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

Evaluación Conductual de la Función Ejecutiva - Versión Infantil

BRIEF-P

Prueba valorada en la ronda de evaluaciones del año 2019

RESUMEN DE LA VALORACIÓN DEL TEST

Descripción general

Característica	Descripción
Nombre del test	Evaluación Conductual de la Función Ejecutiva -Versión Infantil-
Autor	Gerard A. Gioia, Kimberly A. Espy y Peter K. Isquith
Autor de la adaptación española	Esperanza Bausela y Tamara Luque
Variable(s)	Función ejecutiva
Áreas de aplicación	Neuropsicología, psicología clínica, psicología
Soporte	Lápiz y papel, online

Valoración general

Característica	Valoración	Puntuación
Materiales y documentación	Buena - Excelente	4,8
Fundamentación teórica	Buena - Excelente	4,5
Adaptación	Excelente	5
Análisis de ítems	Bueno	4
Validez: contenido	Adecuada - Buena	3,3
Validez: relación con otras variables	Adecuada - Buena	3,3
Validez: estructura interna	Adecuada pero con algunas carencias - Adecuada	2,5
Validez: análisis del DIF	-	-
Fiabilidad: equivalencia	-	-
Fiabilidad: consistencia interna	Excelente	5
Fiabilidad: estabilidad	Buena	4
Fiabilidad: TRI	-	-
Fiabilidad inter-jueces	-	-
Baremos e interpretación de puntuaciones	Adecuada - Buena	3,7

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

Comentarios generales

El test BRIEF-P pretende evaluar la función ejecutiva de infantes de entre 2 y 6 años mediante cinco escalas clínicas denominadas Inhibición, Flexibilidad, Control emocional, Memoria de trabajo y Planificación y organización que, a su vez, pueden agruparse en tres índices denominados de Autocontrol inhibitorio, de Flexibilidad y de Metacognición emergente, e incluso puede llegar a calcularse un índice global de función ejecutiva. El test está formado por 63 ítems que describen conductas y se contestan en una escala de respuesta de tres categorias (nunca, a veces, frecuentemente) por parte de padres, madres, cuidadores/as o profesorado en relación con un niño o una niña.

Según el manual, el test se puede utilizar como instrumento de cribado, como ayuda en el diseño y planificación de estrategias de intervención, así como para valorar el seguimiento o evolución en los ámbitos clínico, educativo, social, legal y de investigación.

Entre sus puntos fuertes destacan el hecho de que forme parte de una familia de tests BRIEF aplicables a distintas edades, el que esté disponible en distintos idiomas, la calidad de los materiales, la facilidad de administración y corrección, la buena fundamentación teórica en el desarrollo del test original, así como un excelente proceso de adaptación lingüística del test a los idiomas castellano y catalán. También se valora muy positivamente la presencia de criterios de validez del proceso de respuesta, atendiendo a los ítems no contestados o doblemente marcados, a patrones atípicos de respuesta e incluyendo dos escalas de validez, una de inconsistencia de las respuestas y otra de negatividad en el conjunto de la evaluación.

La interpretación de puntuaciones se basa en el reporte emitido tras la corrección online; un reporte claro que contiene las puntuaciones directas y las puntuaciones T obtenidas en cada escala, así como su situación respecto a dos puntos de referencia que son los valores T = 50 y T = 65. En el manual se especifica con claridad que una puntuación más elevada indica presencia de mayor problemática en todas las escalas y el valor T = 65 se trata como un umbral para considerar una puntuación como con significación clínica potencial. Aunque queda claro que se propone la interpretación relativa a un grupo normativo, dos aspectos se podrían mejorar en el futuro. Por una parte, las puntuaciones T y también las puntuaciones directas que se incluyen en el reporte serían más interpretables si se indicara explícitamente en el manual que se obtienen por simple suma de las puntuaciones dadas a las respuestas a los ítems. Por otra parte, sería deseable que se realizaran estudios con muestras amplias de personas diagnosticadas para apoyar la significación clínica potencial del umbral propuesto.

En el manual se presentan pruebas de validez de tipo cuantitativo que son adecuadas. Para justificar los tres índices en que se agrupan las cinco escalas clínicas, se proporcionan los resultados de análisis factoriales realizados en muestras de cuidadores/as y profesorado que han contestado el cuestionario en inglés, en castellano y en catalán. Se valora muy positivamente la transparencia al presentar todos los resultados y, sin embargo, se sugiere discutir en una futura edición del manual el hecho de que las estructuras factoriales de las respuestas obtenidas en España concuerdan sólo parcialmente con las estructuras factoriales obtenidas en las muestras estadounidenses.

Las evidencias de validez convergente y discriminante son adecuadas pero deberían ampliarse obteniendo más datos con muestras obtenidas en España puesto que actualmente se basan principalmente en datos obtenidos en muestras estadounidenses y en alguna muestra española de tamaño reducido. Los tests que se han usado como marcadores externos se consideran adecuados aunque sería interesante que en el manual se incluyeran datos o razonamientos que los avalen para este uso, sobre todo pensando en proporcionar datos convincentes para profesionales sin larga experiencia en el campo. El conjunto de resultados que se presentan apuntan en la dirección esperada para sustentar tanto la validez convergente como la discriminante de la escalas del BRIEF-P y, sin embargo, los comentarios en el manual se centran sólo en una parte de los datos. Sería deseable que se clarifique y se comente la interpretación de los muchos coeficientes de correlación que se publican. Los datos de comparación entre grupos clínicos y población comunitaria apuntan también en la dirección esperada pero en este caso se sugiere aumentar la potencia de las pruebas estadísticas utilizando muestras de tamaño más grande en el futuro. Por otra parte, tal como se ha indicado antes, estos datos serían muy

relevantes para apoyar la interpretación de un umbral con significación clínica potencial y para sustentar el uso de este test con el objetivo previsto de cribado.

También se valora positivamente la presentación de evidencias de fiabilidad variadas, incluyendo datos sobre consistencia interna y estabilidad temporal, así como los errores estándar de medida basados en las distintas muestras de informantes. Los valores obtenidos son entre adecuados y excelentes para la mayoría de usos del test y se valora muy positivamente la publicación de los errrores estándar de medida que transparentan la incentidumbre asociada a las puntuaciones individuales obtenidas mediante el test. Ello es muy útil en todos los ámbitos de aplicación pero especialmente en el social y el legal donde la especialización en psicometría es menos abundante. Los datos presentados son un buen apoyo a la fiabilidad por estabilidad temporal del BRIEF-P. Si además se quiere apoyar el uso de este test como medida de seguimiento o evolución sería oportuno incluir estudios de la sensibilidad al cambio de sus escalas.

Los coeficientes de acuerdo entre los dos grupos informadores, uno formado por pa/madres o cuidadores/as y otro formado por profesorado, son de tamaño reducido; un hecho que se explica adecuadamente en el manual puesto que provienen de observaciones realizadas en contextos donde las propias conductas observadas pueden ser diferentes.

Resumiendo, la información aportada en el manual del BRIEF-P sustenta en conjunto la calidad psicométrica de este test, tanto en lo que se refiere a sus contenidos como a su fiabilidad. Las mejoras más importantes se producirían al incrementar las evidencias de validez de tipo cuantitativo para avalar todos y cada uno de los usos previstos del mismo en población española.

ANÁLISIS DETALLADO DE LA PRUEBA

1. DESCRIPCIÓN GENERAL DEL TEST

1.1. Nombre del test:

BRIEF-P. Evaluación Conductual de la Función Ejecutiva -Versión Infantil

1.2. Nombre del test en su versión original:

BRIEF-P. Behavior Rating Inventory of Executive Function-Preschool Version

1.3. Autor/es del test original:

Gerard A. Gioia, Kimberly A. Espy y Peter K. Isquith

1.4. Autor/es de la adaptación española:

Esperanza Bausela y Tamara Luque

1.5. Editor del test en su versión original:

Psychological Assessment Resources (PAR)

1.6. Editor de la adaptación española:

TEA Ediciones

1.7. Fecha de publicación del test original:

2003

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

2016

1.9. Fecha de la última revisión del test:

2016

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

Escalas clínicas, función ejecutiva

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

Se pretende evaluar la función ejecutiva de infantes mediante cinco escalas clínicas: Inhibición, Flexibilidad, Control emocional, Memoria de trabajo y Planificación y organización. La primera evalúa el control de impulsos y regulación conductual; la segunda, la realización de transiciones; la tercera, la modulación de respuestas emocionales; la cuarta, el mantenimiento de información en mente, y la última, la capacidad de anticipación.

Estas cinco escalas se agrupan en tres índices: Autocontrol inhibitorio, formada por las escalas de Inhibición y Control emocional; Flexibilidad, formada por las escalas de Flexibilidad y, de nuevo, Control emocional; y Metacognición emergente, formada por las escalas Memoria de trabajo y Planificación y organización.

Las cinco escalas pueden resumirse también en un Índice global de función ejecutiva siempre que no haya discrepacias entre los tres índices.

Se incluyen también dos indicadores de validez de las respuestas. La Inconsistencia entre las respuestas proporcionadas a ítems con contenido similar, y la Negatividad, obtenida como el nivel de respuestas inusualmente negativas a varios ítems.

1.12. Áreas de aplicación:

Neuropsicología, psicología clínica, psicología educativa, psicología forense, servicios sociales, investigación

1.13. Formato de los ítems:

Respuesta graduada

1.14. Número de ítems:

63 ítems en total. Por escalas son los siguientes: Inhibición: 16, Flexibilidad: 10, Control emocional: 10, Memoria de trabajo: 17, y Planificación y organización: 10. Inconsistencia: diferencias entre 20 ítems tomados dos a dos, Negatividad: 10 ítems. Cabe destacar que para el cálculo de la inconsistencia y de la negatividad se utilizan ítems que forman parte de las escalas clínicas.

1.15. Soporte:

Papel y lápiz,online

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

Nivel C.

La aplicación y corrección puede ser realizada por personas auxiliares formadas para ello, pero la interpretación de las puntuaciones y los perfiles obtenidos debe realizarse por personas profesionales tituladas con formación específica en el área de la función ejecutiva y la evaluación mediante tests psicológicos.

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

Entre 2 y 5 años para la versión en castellano, y entre 3 y 6 años para la versión en catalán. Según el manual, el test es aplicable a niños y niñas de distintas procedencias y niveles socioeconómicos, con baremos distintos para cada idioma. Su ámbito de aplicación incluye la evaluación de niños y niñas con problemas de aprendizaje y trastornos atencionales incipientes, trastornos del lenguaje, lesiones cerebrales traumáticas, problemas relacionados con la exposición a tóxicos, trastornos generalizados del desarrollo, o del espectro del autismo, y otras alteraciones del desarrollo, neurológicas, psiquiátricas o médicas.

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

Existe una única forma o ejemplar para ser respondido por madres, padres, maestras, maestros u otras personas cuidadoras. Adicionalmente, existen otras versiones aplicables a distintos grupos de edad: BRIEF-2 para la evaluación de 5 a 18 años y BRIEF-A para la evaluación de personas adultas mayores de 18 años.

1.19. Procedimiento de corrección:

Automatizado por ordenador

1.20. Puntuaciones:

Los/as profesores/as o familiares responden directamente a los 63 Items del cuestionario. Después, la persona responsable de la aplicación deberá verificar que se han respondido a todas las frases, sin dejar ninguna en blanco, que se ha dado una sola respuesta a cada uno de los items y aclarar la veracidad de las respuestas en caso que se observen patrones atípicos. Tras ello, se procede a la corrección informatizada de la prueba. En esta corrección se invalidan aquellos cuestionarios que presentan 12 o más omisiones o respuestas dobles. También se calculan las inconsistencias, alertando cuando se observan diferencias de 8 puntos o más sobre 20, y la negatividad, alertando cuando se observa una puntuación de 4 o más respuestas extremas sobre 10. Estos puntos de corte se basan en datos obtenidos a partir de la muestra clínica española que se describe en el apartado sobre validez de este informe.

En el manual no se explicita el cálculo e interpretación de las puntuaciones directas de cada escala clínica, aunque las puntuaciones directas publicadas en los casos ilustrativos son compatibles con su obtención mediante la suma de las respuestas a cada uno de los ítems, que se contestan en una escala de tres categorías (nunca, a veces, frecuentemente). Sí se especifica claramente que una puntuación más elevada en una escala implica mayor problemática en el ámbito.

La interpretación de las puntuaciones de las escalas clínicas se hace de forma normativa y se comentará más adelante en el apartado de este informe previsto para ello.

1.21. Escalas utilizadas:

Puntuaciones T

1.22. Posibilidad de obtener informes automatizados:

Sí. En el perfil de resultados se presentan las puntuaciones directas y puntuaciones T obtenidas por el niño o niña, tanto en las escalas de validez como en cada una de las escalas clínicas e índices de la prueba. Además, las puntuaciones T se representan gráficamente diferenciando diversos tramos de puntuación mediante líneas y coloraciones.

1.23. Tiempo estimado para la aplicación del test:

15 minutos

1.24. Documentación aportada por el editor:

Manual, ejemplar con los ítems seguidos del espacio para anotar la respuesta y código de acceso a la corrección informatizada

1.25. Precio de un juego completo de la prueba:

66,20 € (13/06/19)

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

No aplicable

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

47,49 € por 25 ejemplares (13/06/19)

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

Incluido en el precio del paquete de hojas de respuesta

2. VALORACIÓN DE LAS CARACTERÍSTICAS DEL TEST

2.1. Aspectos generales:

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

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 - Excelente	4,5
2.11.1.2. Consultas a expertos	Adecuada, con algunas carencias	2

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 por sexo y edad	
2.11.2.1.2. Tamaño de las muestras	Un estudio con una muestra moderada (200 ≤ N < 500) o varios estudios con muestras pequeñas (N < 200)	2
2.11.2.1.3. Procedimiento de selección de las muestras	Incidental. La participación de padres, madres y profesionales fue voluntaria y se pidió a cada profesional información sobre su cualificación, experiencia y tipo de centro. Se presentan varios estudios para el test original con algunas muestras grandes y otras pequeñas y un estudio para la adaptación española con muestra pequeña.	
2.11.2.1.4. Calidad de los tests empleados como criterio o marcador	Buena	4
2.11.2.1.5. Promedio de las correlaciones con otros tests que miden constructos similares	Adecuada - Buena	3,5
2.11.2.1.6. Promedio de las correlaciones con otros tests que miden constructos no relacionados	Adecuada	3
2.11.2.1.7. Resultados de la matriz multirrasgo-multimétodo	-	-
2.11.2.1.8. Resultados de las diferencias intergrupo	Adecuada	3

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	personas evaluadas clínicamente vs grupo comunitario igualado por sexo y edad	
2.11.2.2.2. Calidad de los criterios empleados	Buena	4
2.11.2.2.3. Relación temporal entre test y criterio	retrospectivo	
2.11.2.2.4. Tamaño de las muestras	Varios estudios con muestras de tamaño moderado o con alguna muestra grande y otras pequeñas	4
2.11.2.2.5. Procedimiento de selección de las muestras	Incidental. No especificado el procedimiento de incorporación al estudio de las personas evaluadas clínicamente.	
2.11.2.2.6. Promedio de las correlaciones del test con los criterios	Adecuado (0.35 ≤ r < 0.45)	3

2.11.3. Evidencias de validez basadas en la estructura interna:

Contenido	Valoración	Puntuación
2.11.3.1. Resultados del análisis factorial	Adecuado, con algunas carencias	2,5
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	No	

2.11.5. Comentarios generales sobre evidencias de validez:

Se presentan evidencias de validez basadas en el contenido, el proceso de respuesta, la estructura interna y en la relación con variables externas. Cabe destacar que, como sería de esperar, en general se aportan estas evidencias dando más peso a datos obtenidos directamente con el cuestionario adaptado a nuestra cultura. La excepción son las evidencias en relación con variables externas que se apoyan mayormente en datos obtenidos durante el desarrollo del test original, siendo muy limitada la evidencia obtenida en la adaptación española y pendiente de abordar en la adaptación catalana. En este sentido, se valora muy positivamente la transparencia del manual al comentar los resultados.

En cuanto al contenido, hay una presentación clara del dominio que se va a evaluar, la función ejecutiva en infantes, y se proporciona información precisa sobre el proceso de construcción de los ítems con el objetivo de clarificar la representación el dominio. También se especifica que se consultó a especialistas aunque sería conveniente que en una próxima edición se detallara el procedimiento utilizado para ello. Los procesos de adaptación al idioma castellano y al idioma catalán están bien explicados y son convincentes en el aspecto lingüístico. En la versión catalana se echa de menos la participación de la población diana durante este proceso, un aspecto bien resuelto en la versión castellana en la que la adaptación lingüística sí incluyó el recomendable estudio piloto. Otro aspecto a tener presente es que en la adaptación al castellano se valoró la inclusión de ítems específicos redactados en español y también se eligieron de nuevo los ítems para formar parte de la escala de inconsistencia, mientras que la adaptación catalana se realizó de forma mucho más dependiente de los ítems originales. Sería interesante disponer de una discusión sobre la equivalencia lingüística de las dos versiones adaptadas más allá de la indicación que se da en el manual de que sus diferencias no afectan a su fiabilidad y validez y sólo implican que cada versión debe corregirse con sus propios baremos.

Durante la corrección del test se incluyen diversos indicadores de validez del proceso de respuesta. Además de las mencionadas escalas de inconsistencia y negatividad, se valoran los patrones inusuales de respuesta, la competencia de quien informa para hacerlo y la coherencia de los resultados obtenidos con otras informaciones externas. Todo ello ofrece garantías suficientes sobre la credibilidad del proceso de respuesta.

En relación con la estructura interna, en el manual se presentan los resultados del cuatro análisis factoriales exploratorios realizados con los datos de las muestras de pa/madres y de maestros/as en los idiomas castellano y catalán. Se parte de la hipótesis que los tres índices del BRIEF-P medirán tres factores correlacionados. El estudio del ajuste se basó en la comparación de resultados con 2, 3 y 4 factores y en el porcentaje de variancia explicada por los tres factores hipotéticos, que resultó muy elevado en ambas muestras (igual o superior al 90%), no se utilizó ningún otro de los indicadores de ajuste recomendados como por ejemplo el análisis paralelo, aunque en la versión en catalán se hicieron estudios complementarios utilizando análisis factorial confirmatorio para evaluar el ajuste del modelo unidimensional para el índice global de función ejecutiva. El dato más destacable es que la estructura interna obtenida en los estudios originales americanos se reproduce sólo parcialmente en las cuatro muestras obtenidas en nuestra cultura independientemente del idioma. Concretamente, en los cuatro análisis, las escalas de Inhibición, Memoria de trabajo y Palnificación y organización pesan de forma elevada en el factor previsto pero también presentan cargas factoriales destacables en otro factor no previsto. Este resultado tan consistentemente observado y sus posibles consecuencias deberían discutirse en el manual. Además, en este caso sería especialmente interesante complementar la discusión sobre la equivalencia métrica de los cuestionarios aportando los resultados de un análisis de funcionamiento diferencial de los ítems entre culturas.

Tal como hemos avanzado, las evidencias de validez convergente y discriminante del test se basan principalmente en estudios realizados con muestras americanas. Los resultados basados en muestras españolas deben interpretarse con precaución porque mayoritariamente son datos obtenidos con muestras reducidas. Los test que se utilizan como marcadores externos (ADHV-IV-P, CBCL/BASC y SENA) se consideran adecuados aunque sería deseable que en el manual se incluyeran explícitamente datos o razonamientos que los avalen para este uso. En cuanto a los datos presentados en las diversas tablas, se comentan sólo en parte indicando que el patrón de correlaciones entre las escalas e índices de los distintos tests es coherente con lo que cabría esperar. También se ofrece una explicación coherente para los resultados del análisis de factores principales presentados en las tablas 5.24 y 5.25, aunque sería conveniente aclarar las hipótesis previas sobre el papel de las escalas del CBCL en los distintos análisis. Sería deseable que se especificaran todas las hipótesis de relación entre las escalas del BRIEF-P y las escalas de cada uno de los tests utilizados para la validación más allá de las que se apuntan en la página 31 del manual. Ello ayudaría a la comprensión y valoración de los muchos coeficientes de correlación que se publican y que toman

valores en un rango que va desde positivos y elevados pasando por valores nulos y llegando hasta valores negativos destacables. Finalmente, se aconseja eliminar la referencia a la matriz multirasgo-multimétodo que se hace al principio de este apartado del manual. Ello evitaría posibles confusiones puesto que los comentarios incluidos en las páginas siguientes no se inspiran en los conceptos relacionados con este diseño.

La comparación de las puntuaciones por grupos de edad y sexo lleva a la conclusión de que es necesario elaborar baremos diferenciados por sexo y dos grupos de edad tanto para la evaluación parental como para la evaluación por parte de profesionales. Convendría aclarar el criterio utilizado para formar los dos grupos de edad y porqué se agrupan de diferente forma en la versión castellana y en la catalana.

Finalmente, se aportan también datos de comparación entre un grupo clínico y un grupo comunitario comparable en cuanto a sexo y edad. Sería deseable mayor concreción en relación con la calidad del diagnóstico en la muestra clínica. Los resultados que se presentan son favorables a la interpretación pretendida del test por cuanto se observan puntuaciones medias más elevadas en el grupo clínico que en el grupo comunitario en todas las escalas, tanto en las evaluaciones de pa/madres como en las de maestros/as, siendo los tamaños del efecto grandes en ambas muestras (eta cuadrado de ,36 y ,35, respectivamente). Los datos promedio utilizados para caracterizar los grupos con diagnósticos distintos van en la dirección esperada por las autoras, aunque deben interpretarse con precaución dado el reducido tamaño de las muestras, de entre 9 y 24 casos, en los que se basan.

2.12. Fiabilidad:

Contenido	Valoración	Puntuación
2.12.1. Datos aportados sobre fiabilidad	varios coeficientes de fiabilidad para cada escala, para diversos grupos de personas y errores típicos de medida	

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	Varios estudios con muestras grandes	5
2.12.3.2. Coeficientes de consistencia interna presentados	alfa de Cronbach	
2.12.3.3. Promedio de coeficientes de consistencia	Excelente (r ≥ 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 grande (N ≥ 200)	3
2.12.4.2. Coeficientes de estabilidad	Excelente (r ≥ 0.80)	5

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 valora positivamente la aportación de evidencias de fiabilidad variadas, incluyendo coeficientes de consistencia interna y de estabilidad para cada escala y para cada grupo evaluador, sean progenitores o profesorado. Se valora positivamente la inclusión del error estándar de medida asociado a cada escala y grupo evaluador.

Para el estudio de la consistencia interna, se utilizaron las cuatro muestras de tipificación de pa/madres y de profesorado en ambos idiomas. Obtuvieron buenos datos de consistencia interna con resultados del coeficiente alfa de Cronbach comprendidos entre ,72 y ,95 para progenitores, y entre ,83 y ,96 para profesorado. Estos valores de fiabilidad se reflejaron en errores estándar de medida para las puntuaciones T de entre 3 y 5 puntos para cada una de las escalas clínicas. La información contenida en el manual a este respecto permite hacerse una idea perfectamente transparente de la incertidumbre asociada a los valores posibles de una puntuación verdadera individual, lo que es muy útil en la comunicación con personas no especialistas en psicometría.

La estabilidad test-retest se calculó sólo en la adaptación al castellano con muestras más reducidas de pa/madres y profesorado y utilizando un intervalo temporal de 2 semanas entre ambas mediciones. Los coeficientes de correlación test-retest fueron excelentes, con valores que oscilaron entre ,82 y ,91 en las escalas clínicas. Se obtuvo también la comparación de medias entre las dos mediciones y éstas resultaron también estables. Tras esta revisión se considera que las conclusiones que se extrean en el manual son excesivas. Efectivamente, estos datos son un buen apoyo a la fiabilidad por estabilidad temporal del BRIEF-P y, sin embargo, habría que incluir otras evidencias si se quiere defender el uso del test como medida de seguimiento o evolución. Para este uso debería demostrarse no sólo que las medidas son estables cuando las condiciones son estables, sino también que las medidas son sensibles al cambio cuando las condiciones son cambiantes.

Finalmente, en el apartado 2.5 del manual dedidcado a sustentar la fiabilidad de la adaptación al castellano, se presentan también datos relativos al acuerdo entre evaluadores. Se trata de coeficientes de correlación entre las evaluaciones de las mismas personas realizadas por ambos grupos evaluadores. El coeficiente de correlación promedio fue de ,36, y para valorarlo hay que tener en cuenta que las distintas personas que realizan la observación lo hacen en contextos distintos, unas dentro de la escuela y otras fuera de ella, contextos en los que no es de esperar que las manifestaciones comportamentales

sean perfectamente comparables. En el manual se comenta adecuadamente este resultado, pero sugerimos que en una futura edición esta información se sitúe en un lugar del manual donde no pueda confundirse con una prueba de fiabilidad puesto que no puede considerarse que los dos tipos de informadores sean intercambiables, sino más bien complementarios.

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	Suficiente (150 < N < 300)	2
2.13.1.3. Aplicación de estrategia de tipificación continua	No	
2.13.1.4. Procedimiento de selección de las muestras	Incidental. Con criterios de inclusión detallados. La participación de padres, madres y profesionales fue voluntaria y se pidió a cada profesional información sobre su cualificación. En el manual se proporciona información sobre la composición de las muestras por sexo y edad de los niños y niñas, el nivel educativo parental y el número de hijos/as, así como un listado de las zonas geográficas de procedencia de las personas participantes .	
2.13.1.5. 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	-	-
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 intepretación de resultados se hace utilizando baremos distintos por idioma, informador/a, sexo y edad de la persona evaluada. Las muestras de tipificación para cada grupo informador son, en principio, suficientemente grandes, ya que están formadas por alrededor de 1000 casos en la adaptación castellana y de 400 en la adaptación catalana, equilibrados por sexo y edad. Sin embargo, los números por baremo se reducen a valores entre 100 y 200 casos al presentarse cuatro baremos distintos obtenidos de la clasificación por sexo y dos grupos de edad.

La representatividad de la muestra castellana está muy documentada en el manual, teniendo en cuenta no sólo el sexo y la edad, sino también el número de hermanos y hermanas y el nivel educativo de las figuras parentales, así como la distribución en toda la geografía española. En cambio, la muestra de tipificación catalana se obtuvo en una zona muy restringida y sería adecuado que en el manual se incluyera algún razonamiento encaminado a justificar su representatividad. Sería deseable aclarar cómo ambas muestras pueden representar correctamente la variabilidad del fenómeno en su ámbito teniendo en cuenta el gran esfuerzo geográfico hecho en una de las adaptaciones y el menor esfuerzo que se ha hecho en este sentido para desarrollar la otra.

También merece más explicación el hecho de que los baremos se basen en agrupaciones distintas en función de la edad en ambas adaptaciones. Efectivamente, en la adaptación castellana, un grupo incluye infantes de 2 a 3 años y el otro de 4 a 5 años. En cambio, en la adaptación catalana, un grupo es de 3 a 4 años y el otro de 5 a 6. La confusión sobre cuál sería el grupo normativo adecuado en cada caso aumenta si tenemos en cuenta que en el manual se presentan adicionalmente los baremos por grupo de edad sin distinción de sexo y que, además, se aconseja utilizar precisamente estos últimos. Todo ello no parece desprenderse directamente de los resultados de los análisis en función del sexo y la edad que se realizaron con este fin. Según los resultados presentados, ambas variables tuvieron efectos significativos pero con tamaños del efecto pequeños, de manera que no queda claro qué papel jugaron estos resultados en las agrupaciones que acabaron formando los distintos baremos.

Otro aspecto que sería importante clarificar es el umbral de una puntuación T= 65 recomendado en el manual para "considerar una puntuación como con significación clínica potencial" (página 30). En el manual se explica que dicho umbral se basa en el hecho que esta puntuación le corresponde a una persona cuyo resultado se sitúa 1,5 desviaciones estándar por encima de la media del grupo normativo. Esta explicación resulta sencilla, pero se necesita más apoyo para invocar una posible signficación clínica y en el manual se aportan datos poco conclusivos en este sentido. Los porcentajes de niños y niñas de las muestras control y clínica que obtienen puntuaciones por encima del punto de corte sugerido aportan ciertas evidencias en su apoyo, pero, a juzgar por los valores promedio, sólo superarían este criterio los grupos diagnosticados con TDAH y TEA.