Consejo General de la Psicología
ESPAÑA

Miembro de la Federación Europea de Asociaciones de Psicólogos

Conde Peñalver, 45. 5º Izq.
28006 Madrid – España
Tel.: +34 91 444 90 20
Fax: +34 91 309 56 15
secop@cop.es
www.cop.es

Evaluación de la

Evaluación clínica de los fundamentos del lenguaje

CELF-5

Prueba valorada en la ronda de evaluaciones del año 2019

RESUMEN DE LA VALORACIÓN DEL TEST

Descripción general

Característica	Descripción
Nombre del test	CELF-5, Evaluación clínica de los fundamentos del lenguaje
Autor	Elisabeth H.Wiig, Eleonor Semel y Wayne A. Secord
Autor de la adaptación española	Departamento de I+D de Pearson Clinical & Talent Assessment: Cristina Aguilar, Ana Hernández, Érica Paradell y Frédérique Vallar
Variable(s)	Dimensiones léxico-semántica, morfosintáctica y pragmática del lenguaje
Áreas de aplicación	Aptitudes, Escalas de desarrollo, Escalas clínicas, Lenguaje
Soporte	Administración oral (acompañada con un libro de imágenes) o informatizada.

Valoración general

Característica	Valoración	Puntuación
Materiales y documentación	Buena	4
Fundamentación teórica	Buena	4
Adaptación	Excelente	5
Análisis de ítems	Adecuada, pero con algunas carencias	2
Validez: contenido	Adecuada - Buena	3,5
Validez: relación con otras variables	Adecuada - Buena	3,4
Validez: estructura interna	-	-
Validez: análisis del DIF	-	-
Fiabilidad: equivalencia	-	-
Fiabilidad: consistencia interna	Buena - Excelente	4,5
Fiabilidad: estabilidad	Adecuada, pero con algunas carencias - Adecuada	2,5
Fiabilidad: TRI	-	-
Fiabilidad inter-jueces	-	-
Baremos e interpretación de puntuaciones	Adecuada - Buena	3,3

Nota. El signo – se interpreta como que no se aporta información o bien que no procede.

Comentarios generales

La batería CELF-5 mide las aptitudes lingüísticas de niños, niñas y adolescentes (desde 5 hasta 15 años y 11 meses) e identifica y diagnostica posibles trastornos del lenguaje y la comunicación. Puede usarse para hacer una exploración completa de los diferentes fundamentos del lenguaje o sólo de los aspectos que profesionalmente se considere que son relevantes para la exploración que se esté realizando.

Era necesario un test de lenguaje tan completo como la CELF-5, en la que sólo falta la evaluación de la dimensión fonológica, y además baremado con muestras españolas ya que su anterior versión, la CELF-4, estaba baremada con muestras hispanoamericanas.

El proceso de adaptación de la CELF-5 ha sido minucioso. Las personas encargadas de realizar la adaptación y las expertas que han participado en este proceso han hecho un meritorio esfuerzo por conseguir una prueba aplicable a población española comparable en el máximo de características con la original y además poniendo especial interés en aspectos relacionados con conseguir un adecuado proceso de respuesta. No se aportan, sin embargo, suficientes detalles sobre el análisis de ítems realizado en la fase experimental de la adaptación, por lo que no se puede valorar la adecuación del proceso de selección de los mismos, aunque los argumentos que se proporcionan son convincentes. A pesar de ello, a continuación se exponen algunas limitaciones en cuanto al contenido y extensión de algunas de las pruebas.

Morfosintaxis: es una prueba excesivamente larga, pesada para los niños y las niñas. Se ha tratado de abarcar buena parte de la morfología del español, pretendiendo con ello identificar aquellos elementos en los que un/a niño/a falla para tomarlos como objetivos de intervención. Sin duda, es buena la intención, pero desde hace tiempo se sabe que los tests de lenguaje, en sus dimensiones semántica y morfosintáctica, fechan las adquisiciones mucho más tarde de lo que se da en las interacciones cotidianas. Además, en algunos ítems se exigen respuestas que contravienen las reglas normales de uso del lenguaje. Por ejemplo, si a un niño de 6 años ante un dibujo con 3 atletas corriendo se le dice: "esta chica es rápida y ésta (señalando a la que está delante de la otra)…", las normas del test proponen como correcta la respuesta de "más rápida", cuando cualquier hablante respondería sólo "más". Por otro lado, hay algún ítem como por ejemplo el ítem 28 muy difícil de responder correctamente porque probablemente resultará ambiguo hasta para una persona adulta.

Palabras relacionadas: la mayor complejidad de los ítems quizá se podría conseguir haciendo más infrecuente o más compleja la relación, y no proponiendo palabras desconocidas, que es lo que se hace en este test. En efecto, ya que se busca conocer cómo el niño o la niña es capaz de activar palabras asociadas a las palabras propuestas como estímulos, éstas últimas deberían ser conocidas por el/la niño/a porque, en caso contrario, se corre el riesgo de convertirse en una prueba de léxico y no de asociación de palabras.

En cuanto al sistema de puntuación de la batería, en la forma 1 (de 5 a 8 años y 11 meses), el índice general (Puntuación principal del lenguaje; PPL) está constituido por las mismas pruebas que el Índice de estructura lingüística (IEL) pero están expresadas en escalas diferentes (40-160 para PPL y 45-155 para IEL). Con ello se quiere proporcionar un indicador (IEL) que pueda compararse fácilmente con el resto de indicadores, y un índice que pueda utilizarse con finalidades diagnósticas (PPL). El cambio en el rango de

valores es pequeño, los errores estándar de medida similares y las correlaciones de ambas puntuaciones con otras puntuaciones y criterios parecidas, por lo que la duplicación de puntuaciones no parece necesaria.

Quizás hubiera sido conveniente incorporar la prueba Conceptos lingüísticos al Índice de lenguaje receptivo tal como sucedía en la anterior versión de la batería. Esta prueba trata de que el/la niño/a, ante demandas de la persona encargada de hacer el examen, señale figuras que contienen conceptos básicos y relaciones sintácticas. Respecto de la versión anterior, sin embargo, sí que se ha incorporado un nuevo test, Ejecución de indicaciones, que se solapa con el de Conceptos lingüísticos anteriormente comentado. De la misma forma, hubiera sido indicado que la comprensión oral de textos (en la forma 1) también se hubiera incorporado a ese Índice de lenguaje receptivo.

Debe tenerse en cuenta que la valoración de la dimensión pragmática deberá ser completada con otros instrumentos de medida, o con observaciones sistemáticas de larga duración, porque el perfil ofrecido por la CELF-5 es, en este aspecto, bastante limitado.

Otra carencia del test, comprensible por la dificultad para su valoración, es la evaluación del nivel discursivo. Este test, como todos en este ámbito, se queda en la oración, tanto en la vertiente comprensiva como productiva, sin llegar a unidades lingüísticas supraoracionales como son la narración, explicación, comparación, descripción, relaciones problema-solución, argumentación, etc. Así pues, emitir un juicio sobre el lenguaje de un niño o una niña basándose sólo en los resultados obtenidos en este test sería inapropiado. Debe conocerse el nivel discursivo, porque, en todo caso, cuando se habla del lenguaje de alguien se hace referencia a ese nivel, no a si relaciona palabras o forma oraciones con una conjunción determinada.

Esta batería se va a aplicar a personas con trastornos de lenguaje y por ello se ha mostrado un interés especial hacia niños, niñas y adolescentes con trastorno específico del lenguaje (actualmente llamado también trastorno evolutivo del lenguaje). Hay razones para considerar la CELF-5 como la mejor herramienta disponible para el diagnóstico de este trastorno y así lo atestiguan los valores de sensibilidad y especificidad, aunque sería deseable que se clarificara mejor el criterio diagnóstico que se utilizó.

Las evidencias de validez y de fiabilidad de las puntuaciones de la CELF-5 aportadas son, en términos generales, adecuadas, y aunque proceden básicamente del análisis de una misma muestra de casos, esto queda justificado por la recencia de su adaptación. También se aportan datos de estudios realizados con la versión original en muestras estadounidenses que por ello, no son directamente generalizables. Es el caso de los coeficientes de correlación con las puntuaciones de la CELF-4 y con las pruebas de vocabulario PPVT-4 y EVT-2.

La adopción de los mismos puntos de corte que en la versión original vendría apoyada por la excelente capacidad diagnóstica de la versión adaptada, aunque debería justificarse mejor el criterio diagnóstico utilizado.

Quizás el punto más destacable de la batería adaptada, a nivel aplicado, son los baremos que proporciona, con la inclusión de los niveles de confianza teniendo en cuenta los errores estándar de medida para cada grupo de edad.

La prueba tiene una gran complejidad y se ha hecho un gran esfuerzo para describirla en los dos manuales, el técnico y el de aplicación y corrección. Son de agradecer las tablas que resumen las pruebas y recursos de la batería así como la combinación de éstas en los diferentes grupos de edad, y también la figura que asocia las diferentes acciones evaluativas de la CELF-5 con diferentes preguntas que puede

hacerse la persona que vaya a administrarlo. También resulta de mucha utilidad la concreción con que se han descrito todos los aspectos que afectan a la administración de la batería y a la corrección de la misma. Sin embargo, en algunas ocasiones no se ha podido evitar que la información relevante se encuentre en capítulos diferentes, incluso de manuales distintos, y en otras ocasiones la información se repita, lo que a veces dificulta un poco su comprensión.

En la revisión realizada se ha detectado un error que afecta a la corrección de la prueba Comprensión oral de textos: en la página 16 del cuadernillo 2, Texto B (El cambio climático), la respuesta correcta dada al ítem 6 es la del ítem 7 y la de éste es la del ítem 6.

ANÁLISIS DETALLADO DE LA PRUEBA

1. DESCRIPCIÓN GENERAL DEL TEST

1.1. Nombre del test:

Evaluación Clínica de los Fundamentos del Lenguaje (CELF-5)

1.2. Nombre del test en su versión original:

Clinical Evaluation of Language Fundamentals, fifth edition

1.3. Autor/es del test original:

Elisabeth H. Wiig, Eleanor Semel, Wayne A. Secord

1.4. Autor/es de la adaptación española:

Departamento de I+D de Pearson Clinical & Talent Assessment: Cristina Aguilar, Ana Hernández, Érica Paradell y Frédérique Vallar

1.5. Editor del test en su versión original:

NCS Pearson Inc.

1.6. Editor de la adaptación española:

Pearson Educación

1.7. Fecha de publicación del test original:

2013

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

2018

1.9. Fecha de la última revisión del test:

2018

1.10. Área general de la/s variable/s que pretende medir el test:

Aptitudes, Escalas de desarrollo, Escalas clínicas, Lenguaje

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

La batería CELF-5 pretende medir las aptitudes lingüísticas de niños, niñas y adolescentes de entre 5 años y 15 años y 11 meses, con el objetivo de identificar posibles trastornos del lenguaje y la comunicación y realizar su seguimiento.

Está formada por 12 pruebas: CF (comprensión de frases), CL (conceptos lingüísticos), M (morfosintaxis), PR (palabras relacionadas), EI (ejecución de indicaciones), EF (elaboración de frases), RF (repetición de frases), COT (comprensión oral de textos), DP (definición de palabras), PP (puzle de palabras: gramática y semántica de palabras o grupos de palabras/frases), RS (relaciones semánticas), PHP (perfil de habilidades pragmáticas). Las 11 primeras se aplican directamente a niños, niñas o adolescentes y la última la cumplimenta la persona profesional encargada de administrarla a partir de la observación de las conductas verbales y no verbales y si es necesario de la información proporcionada por padres y madres, docentes u otras personas informantes.

La batería está dividida en dos partes según la edad: de 5 a 8 años y de 9 a 15. La mayoría de las pruebas que la constituyen son comunes. Sólo se diferencian en que las pruebas DP, RS y PP son exclusivas de los niños y niñas mayores, y CL es exclusiva de los niños y niñas de menos edad.

Las 11 primeras pruebas se agrupan en 5 índices generales que permiten identificar diferentes problemas en el lenguaje, describir la naturaleza del trastorno y determinar la posterior intervención: Puntuación principal de lenguaje (PPL: aptitud general lingüística), Índice de lenguaje receptivo (ILR: comprensión), Índice de lenguaje expresivo (ILE: expresión), Índice de contenido lingüístico (ICL: dimensión léxico-semántica), e Índice de estructura lingüística (IEL: que en la parte para los de más edad es llamado Índice de memoria lingüística y representa la dimensión morfosintáctica).

El recurso complementario Verificación de las habilidades pragmáticas (VHP), que permite evaluar las destrezas verbales y no verbales en contextos de interacción social, complementa al PHP, y ambos permiten diagnosticar la presencia de un trastorno pragmático del lenguaje y determinar cuáles son las necesidades específicas de mejora.

Además, la batería viene acompañada de otro recurso complementario, el Cuestionario de competencia lingüística (CCL), que en caso de administrarse puede hacerse antes o después de aplicar la batería, y que es respondido por los padres, profesores o cuidadores y proporciona una amplia imagen del rendimiento comunicativo y lingüístico (estructura, comprensión producción y lenguaje en uso) en el entorno educativo y familiar. Permite identificar déficits del lenguaje que afectan al rendimiento académico.

1.12. Áreas de aplicación:

Psicología clínica, Psicología educativa

1.13. Formato de los ítems:

Respuesta construida, elección múltiple, respuesta graduada/tipo Likert

1.14. Número de ítems:

El test tiene 12 pruebas que pueden ir dirigidas a distintos grupos de edad:

• Comprensión de frases: 26 ítems (5-8 años).

• Conceptos lingüísticos: 25 ítems (5-8 años).

• Morfosintaxis: 33 ítems (5-8 años).

• Palabras relacionadas: 40 ítems (5-15 años y 11 meses).

• Ejecución de indicaciones: 33 ítems (5-15 años y 11 meses).

• Elaboración de frases: 24 ítems (5-15 años y 11 meses)

• Repetición de frases: 26 ítems 5-15 años y 11 meses).

• Comprensión oral de textos: 20 ítems (5-15 años y 11 meses)

• Definición de palabras: 21 ítems (9-15 años).

• Puzle de palabras: 20 ítems (9-15 años y 11 meses).

• Relaciones semánticas: 20 ítems (9-15 años y 11 meses).

• Perfil habilidades pragmáticas: 48 ítems (5-15 años y 11 meses).

Contiene también dos recursos complementarios:

• Verificación de habilidades pragmáticas: 32 ítems (5-15 años y 11 meses)

• Cuestionario de competencia lingüística: 40 ítems (puede aplicarse a los niños, niñas o adolescentes, sus padres, sus madres, educadores, educadoras,…).

1.15. Soporte:

Administración oral (en algunas pruebas acompañadas de estímulos visuales) y respuestas en papel.

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

Nivel B

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

Aplicable a niños, niñas y adolescentes de entre 5 y 15 años y 11 meses de edad, en especial con trastornos de lenguaje, sean éstos primarios o secundarios. También se puede aplicar a personas cuya edad cronológica se encuentra fuera de estos rangos pero cuyo funcionamiento sea propio de un nivel de desarrollo inferior.

1.18. Existencia o no de diferentes formas del test y sus características:

La batería tiene dos formas aplicables a dos grupos de edad: entre 5 y 8 años y 11 meses, y entre 9 y 15 años y 11 meses. No hay una versión abreviada, pero según los objetivos específicos de la evaluación, la persona profesional puede decidir aplicar la batería completa o solo las pruebas y los recursos que necesite para su objetivo evaluativo.

1.19. Procedimiento de corrección:

Manual o automatizada por ordenador

1.20. Puntuaciones:

Los ítems de algunas pruebas se valoran con 0-1 puntos, los de otras pruebas con 0-2, otras con 0-3, y la de habilidades pragmáticas con 1-4 (Likert). La puntuación en cada prueba se obtiene sumando sus puntos, sin correcciones del azar, ni inversión de ítems. Las puntuaciones de cada prueba se transforman en puntuaciones escalares con media 10 y desviación típica 3. La PPL y los índices generales son puntuaciones compuestas, con media 100 y desviación típica 15, que se obtienen sumando la puntuación escalar de determinadas pruebas. A su vez, estas puntuaciones compuestas pueden transformarse en percentiles y estaninas para su mejor interpretación, con la excepción de la VHP que se interpreta de manera criterial. También se ofrece una escala de valores de desarrollo (media = 500, DT = 25) para medir el progreso de manera estandarizada.

1.21. Escalas utilizadas:

Puntuaciones estandarizadas (media 10 y desviación típica 3 para las pruebas, media 100 y desviación típica 15 para los índices generales; media 500 y desviación típica 25 para los valores de desarrollo.

Puntuaciones basadas en percentiles (Centiles)

Equivalencia en edad (años y meses)

1.22. Posibilidad de obtener informes automatizados:

Se ofrece la plataforma online Q-global que se debe activar.

A partir de un ejemplo de un informe ficticio (https://www.pearsonclinical.es/Portals/0/DocProductos/CELF-5_%20Informe_Ficticio.pdf), el informe es claro, da la información necesaria y se acompaña con un informe narrativo en el que se da la información relativa a cada puntuación.

1.23. Tiempo estimado para la aplicación del test:

Los tiempos estimados se han basado en la administración de la versión original, no la adaptada. Depende de varios factores como la edad, el nivel de aptitud, el estilo de aplicación y la experiencia de quien lo aplica. Si se aplican las 12 pruebas el tiempo puede llegar a los 90 minutos en los grupos de más edad. El tiempo estimado para poder obtener solo la PPL es de 34 minutos para edades de 5 a 8-11 y de 42 minutos para edades de 9 a 15-11. Si también se desea obtener el ILP, el tiempo promedio aumenta 16 minutos para edades de 5 a 8-11 o 9 minutos para edades de 9 a 15-11. Si también se desea obtener el ILE, el tiempo no aumenta para edades de 5 a 8-11 pero se incrementa en 12 minutos para edades de 9-0 a 15-11.

1.24. Documentación aportada por el editor:

Manual técnico, Manual de aplicación y corrección

1.25. Precio de un juego completo de la prueba:

782 € (26-1-20)

1.26. Precio y número de ejemplares del paquete de cuadernillos:

-

1.27. Precio y número de ejemplares del paquete de hojas de respuesta:

Cuadernillo de Anotación 1 (5-8:11 años) 25 usos 75,75 € (consultado el 26-1-20)

Cuadernillo de Anotación 2 (9-15:11 años) 25 usos 75,75 € (consultado el 26-1-20)

Cuestionario de Competencia, 50 unidades, 55,55 € (consultado el 26-1-20)

1.28. Precio de la administración y/o corrección, y/o elaboración de informes por parte del editor:

25 Recargas perfiles online: 52 € (consultado el 26-1-20)

2. VALORACIÓN DE LAS CARACTERÍSTICAS DEL TEST

2.1. Aspectos generales:

Contenido	Valoración	Puntuación
2.1. Calidad de los materiales del test	Buena	4
2.2. Calidad de la documentación aportada	Buena	4
2.3. Fundamentación teórica	Buena	4
2.4. Adaptación del test	Excelente	5
2.5. Desarrollo de los ítems del test	Buena	4
2.6. Calidad de las instrucciones para el participante	Excelente	5
2.7. Calidad de las instrucciones (administración, puntuación, interpretación)	Excelente	5
2.8. Facilidad para registrar las respuestas	Excelente	5
2.9. Bibliografía del manual	Excelente	5
2.10. Datos sobre el análisis de los ítems	Adecuados, pero con carencias	2

2.11. Validez:

2.11.1. Evidencias de validez de contenido:

Contenido	Valoración	Puntuación
2.11.1.1. Calidad de la representación del contenido o dominio	Buena	4
2.11.1.2. Consultas a expertos	Se ha consultado a un pequeño número de expertos mediante un procedimiento sistematizado (N < 10)	3

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

2.11.2.1. Evidencias de validez basadas en la relación entre las puntuaciones del test y otras variables:

Contenido	Valoración	Puntuación
2.11.2.1.1. Diseños empleados	Correlaciones con otros tests, Diferencias entre grupos	
2.11.2.1.2. Tamaño de las muestras	Un estudio con una muestra grande	3
2.11.2.1.3. Procedimiento de selección de las muestras	Incidental	
2.11.2.1.4. Calidad de los tests empleados como criterio o marcador	Adecuada pero con algunas carencias	2
2.11.2.1.5. Promedio de las correlaciones con otros tests que miden constructos similares	Excelente ($r \geq 0.70$)	5
2.11.2.1.6. Promedio de las correlaciones con otros tests que miden constructos no relacionados	-	-
2.11.2.1.7. Resultados de la matriz multirrasgo-multimétodo	-	-
2.11.2.1.8. Resultados de las diferencias intergrupo	Excelente	5

2.11.2.2. Evidencias de validez basadas en la relación entre las puntuaciones del test y un criterio:

Contenido	Valoración	Puntuación
2.11.2.2.1. Criterios empleados	Para la versión adaptada y con muestra española, identificación de niños, niñas y adolescentes con trastorno de lenguaje F80.2 (trastorno específico / evolutivo del lenguaje): un grupo de 89 participantes con trastorno del lenguaje y otro control del mismo tamaño y características.	
2.11.2.2.2. Calidad de los criterios empleados	Adecuada	3
2.11.2.2.3. Relación temporal entre test y criterio	Concurrente	
2.11.2.2.4. Tamaño de las muestras	Un estudio con una muestra pequeña (N < 100)	1
2.11.2.2.5. Procedimiento de selección de las muestras	Se aportan estudios obtenidos con datos estadounidenses y españoles, en este último caso con 89 casos de entre 5 y 15 años diagnosticados con un trastorno del lenguaje (es valorable el tamaño porque resulta difícil disponer de una muestra elevada de participantes con Trastornos del lenguaje) y un grupo control del mismo tamaño y características sociodemográficas. No se especifica el procedimiento de selección que se supone incidental. Con respecto al procedimiento, no se especifican claramente los instrumentos estandarizados utilizados para determinar Trastorno del lenguaje. Si por ejemplo, se utilizó el Peabody o el ITPA, habitualmente utilizados para este fin, la calidad de la selección de la muestra sería dudosa.	
2.11.2.2.6. Promedio de las correlaciones del test con los criterios	Excelente	5

2.11.3. Evidencias de validez basadas en la estructura interna:

Contenido	Valoración	Puntuación
2.11.3.1. Resultados del análisis factorial	-	-
2.11.3.2. Funcionamiento diferencial de los ítems	-	-

2.11.4. Acomodaciones en la administración del test:

Contenido	Valoración	Puntuación
2.11.4. El manual del test informa sobre las acomodaciones en la administración del test	Se indican recomendaciones claras y justificadas para aplicar las pruebas e interpretar las respuestas a personas de entornos culturales y lingüísticos diversos (por ejemplo, personas que por cualquier motivo no utilizan el español estándar), a personas con necesidades especiales como limitaciones motoras, sensoriales o cognitivas, o también a personas con edad cronológica fuera del rango de las pruebas pero con un funcionamiento propio de niveles de desarrollo inferior.	

2.11.5. Comentarios generales sobre evidencias de validez:

Se aportan varias evidencias de validez de las puntuaciones del CELF-5 que en general son adecuadas.

La relevancia y representatividad de las distintas pruebas que constituyen la batería adaptada respecto al dominio que pretende evaluar está adecuadamente documentada. Con respecto a la versión original, se han modificado algunos ítems, eliminado otros y creado nuevos, todo ello con la supervisión de las expertas que han participado en la adaptación y que han velado por la adecuación al contenido de las pruebas adaptadas.

Para la validez de la estructura interna se recurre a la interpretación de las correlaciones entre las distintas pruebas de la batería y entre éstas y los índices principales. Por lo general las correlaciones entre las diferentes pruebas e índices son, como era de esperar, moderadas-altas, excepto para el PHP, resultado que se puede explicar por el hecho diferencial del perfil respecto al resto de pruebas. Un análisis factorial de la matriz de correlaciones analizada podría aportar una visión más objetiva de la misma.

En cuanto a las evidencias de validez relativa al criterio o de pronóstico, se proporcionan resultados de sensibilidad, especificidad y valor predictivo positivo y negativo de las tres puntuaciones fundamentales del test (PPL, ILR e ILE) con diferentes puntos de corte (-1,-1,3, -1,5 y -2 desviaciones típicas). En todos los casos la sensibilidad es total o muy alta y la especificidad también es alta, entre ,91 y 1, solo en los tres primeros puntos de corte. El problema en este punto quizás se encuentre en el criterio de selección de la muestra de niños, niñas y adolescentes con trastorno de lenguaje: -1.5 desviación típica o menos en "algún" test estandarizado de aptitud lingüística. A partir de estudios sobre la práctica logopédica en España se sabe que se tiende a utilizar tests que difícilmente se pueden considerar como de "aptitud lingüística", como el Peabody y el ITPA. El primero no es más que un test de designación de la imagen, y el segundo es una mezcla de habilidades visomotoras que nada tienen que ver con el lenguaje y de pruebas lingüísticas que no serían representativas de lo que pretenden medir.

Las evidencias aportadas respecto a la relación con otras variables corresponden a estudios realizados con la versión original y muestras estadounidenses y no son directamente generalizables a la versión adaptada. Además, los tests seleccionados podrían no ser del todo adecuados para esta finalidad. Se trata del Peabody PPVT-4 (vocabulario receptivo) y el Expressive Vocabulary Test (vocabulario expresivo) y por tanto ambos tests miden solo una pequeña parte de la dimensión semántica, mientras que el CELF-5 pretende medir las dimensiones léxico-semántica, morfosintáctica y pragmática. Otra evidencia la obtienen relacionando el test que se está valorando, CELF-5, con su versión anterior, CELF-4, con la que la versión actual mantiene muchas igualdades (más allá de similitudes), pero también en este caso el estudio no se ha realizado con la versión adaptada ni con una muestra española. Hay que tener en cuenta, sin embargo, la dificultad de obtener evidencia de relación con otras variables dada la escasez de pruebas de lenguaje validadas en población española para la franja de edad a la que va dirigido el CELF-5.

No se aportan estudios de DIF con datos de la población española, cuando según se indica, en el proceso de desarrollo de la versión original sí se tuvieron en cuenta.

Finalmente, un aspecto a destacar es el hecho de que en el manual de aplicación se observa un esfuerzo por proporcionar información muy detallada acerca de las condiciones óptimas de la aplicación de las pruebas, lo que denota un interés por favorecer que el proceso de respuesta seguido por todos los informadores sea correcto.

2.12. Fiabilidad:

Contenido	Valoración	Puntuación
2.12.1. Datos aportados sobre fiabilidad	Coeficientes de fiabilidad para diferentes grupos de personas; Error típico de medida para diferentes grupos de personas.	

2.12.2. Equivalencia formas paralelas:

Contenido	Valoración	Puntuación
2.12.2.1. Tamaño de las muestras	-	-
2.12.2.2. Puesta a prueba de los supuestos de paralelismo	-	-
2.12.2.3. Promedio de coeficientes de equivalencia	-	-

2.12.3. Consistencia interna:

Contenido	Valoración	Puntuación
2.12.3.1. Tamaño de las muestras	Un estudio con una muestra grande y otro con una muestra pequeña	4
2.12.3.2. Coeficientes de consistencia interna presentados	Dos mitades	
2.12.3.3. Promedio de coeficientes de consistencia	Excelente ($r \geq 0.85$)	5

2.12.4. Estabilidad (test-retest):

Contenido	Valoración	Puntuación
2.12.4.1. Tamaño de las muestras	Un estudio con una muestra pequeña ($N < 100$)	1
2.12.4.2. Coeficientes de estabilidad	Buena ($0.75 \leq r < 0.80$)	4

2.12.5. Cuantificación de la precisión mediante TRI:

Contenido	Valoración	Puntuación

2.12.5.1. Tamaño de las muestras	-	-
2.12.5.2. Coeficientes proporcionados	-	
2.12.5.3. Tamaño de los coeficientes	-	-

2.12.6. Fiabilidad inter-jueces:

Contenido	Valoración	Puntuación
2.12.6.1. Tipos de coeficientes presentados	-	
2.12.6.2. Promedio de los coeficientes	-	-

2.12.7. Comentarios generales sobre evidencias de fiabilidad:

Se proporciona información sobre la consistencia interna y la estabilidad temporal de las diferentes puntuaciones de la CELF-5.

Los valores de consistencia interna se calcularon mediante el método de las dos mitades en una muestra de 922 casos repartidos en 13 grupos de edad (entre 40 y 94 casos por grupo) representativos de la población española según el censo del INE de 2011. Aunque el método de dos mitades es adecuado para analizar la consistencia interna, de manera complementaria podría haberse publicado también el valor del coeficiente alfa de Cronbach. Respecto a las dos mitades no se indica de qué manera se han elegido los ítems para configurarlas. La mayoría de las pruebas de la batería están compuestas por un número elevado y homogéneo de ítems, con lo que el estudio de la consistencia interna está justificado, pero su adecuación puede ser dudosa en pruebas como Morfosintaxis, porque sus 33 ítems están distribuidos en 19 grupos que evalúan, cada uno de ellos, un elemento morfológico o sintáctico diferente y además 6 de estos grupos están constituidos solo por un ítem.

Se proporcionan valores de consistencia interna para todas las pruebas y puntuaciones compuestas, tanto para el conjunto de la muestra como para cada grupo de edad estudiado. Los valores promedio, una vez transformados con la z de Fisher, oscilan entre ,79 (DP) y ,97 (PHP) en las doce pruebas y entre ,91 a ,94 en los índices generales. En un grupo de 89 casos con trastorno de lenguaje, el rango de valores del coeficiente de fiabilidad, entre ,75 y ,97, ha sido similar a los encontrados en la población general.

Un aspecto destacable es la publicación de los errores típicos de medida de las puntuaciones con promedios entre 0,50 a 1,39 para las pruebas, y de entre 3,83 a 4,46 para los índices generales. Los errores típicos ayudan a precisar la interpretación de las puntuaciones.

La estabilidad temporal ha sido evaluada en una muestra de 88 niños, niñas y adolescentes evaluados en dos ocasiones separadas por 22 días en promedio (entre 7 y 46 días). Los valores promedio corregidos por la variabilidad de la muestra, variaron entre ,64 y ,89 en las doce pruebas y entre ,79 a ,84 en los índices generales. Las puntuaciones medias de todas las pruebas e índices fueron más altas en la segunda aplicación (retest) que en la primera aplicación (test). Los tamaños del efecto para cada prueba (entre 0,13 y 0,58) y para los índices (entre 0,41 y 0,56) son en general pequeños, pero destacan los valores moderados en la prueba Comprensión oral de textos y en los índices IEL y PPL. Estos resultados apuntan, como se indica en el manual, a un posible efecto del aprendizaje en la prueba, por lo que quizás sería conveniente realizar otros estudios ampliando el tiempo

entre aplicaciones (en el estudio presentado, algunos casos repitieron la prueba solo una semana después de la primera evaluación).

En general tanto las puntuaciones de las pruebas como las de los índices generales son muy consistentes, más en este último caso como era de esperar. La estabilidad temporal de las puntuaciones es, en general, buena, aunque en este punto hay que destacar que en algunos casos la administración de las dos pruebas se realizó solo con una diferencia de una semana, lo que ha podido favorecer el aprendizaje en algunos casos tal y como dan a entender algunos tamaños del efecto. Por este motivo, la valoración de la estabilidad temporal debe realizarse con cautela.

2.13. Baremación e interpretación de las puntuaciones:

2.13.1. Interpretación normativa de las puntuaciones:

Contenido	Valoración	Puntuación
2.13.1.1. Calidad de las normas	Varios baremos dirigidos a diversos estratos poblacionales	4
2.13.1.2. Tamaño de las muestras	Pequeño (N < 150)	1
2.13.1.3. Aplicación de estrategia de tipificación continua	Si	
2.12.1.4. Procedimiento de selección de las muestras	Incidental	
2.12.1.5. Actualización de baremos	Excelente (menos de 10 años)	5

2.13.2. Interpretación referida a criterio:

Contenido	Valoración	Puntuación
2.13.2.1. Adecuación del establecimiento de los puntos de corte	Adecuado	3
2.13.2.2. Procedimiento empleado para fijar los puntos de corte	-	
2.13.2.3. Procedimiento de obtención del acuerdo inter-jueces	-	
2.13.2.4. Valor del coeficiente de acuerdo inter-jueces	-	-

2.13.3. Comentarios generales sobre baremación e interpretación de las puntuaciones:

La información sobre cómo interpretar las puntuaciones es amplia y detallada. Los baremos realizados con muestra española son útiles. Se ofrecen puntuaciones escalares (media = 10, DT = 3) de todas las pruebas y puntuaciones compuestas (media = 100, DT = 15) para los índices generales. Estos baremos incluyen intervalos de confianza de todas estas puntuaciones en 3 niveles de confianza; 68%, 90% y 95%. Los baremos de los índices generales ofrecen correspondencias de las puntuaciones compuestas con puntuaciones centiles.

También se incluyen equivalencias de edad de las puntuaciones directas de cada una de las 12 pruebas, y una tabla de correspondencias de puntuaciones escalares, puntuaciones compuestas, centiles y estaninas.

Se ofrece también valores de desarrollo (media = 500, DT = 25), muy útiles para medir el progreso del niño en las distintas pruebas y, por tanto, la eficacia de una eventual intervención. Estos valores están basados en puntuaciones directas y permiten medir el progreso de manera más precisa que dichas puntuaciones directas.

La adopción de los mismos puntos de corte que en la versión original vendría apoyada por la excelente capacidad diagnóstica de la versión adaptada, aunque, como se ha comentado en el apartado de validez, debería justificarse mejor el criterio diagnóstico utilizado.

La muestra se reparte en 13 grupos de edad, entre 5 y 15 años 11 meses. Aunque el tamaño de la muestra utilizado para realizar la baremación, en su conjunto, es elevado, al dividirla por grupos de edad queda pequeño. La muestra de tipificación se planificó para 11 grupos de edad (5-5,11 y 6-6,11) aunque, sin que los motivos sean explícitos, finalmente los dos grupos de edad más baja se subdividieron por lo que los baremos en estos casos se basan en muy pocos casos (entre 40 y 52). Sin embargo, la limitación en el tamaño de las muestras para cada grupo de edad queda compensada con el uso de una técnica de tipificación continua aplicada a datos de muestras representativas de todas las edades.

La Verificación de habilidades pragmáticas (denominada erróneamente Verificación de habilidades lingüísticas en la página 211 del manual de aplicación y corrección) se interpreta de forma criterial. Para la versión adaptada se ha adoptado el mismo punto de corte que para la versión original, propuesto éste a partir de un estudio realizado con una muestra estadounidense.