30$)

2.11.2. Evidencias basadas en la relación entre las puntuaciones del test y otras variables:

2.11.2.1. Relaciones con otras variables. Diseños y/o técnicas empleados: *(Es posible marcar más de una opción)*

- ☐ No se aporta información en la documentación
- ☐ Correlaciones con otros tests
- ☐ Diferencias entre grupos
- ☐ Matriz multirrasgo-multimétodo
- ☐ Diseños experimentales o cuasi experimentales
- ☐ Otros (Indique cuáles:).

2.11.2.1.1. Tamaño de las muestras:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña ($N < 200$)
- ★★ ☐ Un estudio con una muestra moderada ($200 \leq N \leq 500$) o varios estudios con muestras pequeñas ($N < 200$)
- ★★★ ☐ Un estudio con una muestra grande ($N > 500$)
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes

En caso de que una muestra tuviera alguna característica que pudiera justificar su tamaño reducido *(por ejemplo, su carácter clínico),* ***indíquela:***

.

2.11.2.1.2. Procedimiento de selección de las muestras:

- ☐ No se aporta información en la documentación
- ☐ Incidental
- ☐ Aleatorio, aunque las muestras no son representativas de la población objetivo
- ☐ Aleatorio, con muestras representativas de la población objetivo

Describa brevemente el procedimiento de selección:

2.11.2.1.3. Calidad de los tests marcadores empleados para evaluar las relaciones:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Se justifica adecuadamente la selección de los tests marcadores y sus propiedades psicométricas son satisfactorias)

2.11.2.1.4. Promedio de las correlaciones del test con otros tests que midan el mismo constructo o constructos con los que se esperen relaciones altas:

Se ofrecen puntos de corte para la evaluación de los coeficientes de correlación cuando se trata del mismo constructo. Dado que se esperan correlaciones de menor tamaño cuando se correlaciona el test con un constructo diferente, reduzca en 0.15 puntos los topes anteriores cuando haya de aplicarlos en esta situación.

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada ($r < 0.35$)
- ★★ ☐ Adecuada, pero con algunas carencias ($0.35 \leq r < 0.50$)
- ★★★ ☐ Adecuada ($0.50 \leq r < 0.60$)
- ★★★★ ☐ Buena ($0.60 \leq r < 0.70$)
- ★★★★★ ☐ Excelente ($r \geq 0.70$)

2.11.2.1.5. Promedio de las correlaciones del test con otros tests que midan constructos con los que el test no debería estar relacionado:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Las correlaciones estimadas con muestras de tamaño adecuado son próximas a 0, no estadísticamente significativas o, siendo significativas, los tamaños del efecto son bajos)

2.11.2.1.6. Resultados del análisis de la matriz multirrasgo-multimétodo:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Los resultados apoyan tanto la validez convergente como discriminante)

2.11.2.1.7. Resultados de las diferencias entre grupos (pueden ser grupos naturales —por ejemplo, grupos demográficos— o experimentales):

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Se establecen hipótesis de validación claras y adecuadas, se observan diferencias significativas en el sentido esperado y se presta atención al tamaño del efecto)

2.11.2.2. Evidencias basadas en las relaciones entre las puntuaciones del test y un criterio

2.11.2.2.1. Describa los criterios empleados y las características de las poblaciones:

2.11.2.2.2. Calidad de los criterios empleados:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Se justifica adecuadamente la selección del criterio y, cuando se mida mediante un test, las propiedades psicométricas del mismo son satisfactorias)

2.11.2.2.3. Atendiendo a la relación temporal entre la aplicación del test y la medida del criterio, indique el tipo de diseño: *(Es posible marcar más de una opción)*

- ☐ Retrospectivo
- ☐ Concurrente
- ☐ Predictivo

2.11.2.2.4. Tamaño de las muestras en las evidencias basadas en las relaciones con un criterio:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña ($N < 100$)
- ★★ ☐ Un estudio con una muestra moderada ($100 \leq N < 200$) o varios estudios con muestras pequeñas ($N < 100$)
- ★★★ ☐ Un estudio con una muestra grande ($N \geq 200$)
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes o estudios meta-analíticos

Indique si alguna característica de las muestras (por ejemplo, su carácter clínico) pudiera justificar el tamaño reducido de la o las muestras:

2.11.2.2.5. Procedimiento de selección de las muestras:

- ☐ No se aporta información en la documentación
- ☐ Incidental
- ☐ Aleatorio, aunque las muestras no son representativas de la población objetivo
- ☐ Aleatorio, con muestras representativas de la población objetivo

Describa brevemente el procedimiento de selección y proporcione información relevante sobre el grado de representatividad de las muestras:

2.11.2.2.6. Promedio de las correlaciones del test con los criterios:

Los rangos de valores mostrados abajo se refieren a la correlación entre el test y un criterio, que es la forma más habitual de obtener evidencias de validez referida a un criterio. Sin embargo, en ciertas situaciones clínicas, como cuando se usan tests de "screening" en un proceso diagnóstico, puede resultar más útil proporcionar información sobre la sensibilidad y especificidad del test (por ejemplo, mediante curvas ROC) que correlaciones. En estos casos, debe tener en cuenta la sensibilidad y especificidad del test a la hora de evaluar su utilidad para tomar decisiones diagnósticas y así determinar el nivel de adecuación del test, añadiendo los comentarios pertinentes en la sección 2.11.5.

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada ($r < 0.20$)
- ★★ ☐ Adecuada, pero con algunas carencias ($0.20 \leq r < 0.35$)
- ★★★ ☐ Adecuada ($0.35 \leq r < 0.45$)
- ★★★★ ☐ Buena ($0.45 \leq r < 0.55$)
- ★★★★★ ☐ Excelente ($r \geq 0.55$)

2.11.3. Evidencias basadas en la estructura interna del test:

2.11.3.1. Resultados del análisis factorial (exploratorio y/o confirmatorio):

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada
- ★★ ☐ Adecuada, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Buena
- ★★★★★ ☐ Excelente (Los resultados apoyan la estructura del test tanto en lo que se refiere al número de factores extraídos como a su interpretación. Además, se proporciona información suficiente y adecuada para evaluar la calidad de las decisiones tomadas al aplicar la técnica —AFE y/o AFC, método de factorización, rotación, software empleado, etc.— e interpretar los resultados)

2.11.3.2. Datos sobre el funcionamiento diferencial de los ítems:

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuados
- ★★ ☐ Adecuados, pero con algunas carencias
- ★★★ ☐ Adecuados
- ★★★★ ☐ Buenos
- ★★★★★ ☐ Excelentes (Información detallada sobre diversos estudios acerca del sesgo de los ítems relacionado con el sexo, la lengua materna, etc. Empleo de la metodología apropiada)

2.11.4. Indique si el manual del test informa de las acomodaciones a introducir en la administración del test, para la correcta evaluación de personas con limitaciones o diversidad funcional:

No *En caso afirmativo*, **indique cuáles y si se han justificado adecuadamente en el manual:**

2.11.5. Comentarios sobre la validez en general:

Resuma, por favor, las principales evidencias de validez que la documentación examinada aporta. Valore su calidad y justifique las puntuaciones otorgadas en las preguntas previas. En caso de que haya revisado información sobre la sensibilidad y especificidad del test (por ejemplo, mediante curvas ROC), los resultados a la hora de determinar la utilidad diagnóstica del test serán comentados en este punto. También se comentará cualquier otro tipo de evidencia diferente de las consideradas en el modelo, por ejemplo, basadas en el proceso de respuesta, si las hubiere.

2.12. Fiabilidad:

2.12.1. Datos aportados sobre la fiabilidad: *(Es posible marcar más de una opción)*

- ☐ Un único coeficiente de fiabilidad (para cada escala o subescala)
- ☐ Varios coeficientes de fiabilidad (para cada escala o subescala)
- ☐ Un único error típico de medida (para cada escala o subescala)
- ☐ Coeficientes de fiabilidad para diferentes grupos de personas
- ☐ Error típico de medida para diferentes grupos de personas
- ☐ Cuantificación del error mediante TRI (Función de información u otros)
- ☐ Otros indicadores de fiabilidad (indique cuáles):

2.12.2. Equivalencia (Formas paralelas): *Rellenar sólo si es aplicable al instrumento*

2.12.2.1. Tamaño de las muestras en los estudios de equivalencia:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña ($N < 200$)
- ★★ ☐ Un estudio con una muestra moderada ($200 \leq N < 500$) o varios estudios con muestras pequeñas ($N < 200$)
- ★★★ ☐ Un estudio con una muestra grande ($N \geq 500$)
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes

2.12.2.2. Resultados de la puesta a prueba de los supuestos de paralelismo:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuados
- ★★ ☐ Adecuados, pero con algunas carencias
- ★★★ ☐ Adecuados
- ★★★★ ☐ Buenos
- ★★★★★ ☐ Excelentes (Se realizan pruebas de significación para poner a prueba la igualdad de las medias y de las varianzas de las formas, así como la igualdad de las correlaciones con otros tests)

2.12.2.3. Promedio de los coeficientes de equivalencia:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada ($r < 0.50$)
- ★★ ☐ Adecuada, pero con algunas carencias ($0.50 \leq r < 0.60$)
- ★★★ ☐ Adecuada ($0.60 \leq r < 0.70$)
- ★★★★ ☐ Buena ($0.70 \leq r < 0.80$)
- ★★★★★ ☐ Excelente ($r \geq 0.80$)

2.12.3. Consistencia interna: *Rellenar sólo si es aplicable al instrumento*

2.12.3.1. Tamaño de las muestras en los estudios de consistencia:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña ($N < 200$)
- ★★ ☐ Un estudio con una muestra moderada ($200 \leq N < 500$) o varios estudios con muestras pequeñas ($N < 200$)
- ★★★ ☐ Un estudio con una muestra grande ($N \geq 500$)
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes

2.12.3.2. Coeficientes de consistencia interna presentados:

☐ No se aporta información
☐ Coeficiente alfa o KR-20
☐ Alfa ordinal
☐ Lambda-2
☐ Otros (indique cuáles:)

2.12.3.3. Promedio de los coeficientes de consistencia:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada ($r < 0.60$)
- ★★ ☐ Adecuada, pero con algunas carencias ($0.60 \leq r < 0.70$)
- ★★★ ☐ Adecuada ($0.70 \leq r < 0.80$)
- ★★★★ ☐ Buena ($0.80 \leq r < 0.85$)
- ★★★★★ ☐ Excelente ($r \geq 0.85$)

2.12.4. Estabilidad (Test-Retest): *Rellenar sólo si es aplicable al instrumento*

2.12.4.1. Tamaño de las muestras en los estudios de estabilidad:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña ($N < 100$)
- ★★ ☐ Un estudio con una muestra moderada ($100 \leq N < 200$) o varios estudios con muestras pequeñas ($N < 100$)
- ★★★ ☐ Un estudio con una muestra grande ($N \geq 200$)
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes

2.12.4.2. Promedio de los coeficientes de estabilidad:

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada ($r < 0.55$)
- ★★ ☐ Adecuada, pero con algunas carencias ($0.55 \leq r < 0.65$)
- ★★★ ☐ Adecuada ($0.65 \leq r < 0.75$)
- ★★★★ ☐ Buena ($0.75 \leq r < 0.80$)
- ★★★★★ ☐ Excelente ($r \geq 0.80$)

2.12.5. Cuantificación de la precisión mediante TRI: Rellenar sólo si es aplicable al instrumento

2.12.5.1. Tamaño de las muestras en los estudios de TRI:

Depende del formato de los ítems y del modelo empleado. Como referencia, en el caso de los modelos para datos dicotómicos, unas recomendaciones generales sobre el tamaño adecuado son 200 casos para el modelo de 1 parámetro, 400 para el modelo de dos parámetros y 700 para el de 3 (Parshall, Spray, Kalohn y Davey, 2002).

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un estudio con una muestra pequeña
- ★★ ☐ Un estudio con una muestra adecuada
- ★★★ ☐ Un estudio con una muestra grande
- ★★★★ ☐ Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas
- ★★★★★ ☐ Varios estudios con muestras grandes

2.12.5.2. Coeficientes proporcionados:

- ☐ No se aporta información
- ☐ Fiabilidad de las puntuaciones en el rasgo latente
- ☐ Función de Información
- ☐ Otros (indique cuáles:)

2.12.5.3. Tamaño de los coeficientes:

Debe tenerse en cuenta que el valor de los coeficientes depende del valor del rasgo latente, existiendo típicamente un rango de puntuaciones latentes para el que el test es óptimo en términos de precisión. Para la valoración del tamaño de los coeficientes, más que ese rango óptimo, se debe tener en cuenta el rango de puntuaciones para el que los resultados del test pueden tener importancia. Si no existe tal rango a priori, la evaluación debe basarse en la información promedio proporcionada (Reise y Haviland, 2005). A continuación, se proporcionan valores orientativos para la información promedio del test, si bien estos valores deben usarse con cautela por la poca experiencia existente en la aplicación de estos puntos de corte y porque dependen del número de ítems del test.

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada (información < 2)
- ★★ ☐ Adecuada, pero con algunas carencias (2 ≤ información < 3.33)
- ★★★ ☐ Adecuada (3.33 ≤ información < 5)
- ★★★★ ☐ Buena (5 ≤ información < 10)
- ★★★★★ ☐ Excelente (información ≥ 10)

2.12.6. Fiabilidad inter-jueces: Rellenar sólo si es aplicable al instrumento

2.12.6.1. Coeficientes de fiabilidad inter-jueces: *(Es posible marcar más de una opción)*

- ☐ Porcentaje de acuerdo
- ☐ Coeficiente Kappa
- ☐ Coeficiente de correlación intraclase (ICC)
- ☐ Coeficiente basado en la Teoría de la Generalizabilidad
- ☐ Otro (indique cuál:)

2.12.6.2. Valor promedio de los coeficientes de fiabilidad inter-jueces:

Se ofrecen a continuación unos puntos de corte orientativos

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Inadecuado ($r < 0.50$)
- ★★ ☐ Adecuado, pero con algunas carencias ($0.50 \leq r < 0.60$)
- ★★★ ☐ Adecuado ($0.60 \leq r < 0.70$)
- ★★★★ ☐ Bueno ($0.70 \leq r < 0.80$)
- ★★★★★ ☐ Excelente ($r \geq 0.80$)

2.12.7. Comentarios sobre la fiabilidad en general:

No olvide cumplimentar este apartado

Resuma los resultados que ha extraído de la documentación sobre la fiabilidad de las puntuaciones. Comente los distintos tipos de indicadores obtenidos, el rango de los coeficientes, si los resultados están basados en muestras adecuadas, para qué poblaciones (en función de los resultados de TRI) resulta el test más preciso, etc. Justifique las valoraciones otorgadas a las preguntas precedentes.

2.13. Baremos e interpretación de puntuaciones:

A la hora de interpretar las puntuaciones se puede diferenciar entre una interpretación normativa o una referida a un criterio. La interpretación normativa se deriva de comparar la puntuación de la persona evaluada con la distribución de las puntuaciones observadas en un grupo de referencia. La interpretación referida a un criterio o dominio requiere el establecimiento de: puntos de corte que reflejen el dominio o no; una serie de competencias o aptitudes; o, en escalas clínicas, si la persona supera un punto de corte que refleje la necesidad de una intervención, por ejemplo. A veces los puntos de corte se establecen a partir del juicio de expertos y otras a partir de investigaciones empíricas que permiten realizar clasificaciones y asignar a las personas a diferentes programas de intervención. Debe responder a uno o a los dos apartados (interpretación normativa y/o interpretación referida a un criterio), en función de la interpretación de las puntuaciones considerada en el manual.

2.13.1. Interpretación normativa: Responder sólo si es aplicable al test

2.13.1.1. Calidad de las normas:

Desde ciertas posiciones teóricas y metodologías como la tipificación continua ("continuous norming"), la generación de un número reducido de baremos no indica necesariamente una baja calidad de la información ofrecida para la interpretación de las puntuaciones. La tipificación continua utiliza la información disponible de todos los grupos para construir el baremo de cada grupo concreto, lo que resulta en baremos más precisos con grupos más reducidos (Evers, Sijtsma, Lucassen y Meijer, 2010; Zachary y Gorsuch, 1985). Tenga en cuenta esta posibilidad a la hora de emitir su valoración en este ítem.

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Un baremo que no es aplicable a la población objetivo
- ★★ ☐ Un baremo aplicable a la población objetivo con cierta precaución, considerando las diferencias entre poblaciones
- ★★★ ☐ Un baremo adecuado para la población objetivo
- ★★★★ ☐ Varios baremos dirigidos a diversos estratos poblacionales
- ★★★★★ ☐ Amplio rango de baremos en función de la edad, el sexo, el nivel cultural y otras características relevantes

2.13.1.2. Tamaño de las muestras: *(Si hay varios baremos, clasifique el tamaño promedio)*

- **0** ☐ No se aporta información en la documentación
- ★ ☐ Pequeño ($N < 150$)
- ★★ ☐ Suficiente ($150 \leq N < 300$)
- ★★★ ☐ Moderado ($300 \leq N < 600$)
- ★★★★ ☐ Grande ($600 \leq N < 1000$)
- ★★★★★ ☐ Muy grande ($N \geq 1000$)

2.13.1.3. Indique si se ha aplicado una estrategia de tipificación continua ("continuous norming") usando diferentes grupos de edad para conseguir baremos de más calidad:

No

2.13.1.4. Procedimiento de selección de las muestras:

- ☐ No se aporta información en la documentación
- ☐ Incidental
- ☐ Aleatorio, aunque las muestras no son representativas de la población objetivo
- ☐ Aleatorio, con muestras representativas de la población objetivo

Describa brevemente el procedimiento de selección:

2.13.1.5. Actualización de los baremos:

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuada (25 años o más)
- ★★ ☐ Adecuada, pero con algunas carencias (entre 20 y 24 años)
- ★★★ ☐ Adecuada (entre 15 y 19 años)
- ★★★★ ☐ Buena (entre 10 y 14 años)
- ★★★★★ ☐ Excelente (menos de 10 años)

2.13.2. Interpretación referida a un criterio: Responder sólo si es aplicable al test

2.13.2.1. Adecuación del establecimiento de los puntos de corte establecidos:

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuado
- ★★ ☐ Adecuado, pero con algunas carencias
- ★★★ ☐ Adecuada
- ★★★★ ☐ Bueno
- ★★★★★ ☐ Excelente (Se cuenta con un grupo de un mínimo de 3-4 jueces con formación y experiencia en el ámbito de estudio, y/o se proporciona evidencia empírica con estudios de calidad que relacionan el punto de corte con un criterio externo, para avalar la adecuación y utilidad de los puntos de corte establecidos)

2.13.2.2. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique el procedimiento empleado para fijar el estándar:

- ☐ Nedelsky
- ☐ Angoff
- ☐ Zieky y Livingston
- ☐ Hofstee
- ☐ Otro (indique cuál:)

2.13.2.3. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique cómo se ha obtenido el acuerdo inter-jueces: *(Es posible marcar más de una opción)*

- ☐ Coeficiente ρ_0
- ☐ Coeficiente Kappa
- ☐ Coeficiente Livingston
- ☐ Coeficiente de correlación intraclase (ICC)
- ☐ Otro (indique cuál:)

2.13.2.4. Si se utiliza el juicio de expertos para establecer los puntos de corte, indique el valor del coeficiente de acuerdo interjueces (e.g., Kappa o ICC):

- 0 ☐ No se aporta información en la documentación
- ★ ☐ Inadecuado ($r < 0.50$)
- ★★ ☐ Adecuado, pero con algunas carencias ($0.50 \leq r < 0.60$)
- ★★★ ☐ Adecuado ($0.60 \leq r < 0.70$)
- ★★★★ ☐ Bueno ($0.70 \leq r < 0.80$)
- ★★★★★ ☐ Excelente ($r \geq 0.80$)

2.13.3. Comentarios sobre los baremos y establecimientos de puntos de corte:

Resuma y evalúe los procedimientos que el test propone para facilitar la interpretación de las puntuaciones y justifique las evaluaciones dadas a las preguntas precedentes.

3

VALORACIÓN GLOBAL DEL TEST

3.1. **Con una extensión máxima de 1000 palabras, exprese su valoración del test, resaltando sus puntos fuertes y débiles, así como recomendaciones acerca de su uso en diversas áreas profesionales. Indique asimismo cuáles son las características de la prueba que podrían ser mejoradas, carencias de información en la documentación, etc.**

3.2. A modo de resumen, rellene las Tablas 1 y 2.

En la Tabla 1 incluya los datos descriptivos del test.

Tabla 1. **Descripción del test.**

Característica	Apartado	Descripción
Nombre del test	1.1	
Autor	1.3	
Autor de la adaptación española	1.4	
Fecha de la última revisión	1.9	
Constructo evaluado	1.11	
Áreas de aplicación	1.12	
Soporte	1.15	

En la Tabla 2 realice una valoración cuantitativa de las características generales del test indicando el promedio de las calificaciones emitidas en los apartados que figuran en la segunda columna de la Tabla 2. El ***número de estrellas*** que acompaña a las opciones de respuesta de los ítems se corresponde con la puntuación correspondiente, de tal modo que a una estrella le correspondería una puntuación de 1 (i.e., inadecuado), a dos estrellas le corresponde una puntuación de 2 y así sucesivamente hasta llegar a una puntuación máxima de 5 (i.e., excelente).

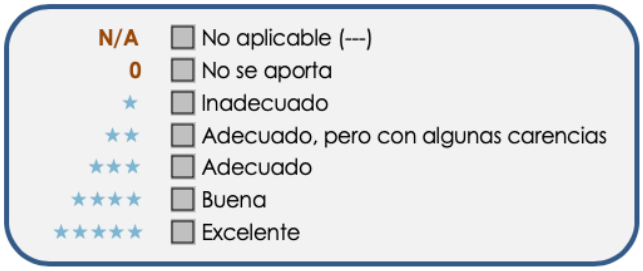

Tabla 2. **Valoración del test.**

Característica	Apartados	Puntuación Media
Materiales y documentación	2.1 y 2.2	
Fundamentación teórica	2.3	
Adaptación	2.4	
Análisis de ítems	2.10	
Validez: contenido	2.11.1	
Validez: relación con otras variables	2.11.2	
Validez: estructura interna	2.11.3	
Validez: análisis del DIF	2.11.3.2	
Fiabilidad: equivalencia	2.12.2	
Fiabilidad: consistencia interna	2.12.3	
Fiabilidad: estabilidad	2.12.4	
Fiabilidad: TRI	2.12.5	
Fiabilidad inter-jueces	2.12.6	
Baremos e interpretación de puntuaciones	2.13	

REFERENCIAS

AERA, APA y NCME. (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

AERA, APA y NCME. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

Evers, A., Sijtsma, K., Lucassen, W. y Meijer, R. R. (2010). The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results. *International Journal of Testing, 10,* 295–317.

Haladyna, T. M., Downing, S. M. y Rodríguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15,* 309-333.

Moreno, R., Martínez, R. y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2,* 65-72.

Moreno, R., Martínez, R. y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27,* 388-394.

Muñiz, J., Elosua, P. y Hambleton, R. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25,* 151-157.

Parshall, C. G., Spray, J. A., Kalohn, J. C. y Davey, T. C. (2002). *Practical considerations in computer-based testing*. New York, NY: Springer.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-72.

Reise, S. P. y Haviland, M. G. (2005). Item Response Theory and the Measurement of Clinical Change. *Journal of Personality Assessment, 84*, 228–238.

Zachary, R. A. y Gorsuch, R. L. (1985). Continuous norming: Implications for the WAIS-R. *Journal of Clinical Psychology, 41*, 86-94.