¿CÓMO VALORAR TESTS PSICOMÉTRICOS?

Errores Conceptuales y Metodológicos en la Evaluación PsicoEducativa

Autores:

E. Manuel García Pérez
Ángela Magaz Lago

Psicólogos Consultores, Clínicos y Educativos

c/. Zubileta, 16; 48903. Barakaldo (Bizkaia-España)

Teléf.: 94 485 0497 Fax: 94 482 0271

e-mail: editor@gac.com.es

En la Web: www.gac.com.es

I.S.B.N.: 978-84-95180-40-7

Depósito Legal:

Imprime: RBS

Impreso en España. Printed in Spain.

Dedicatoria

Son muchas las personas a quienes nos gustaría dedicar esta obra.

Por una parte, a los profesores con quienes aprendimos psicometría, cuyos "apuntes" y enseñanzas hemos seguido[1], desde hace años para elaborar nuestros tests, y ahora para escribir este libro.

Por otra parte, a los profesionales que valoran nuestra formación posgrado, en cuyo programa se incluye un módulo de esta temática, en el cual siempre se sorprenden cuando comentamos estos contenidos.

Pero quienes más se merecen la dedicatoria son, sin ninguna duda, los padres atribulados por los problemas de sus hijos que acuden a nosotros solicitando una ayuda que exige una evaluación conductual y psicométrica en la que podamos fundamentar, con plena confianza, el mejor Plan de Acción posible.

A todos ellos y con un especial agradecimiento a quienes nos han animado a publicarla, les dedicamos esta obra.

Octubre de 2009

[1] Las profesoras Dª. Mariana Martínez, de la Universidad Complutense de Madrid y Dª. Begoña Matellanes, de la Universidad de Deusto

Presentación

Los profesionales a quienes hemos dado a revisar este libro y cuyas sugerencias hemos incluido en algún caso, han coincidido en considerar que su publicación y difusión probablemente conllevará un cierto número de críticas negativas por habernos atrevido, en la segunda parte del mismo, a hacer algo que muchos comentan en "los pasillos", pero que pocos han tomado la decisión de hacer público en artículos u otras publicaciones. A este respecto es muy llamativo que la Comisión de Tests del Consejo de Colegios Oficiales de Psicólogos, pese a su declaración de intenciones de proceder a aplicar el modelo de valoración de tests diseñado por sus miembros en el año 2000 (véase artículo resumido en la página 37), durante estos nueve años, todavía no ha procedido a llevar a cabo ninguna evaluación de los mismos.

El acuerdo entre profesores universitarios de evaluación psicométrica, profesionales con experiencia y estudiantes aventajados, es bastante común: tenemos un elevado número de tests psicométricos de escasa calidad, de base conceptual obsoleta y muy deficientemente baremados. De modo que, al igual que el escritor mencionaba en su columna del diario[2], somos conscientes de que con su publicación "vamos a ganarnos algunos enemigos…", pero confiamos en que el número de "amigos" resulte superior al de aquéllos y que su difusión contribuya a mejorar la calidad de los servicios que prestamos los psicólogos a la población general.

[2] Fernando Sánchez-Dragó, EL LOBO FEROZ, Diario El Mundo de 14 de setiembre de 2009

Índice:

Anexo: Errores conceptuales y metodológicos en la evaluación de los siguientes tests:

Carta a los lectores

Estimado/a lector/a:

Nos gustaría conocer la razón de tu interés por esta publicación, pero dada su naturaleza creemos que sólo caben tres posibilidades:

1. Eres un profesional que realiza evaluaciones psicopedagógicas (actitudes, valores, aptitudes, habilidades,…)
2. Eres un docente universitario que se ocupa de la enseñanza o de las prácticas de evaluación
3. Eres un estudiante de psicometría o de evaluación psicológica

Evidentemente caben otras, pero este libro está escrito pensando especialmente en los estudiantes, en los profesionales noveles y en los más experimentados con una elevada carga o exigencias en su trabajo; en aquel psicólogo o psicóloga que un día fuimos los autores de esta publicación y que comenzamos nuestro trabajo empleando aquellos recursos que nos enseñaron en la Facultad con una "confianza ciega" en ellos. Nunca dejaremos de reconocer el esfuerzo y la dedicación de nuestros docentes, ni apreciaremos suficientemente las enseñanzas de aquéllos que nos dieron lo mejor que tenían en esos momentos.

Ahora bien, la Universidad es quizás el "templo de la sabiduría", pero con el tiempo descubrimos que no es el "templo de la verdad". Ni siquiera llega a ser el "templo de la realidad". Han transcurrido más de veinte años desde nuestra graduación y constatamos que las enseñanzas universitarias distan notablemente de la práctica psicológica y psicoeducativa. La distancia cada vez es mayor y lo que se enseña en las Facultades de Psicología y Educación se encuentra muy alejado de la realidad o de las necesidades cotidianas del profesional.

La población a la que prestamos nuestros servicios cree que empleamos los mejores y más actuales métodos e instrumentos para llevar a cabo nuestro trabajo con la mayor eficacia

posible, desconociendo que la realidad es bien diferente: posiblemente más del 75% de los recursos conceptuales, instrumentales y metodológicos que se emplean en la evaluación Psicopedagógica tienen más de 30 años de antigüedad.

Durante los más de veintiséis años de experiencia de los autores de este texto, hemos seleccionado instrumentos de evaluación a partir de la valoración de su utilidad práctica, después de utilizarlos con múltiples casos y de realizar los ajustes pertinentes. Incluso, desde el año 1991, comenzamos a elaborar nuevos instrumentos, aprovechando los conocimientos y experiencias acumulados a lo largo de los primeros años de ejercicio independiente y de docencia con otros profesionales. Por todo ello, unido a la inquietud para contribuir a dignificar la profesión del psicólogo, disponemos de recursos para proporcionar la mejor asistencia posible a quienes acuden a nosotros con plena o moderada confianza en nuestro saber y entender. Sin embargo, cuando se es un profesional novel no se poseen esos recursos y se confía, bien en las enseñanzas de los profesores de la Universidad, bien en los usos y costumbres tradicionales, normalmente reflejados en publicaciones de todo tipo.

Uno de los problemas más acuciantes a los que nos enfrentamos a principios de los años 80, era la práctica ausencia de instrumentos de evaluación psicométrica elaborados en España, o al menos, adaptados de manera adecuada y bien baremados o tipificados con la población de nuestro país:

Bender, Boehm, McCarthy, Itpa, Wisc, Frostig, Caras, Percepción de Diferencias, Binet-Simon, Kohs, Columbia, y otras decenas más de tests, todos ellos de un antigüedad notable, desarrollados en los Estados Unidos de América de acuerdo a planteamientos conceptuales de principio, o a lo sumo de mediados del siglo pasado, eran los instrumentos que nos enseñaron, con los que practicamos y los que adquirimos para nuestro trabajo profesional.

¿Cuestionamos alguna vez estos tests? ¿Cómo íbamos a hacerlo si carecíamos de criterios o bien éstos entraban en colisión con los consejos aportados por nuestros profesores? ¿Acaso ellos en su trabajo empleaban otros diferentes? No. ¿Cuáles iban a aplicar si no había otros?

Ahora bien, ¿qué sentido tiene que en la actualidad, a punto de entrar en el segundo decenio del siglo XXI, sigamos empleando algunos instrumentos de escasísima calidad psicométrica, de base conceptual superada o falsada por la experiencia, sin validación empírica alguna, con contenidos inadecuados para la realidad social actual, mal o escasamente baremados con nuestra población, …?

Sin embargo, a pesar de las deficiencias notorias de algunos de estos tests, la mayor parte de los profesionales continúan utilizándolos sin considerar la posibilidad de elegir otros de mayores ventajas psicométricas; ahora ya sí disponibles, gracias a que algunos profesionales seguimos las sugerencias de psicólogos tan experimentados e inquietos como el que referimos a continuación:

"... *desde aquí animamos a los investigadores españoles a que generen sus propios instrumentos; y por otro lado, a la búsqueda de tradiciones de pensamiento más propias de nuestra tradición teórica y cultural con el fin de ofrecer una psicología mejor y de mejor aplicación para resolver problemas que existen en nuestro entorno físico y social. Esta especificación no quiere decir que sea de menor calidad ni con menores garantías metodológicas que las que se utilizan en otros entornos. Lo que pide es que sean relevantes y con la misma calidad metodológica posible. Cuando hemos puesto en práctica estas ideas los resultados han sido mejores con instrumentos generados desde nuestro propio contexto que con otros importados. Y pensamos que los usuarios tienen derecho a recibir los mejores servicios posibles.*" (Pelechano, 1997, p. 34)

El resultado de nuestro trabajo, durante los últimos 20 años, se concreta hoy en los Protocolos Magallanes, que pueden encontrar en la web www.protocolomagallanes.es

Este libro se elabora, publica y difunde para ayudarte a valorar los instrumentos que estás utilizando o que te preparas para empezar a utilizar en cuanto inicies tu andadura profesional. Deseamos que seas consciente de las deficiencias de algunos de ellos, lo que denominamos "Errores Conceptuales y Metodológicos", y que adquieras habilidades para seleccionar los ya existentes o los de futura aparición, de acuerdo a criterios de la máxima calidad psicométrica, de utilidad práctica y de eficacia para la elaboración de informes, psicopedagógicos o periciales. Agradeceremos todos los comentarios o valoraciones de otros tests para ir incorporándolos a la web www.preocupados.es

Los usuarios de nuestros servicios esperan y se merecen recibir la mejor asistencia profesional posible, algo a lo que además nos obliga el Código Deontológico del Psicólogo (www.cop.es/cop/codigo.htm)

Artículo 17º

> *La autoridad profesional del Psicólogo/a se fundamenta en su capacitación y cualificación para las tareas que desempeña. El/la Psicólogo/a ha de estar profesionalmente preparado y especializado en la utilización de métodos, instrumentos, técnicas y procedimientos que adopte en su trabajo. Forma parte de su trabajo el esfuerzo continuado de actualización de su competencia profesional. Debe reconocer los límites de su competencia y las limitaciones de sus técnicas.*

Uno de los principales problemas a los que se enfrenta un estudiante de Psicología, un profesional novel o un profesional excesivamente ocupado para llevar a cabo el análisis de los instrumentos que se le ofrecen en el mercado de los tests, es el hecho de que, con frecuencia, los diversos tests psicométricos se mencionan en artículos científicos y semi-científicos, en páginas bibliográficas, en lugares de promoción comercial o en sitios web de diversos profesionales o entidades, sin hacer jamás mención al año en que se publicaron. Este dato que podría alertar sobre su posible inadecuación a los casos que trata, se evita de manera sistemática en los lugares más destacados de sus webs o de las publicaciones, tanto por parte de los editores como por quienes los emplean en sus investigaciones.

Tales tests, convenientemente "remozados" en las portadas de sus manuales o en sus cuadernillos de aplicación (lo cual en algunos casos constituye una auténtica desvirtuación del test) dan la impresión a los profesionales noveles de ser instrumentos de reciente creación. Como ejemplo paradigmático de esta realidad puede visitarse el sitio web de TEA Ediciones para comprobar cómo se ha mejorado la imagen del Test de Formas Idénticas (Thurstone, L. L.) diseñado en el año 1944 y reimpreso en el año 2007 (con 65 años de antigüedad). Este fenómeno también está presente en otras empresas de tests de diversos países europeos, miembros del European Test Group, tal y como se puede comprobar al consultar sus catálogos.

En otras ocasiones, lo que aparece como novedad editorial, confundiendo al profesional que lo toma como un instrumento de reciente creación, dado su formato moderno, es un instrumento de tal antigüedad que sus planteamientos conceptuales o los metodológicos resultaron superados hace tiempo. Ejemplo de este caso lo constituye el test de Atención "**d2**", elaborado en el año 1962 y publicado en España en el año 2002 (40 años más tarde). Probablemente conoces los avances que ha habido en este tema, ligados al desarrollo de los modelos neuropsicológicos, los cuales han dejado obsoleto el citado test.

¿Se imaginan nuestras calles y autopistas llenas de vehículos de más de 30 años de antigüedad? Tales vehículos representaban el mayor avance tecnológico de la Humanidad en el campo de la locomoción, en los mismos años en que Psicólogos y empresas editoras de tests ponían a disposición de los profesionales algunos tests psicométricos.

¿Alguno de los Psicólogos acudiría a su trabajo en uno de estos vehículos si pudiera elegir hacerlo? Probablemente, NO. En tal caso, ¿por qué en el maletero de su moderno y tecnológicamente adelantado coche sigue transportando el mismo test de hace 20, 30, 40 o más años de antigüedad?

Las personas que reciben ayuda de un profesional de la Psicología o de la Educación esperan siempre recibir la mejor respuesta posible. Elegir un instrumento psicométrico con las mejores cualidades de validez (de contenido/constructo), fiabilidad y baremación poblacional es una cuestión deontológica y más aún "de conciencia". Las Facultades de Psicología y los Colegios Oficiales de Psicólogos tienen un elevado grado de responsabilidad en la difusión de este tipo de actitudes y valores en cada nueva promoción de profesionales, y en los veteranos más ocupados. Esperamos que progresivamente la asuman.

Por otra parte, el menosprecio al trabajo de desarrollo de un test lleva a muchos profesionales a la copia ilegal de los componentes esenciales del mismo, sin tener en cuenta, como frecuentemente avisan los editores de tests, del riesgo que conlleva en términos de reducción o desaparición de la investigación para el desarrollo de nuevos y mejores instrumentos. Evidentemente, no nos estamos refiriendo a la copia de hojas de

anotación o incluso a parte de los manuales, como las instrucciones de aplicación o corrección, sino a los elementos esenciales del test, como son los elementos estimulares que lo constituyen. Quizás la falta de rentabilidad económica de los tests sea en parte la explicación al retraso tecnológico de que adolece la Psicología de habla hispana. Frente al gran desarrollo de la psicometría en los Estados Unidos y Canadá, España, que podría liderar el desarrollo de los tests adecuados a la población hispano-hablante, sigue empleando instrumentos elaborados en otras épocas del conocimiento y en países con diferencias culturales que se expresan en lenguas diferentes.

Una breve historia:

La elaboración de tests psicométricos se inicia en el siglo XIX, cuando sir Francis Galton, en el transcurso de sus investigaciones sobre el papel de la herencia, comprendió la necesidad de efectuar mediciones de algunas características humanas. En 1884 estableció un Laboratorio antropométrico en la Exposición Internacional sobre Salud en Londres. Las personas podían pagar 3 peniques y conocer la medida de algunos de sus rasgos físicos y funciones sensoriales y motrices: agudeza visual y auditiva, energía muscular, tiempo de reacción,.... Con ello, consiguió medidas de un grupo suficientemente amplio y diversificado de la población de cuyo análisis pudo establecer las primeras conclusiones estadísticas.

En 1890 James McKeen Cattell empleó por primera vez en un artículo de la revista "Mind" la expresión "test mental" y el término "medidas".

En 1903, Alfred Binet publicó L'etude experimentale de l'intelligence (El Estudio Experimental de la Inteligencia), en el cual explicaba los problemas que presentaba establecer las diferencias entre los niños aventajados y los retrasados y los métodos empleados para identificarlos y evaluar sus diferencias.

La línea central de investigación de Binet fue la elaboración de un test capaz de diferenciar aquellos escolares cuyas capacidades les permitirían adaptarse al sistema educativo normal, de aquellos que necesitarían un refuerzo extra, señalando además las carencias de los mismos. En el año 1905 comenzó la aplicación de tests colectivos en las escuelas francesas.

La Primera Guerra Mundial marcó un hito importante en la historia de los tests, al conllevar la introducción por vez primera de tests de aplicación colectiva en la población. Cuando los Estados Unidos entran en guerra, se ven necesitados de reclutar soldados y suboficiales de forma rápida. En 1917 Arthur S. Otis recibe el encargo de preparar unos tests que permitan clasificar a los reclutados: los tests alfa y

beta que fueron aplicados a cerca de dos millones de personas. Se trataba de dos tests muy sencillos, de aplicación colectiva; el test alfa era una prueba verbal, adecuada para sujetos con capacidad lectora y el test beta una prueba no verbal, apropiada para sujetos con déficit de alfabetización.

Se iniciaba así la elaboración de pruebas de uso individual o grupal para la obtención de información de un sujeto sobre alguna de sus características (habilidades cognitivas, sensoriales o motrices) comparándola con un grupo social, cultural, etc., de referencia.

¿Qué es un test psicométrico?

"Un test psicométrico constituye esencialmente una medida objetiva y tipificada de una muestra de comportamiento" (Anastasi, 1978 pág. 21)

Esta definición implica rigurosamente lo siguiente:

Una Medida Objetiva, lo que significa que el método de medición debe verse afectado lo menos posible por interpretaciones del sujeto o del evaluador.

Así, un buen test se presenta a quien lo tiene que resolver de una manera muy clara y concreta, sin generar dudas sobre la tarea que se le solicita. Igualmente, los criterios de valoración de las respuestas no pueden dejar lugar a dudas por parte del evaluador sobre si la respuesta es acertada o errónea. Y, aún más, cuando una escala de un test se construye con preguntas que pueden responderse con una valoración múltiple: 0, 1, ó 2, una misma puntuación global en la escala puede obtenerse con pocas preguntas de respuesta "2" o con más preguntas de respuesta "1", lo cual supone un incumplimiento del requisito de objetividad.

Una Medida Tipificada, lo que significa que debe compararse con otras medidas de sujetos, que resulten suficientemente representativas de la población. Para ello se tienen en cuenta las puntuaciones medias y la varianza de los resultados de una muestra de referencia. Obviamente, la comparación entre puntuaciones de sujetos diferentes no debe realizarse con puntuaciones directas, sino con puntuaciones "tipificadas".

Lógicamente, los grupos de comparación no pueden ser reducidos (30, 50, 80,… sujetos), ni encontrarse todos ellos formando parte de un mismo grupo social (en el mismo colegio, la misma ciudad,…) Debe tenerse en cuenta que vamos a establecer una comparación entre los resultados obtenidos por un individuo y los obtenidos por un gran grupo de población de características similares. Por ello, resultaría poco

adecuado obtener grupos de baremación (de los que se obtienen las puntuaciones tipificadas) de cuantía reducida o procedentes de un mismo colectivo.

La Federación Europa de Asociaciones de Psicólogos (E.F.P.A.) acordó e hizo públicos (www.efpa.eu/reports) unos criterios técnicos para valorar la calidad de los tests psicométricos. Estos criterios consideran que los grupos de baremación deben estar constituidos por un número de 150 a 200 sujetos para considerarlo de calidad suficiente. Evidentemente, cuanto menos diversificadas sean las muestras de baremación, mayor debe ser el número de sujetos que las integren. Y al contrario, cuando se consiguen unas muestras de baremación muy diversificadas y aleatorizadas, más aceptable es un número reducido de las mismas.

Clases de tests psicométricos

Los tests pueden clasificarse de acuerdo con distintos y diferentes criterios. Una clasificación útil los divide en tests de:

a) **Actitudes, Valores, Preferencias**,... En este grupo se incluyen aquellos tests en los que se trata de medir lo que piensa el sujeto sobre un tema o aspecto de su vida. Se incluirían aquí los tests de Actitudes o Valores ante las relaciones sociales, como el **ADCA**; de Adaptación, como el **TAMAI** o la **EMA**; de Auto-concepto, etc.

b) **Personalidad**. Grupo en el que se incluyen aquellos tests en los que se trata de medir lo que piensa, siente y cómo actúa de manera habitual el sujeto.

Se incluirían aquí tests como los **EPQ**, **EPQ-J**, **16-PF**, **MMPI**, **CPQ**, etc. Obviamente, cada uno de estos tests solamente resulta adecuado para su empleo con el grupo de población con el que se validaron y no con otros grupos.

Un error frecuente de los evaluadores consiste en emplear tests, como el **MMPI**, validado con población psiquiátrica, con personas mentalmente sanas, con el supuesto objetivo de detectar anomalías psicopatológicas.

c) **Rendimiento** o **Aptitudes**. Grupo que incluye a todos los tests que pretenden medir el grado de habilidad o destreza que tiene el sujeto en un campo o área de comportamiento. Dado que estos tests realizan mediciones sobre muestras de comportamiento, todos ellos son tests de Habilidades. El evaluador, *no realiza una medida de la capacidad o aptitud* del sujeto, sino que, a través de la medida de su grado de habilidad, ***infiere la capacidad o aptitud*** del sujeto para realizar algo.

A su vez, los tests de rendimiento pueden sub-clasificarse en otras múltiples categorías. Como pueden ser:

Habilidades intelectuales (RAVEN, IFG, WISC, TONI-2, CERVANTES, EMIN-6)

Habilidades atencionales (CARAS, TOULOUSE-PIÈRON, EMAV)

Habilidades verbales (ITPA, PLON, TALE-2000, ELA-R)

Habilidades lingüísticas (TALE-2000)

Habilidades mecánicas

Habilidades artísticas (creatividad),

Habilidades motrices

Habilidades sensoriales

Etc…

La valoración de un test psicométrico: cómo, por qué y para qué

Todo profesional que desempeñe tareas de evaluación diagnóstica se ve en la necesidad de valorar los tests psicométricos de que dispone en el mercado profesional con la finalidad de seleccionar aquéllos que cumplan los mejores criterios para el uso al que les destina.

Si el profesional se dedica a la intervención psicoeducativa, la evaluación previa de los escolares le resulta imprescindible y, por lo tanto, también necesita valorar los instrumentos de medida.

En el campo de la psicología forense, la gran responsabilidad que conlleva la evaluación pericial de los casos en litigio, exige a los profesionales un cuidado esmerado en la elección de tests como elemento de contraste y evidencia empírica de sus conclusiones.

Ahora bien, como en todos los campos de la ciencia, la Psicometría se encuentra en desarrollo, siguiendo las pautas de los cambios de paradigmas conceptuales de la Psicología. Este fenómeno es observable retrospectivamente, al constatar cómo en una época de predominio de los postulados psicoanalíticos, los tests en uso eran todos ellos de naturaleza "proyectiva" (Fábulas de Duss, Figura Humana, H.T.P., Rorschach, CAT, TAT, test de Luscher, Wartegg,…) mientras que en los últimos 30 años, el abandono progresivo del psicoanálisis ha llevado a los profesionales al mayor empleo de tests psicométricos. Asimismo, un análisis de la realidad actual, transcultural, nos permite comprobar cómo en aquellos países donde predominan todavía los planteamientos psicodinámicos (Argentina, Chile,…) sigue predominando el empleo de tests proyectivos frente a los psicométricos.

Por ello, tanto el profesional con experiencia, como el novel, necesitan disponer de conocimientos prácticos, sencillos y rápidos para valorar la mayor o menor adecuación, a su desempeño profesional, de tests clásicos y nuevos.

El objetivo de este libro es proporcionar un resumen claro y preciso de los criterios de valoración de tests psicométricos que puedan emplear Asesores, Orientadores y Consultores, en sus actividades de Evaluación Psicoeducativa, aunque puede generalizarse a diversos campos de actuación profesional de la Psicología. Evidentemente, cualquier profesional ha cursado en sus estudios de graduación, asignaturas de Estadística, Psicometría, Psicología Experimental, Diseño y Análisis de Datos y/o similares, por lo cual, lo que aquí se expone no será novedoso, ni les resultará de difícil comprensión; únicamente exponemos una visión aplicada, concreta y práctica de sus conocimientos.

Durante los pasados años, en nuestra experiencia con posgraduados, al impartirles un curso sobre Valoración de Tests Psicométricos (Máster en Asesoramiento, Evaluación e Intervención PsicoEducativa, "ASEVINPE", del GAC), pudimos comprobar cómo los participantes agradecían el resumen conceptual y metodológico que suponía la realización de este curso. En sus comentarios, destacaron que "no habían aprendido nada nuevo", pero que "habían integrado sus conocimientos previos de un modo novedoso, sencillo y eficaz". Este es el objetivo que pretendemos con esta publicación, planteado con una finalidad: proporcionar a los usuarios de los servicios de Evaluación Psicológica la posibilidad de recibir la mejor asistencia profesional en cuanto se refiere a la evaluación diagnóstica de las dificultades de niños y adolescentes.

La mejor asistencia profesional favorece tanto al profesional, como al usuario de sus servicios, entendiendo por "mejor" la que resulta más "económica" en términos de coste/beneficio; esto es: la evaluación más rápida, segura, cómoda y, sobre todo, más fiable, aquella en cuyos resultados se puede confiar para tomar decisiones respecto a las ayudas que necesitan los escolares para mejorar su situación de desarrollo.

¿Cómo valorar un test psicométrico?

La Valoración de un test psicométrico es un procedimiento, al que se ve enfrentado en diversas ocasiones todo profesional de la Psicología, que puede llevarse a cabo con distintos criterios, dependiendo del contenido y la finalidad de uso del instrumento que se desea valorar.

De manera general, los aspectos fundamentales a tener en cuenta son los siguientes:

A. VALIDEZ **B. FIABILIDAD** **C. TIPIFICACIÓN** **D. UTILIDAD**

A. VALIDEZ: Se denomina "validez de un test" al grado en que dicho test mide lo que pretende medir. Esta es la primera condición, indispensable, que debe reunir un test de calidad: poseer una validez óptima, dado que de no ser así el resto de criterios de valoración posibles resultan insignificantes. Si un test pretende medir ansiedad, lo esperable es que mida indicadores de ansiedad. Si un test pretende medir autoestima, lo esperable es que mida indicadores de autoestima.

La validez puede estudiarse desde diversas perspectivas:

A.1. Validez de Contenido: un test posee validez de contenido cuando los elementos de cada escala corresponden fielmente a la definición operativa de cada variable a medir.

Lo deseable, aunque no siempre sucede así, es que los autores del test hagan explícita la definición operativa de la variable o variables que mide su test.

Si esta definición no se encuentra de manera explícita en su manual técnico, entonces debemos considerar si, a nuestro juicio profesional, los elementos del test corresponden realmente a manifestaciones de la variable que el profesional desea medir.

Por ejemplo: si un test se denomina de ansiedad pero incluye indicadores de depresión, no es un test válido para medir ansiedad (véase como ejemplo el **STAI,** Spilberger, C.)

Lamentablemente, algunos autores, con innegable buena fe, pero confundidos por la "fantasía de la estadística", construyen tests o escalas siguiendo un procedimiento estadístico: el denominado análisis factorial exploratorio. Este método, que no es otra cosa que un estudio múltiple de correlaciones entre elementos, puede inducir a errores muy graves, ya que puede considerar que los elementos que correlacionan entre sí constituyen una escala que mide una misma variable. Ese error es muy frecuente y se puede detectar rápidamente, si en el manual técnico analizamos el apartado en el que se explica cómo se construyó el test y encontramos que se llevó a cabo mediante análisis factorial. Este dato, añadido a la ausencia de una definición operativa de cada variable (o factor) que mide el test nos permite saber que su construcción no es válida.

Un ejemplo que usamos habitualmente para explicar este error metodológico:

Si tomamos medidas sobre el desarrollo de un niño con indicadores de lenguaje (palabras que conoce), indicadores de conocimientos culturales o de la naturaleza, habilidades de cálculo matemático,… e incluimos, por ejemplo, el tamaño de sus pies y la longitud de sus orejas,…

Al estudiar a niños de edades entre 4 y 8 años, veremos que existe una correlación muy elevada entre:

1. tamaño de los pies y tamaño de las orejas (lo cual sería un indicador de desarrollo físico)
2. amplitud de vocabulario, conocimientos culturales y cálculo aritmético (lo cual sería indicador de desarrollo curricular)

Pero nuestra sorpresa sería enorme al ver que un posible análisis factorial exploratorio podría muy bien indicar que todos los indicadores están incluidos en un único factor que

podríamos denominar "índice de desarrollo". Pues bien, a nadie se le oculta que no podemos sumar la longitud del pie con la cantidad de palabras que conoce un niño, ya que el resultado sería una "media estadística" que lejos de aportar información únicamente confunde y favorece conclusiones erróneas.

Pero tiene perfecto sentido constatar la correlación entre un posible "índice de conocimientos curriculares de vocabulario" y un "índice de desarrollo físico", ya que entre los 4 y los 8 años de edad el niño se desarrolla en ambas áreas.

A.2. Validez de Constructo: un test posee validez de constructo cuando las escalas que lo constituyen, corresponden fielmente al constructo que trata de medir, definido por los autores de test en base a los conocimientos aportados por la Psicología o las disciplinas que corresponda.

*Por ejemplo: si un test trata de medir inteligencia, entendida ésta como capacidad de razonamiento, no puede estar compuesto, como, por ejemplo, el **K-BIT**, de Kaufman por dos escalas independientes: una escala que mide vocabulario y otra que mide razonamiento.*

Por otra parte, si un test trata de medir ansiedad, que es una respuesta fisiológica, no puede incluir una escala de pensamientos depresivos.

Como puede comprobarse fácilmente, la ausencia de validez de contenido afecta gravemente a la validez de constructo.

A.3. Validez Predictiva o Criterial: si el test desea realizar una predicción de conducta o comportamiento, grado en que correlaciona con un criterio externo.

No siempre se utiliza un test para efectuar predicciones, en cuyo caso, no corresponde estudiar su validez predictiva. Sin embargo, en los casos adecuados, esta validez predictiva debe indicarse en el manual y su valor debe ser notablemente superior al 50%.

Entiéndase que una validez predictiva del 50% equivale a lanzar una moneda al aire. El uso del test se justifica si su predicción es notablemente superior a este 50%.

A.4. Validez Concurrente o Convergente: constituye un modo de valorar la validez de constructo. La validez concurrente o convergente de un test hace referencia al grado en que la medida que realiza coincide con la medida proporcionada por otro test que evalúa la misma variable por un procedimiento diferente. Un índice de validez de un nuevo test lo proporciona una alta correlación con otro test de validez ya constatada anteriormente para la misma medida. De modo opuesto se puede analizar la Validez Divergente.

Se podría cuestionar la necesidad de un nuevo test cuando se dispone ya de otro que satisface la necesidad de evaluación. Las razones son varias:

a) un test válido y fiable puede haber quedado obsoleto por aspectos meramente formales (de forma: expresiones culturales, formato de presentación,…) lo cual justifica que se elabore otro nuevo, superando al anterior en estos aspectos y acumulando las bondades del anterior o anteriores.

b) un nuevo test que evalúe lo mismo que otro anterior, válido y fiable, puede justificarse por razones de economía, ser más breve, resultar más cómoda su administración o corrección,…

En todo caso, cuando no se dispone de un test suficientemente válido y fiable, resulta innecesaria, por imposible, establecer la validez convergente del nuevo test. En estos casos, a veces se informa de la escasa validez convergente del test nuevo con respecto a otro al que se pretende sustituir por sus déficits de validez o de fiabilidad.

En ocasiones, para asegurar la validez de un test para evaluar una variable y no otra u otras, se analiza la correlación con otro/s test/s que miden otra/s variable/s y se establece una validez divergente, en términos de muy escasa correlación con ellos.

A5. Validez Discriminante: se refiere a la posibilidad de distinguir a los sujetos con respecto a la variable que mide el test.

Por ejemplo: un test de eficacia atencional debe asegurar que los sujetos que sacan puntuaciones bajas tienen menos eficacia atencional que los que obtienen puntuaciones altas. Un test de inteligencia debe asegurar que las puntuaciones bajas corresponden a sujetos de nivel intelectual significativamente por debajo de la zona media y que las altas corresponden a los situados significativamente por encima de la zona media.

La gran relevancia de esta validez se ha destacado en la X Conferencia de la Asociación Europea de Evaluación Psicológica, en setiembre de 2009 (Müler Jörg Michael, Hospital Universitario de Münster).

B. **FIABILIDAD:** la fiabilidad es el grado en que un test realiza la medición con exactitud.

El estudio de la fiabilidad siempre debe realizarse con posterioridad al estudio de la validez, entendiéndose que si el test no es válido para lo que afirma medir, ¿qué sentido tiene asegurar que es más o menos fiable? De análoga manera a la validez, la fiabilidad de un test se puede analizar de varias maneras.

B.1. Fiabilidad Test-Retest: Grado de estabilidad temporal de la medida; un indicador de que siempre mide igual cuando el constructo es estable. Un test que proporciona una medida cuando se aplica en una ocasión y otra medida cuando se aplica en otra ocasión diferente, no tiene buena fiabilidad.

B.2. Fiabilidad o Consistencia Interna: Grado en el que cada elemento de una escala del test contribuye a dar estabilidad a la medida de cada característica. Obviamente, cada test está constituido por diversos elementos. Se da por supuesto que cada elemento contribuye a medir lo mismo que los demás y que su suma constituye una medida total. Sin embargo,

pudiera ser que algún elemento no contribuya de manera significativa a la misma medida que los otros, lo que debería llevar al autor a desestimarlo.

En este aspecto conviene destacar que los valores de consistencia interna aumentan con el número de elementos del test. Por ello, para valorar la fiabilidad es fundamental estudiar los índices de homogeneidad de cada elemento (correlación de cada elemento con el total de la escala de la que forma parte). Cuando estos índices son mayores de 0,3 podemos estar seguros de la consistencia interna del instrumento.

También se suelen considerar los índices de correlación entre elementos; sin embargo, se debe tener presente que un índice de correlación entre elementos (normalmente entre pares-impares) elevado no es una garantía de fiabilidad del test, ya que puede haber muchos elementos similares, formulados de manera diferente, convenientemente situados en lugares pares o impares para asegurar un elevado índice de correlación.

Por otra parte, cuando se indican elementos *cuyos coeficientes ítem-total arrojan valores menores a 0,35 deben considerarse de baja consistencia*.

Entendida la fiabilidad de un test como la cualidad que permite confiar en sus resultados, ésta también se ve afectada por otros factores; como pueden ser:

1. Una **excesiva cantidad de elementos**, lo que hace que se llegue a los últimos con un estado de fatiga física o bien con un estado de desmotivación. Las respuestas de los últimos elementos pueden ser mucho menos fiables que los primeros, lo que afectaría a la totalidad de la prueba. Este puede ser el caso de las Escalas BASC, bien elaboradas pero de dudosa fiabilidad por la excesiva longitud de las mismas.

Además, un test con más de 15 ó 20 elementos puede aumentar la fiabilidad de forma errónea.

2. La presencia de **elementos del test formulados en términos negativos**:

No me gusta cambiar de planes ( ) Nunca ( ) A veces ( ) A menudo

Se ha comprobado en múltiples ocasiones que en castellano, la formulación negativa afecta seriamente a la comprensión del sentido en el que deben darse las respuestas. De hecho, al revisar las respuestas con los sujetos, con frecuencia éstos indican que lo entendieron en sentido contrario.

3. La inclusión en la formulación de los elementos del test **aspectos relacionados con la frecuencia o intensidad de la variable** a medir:

Frecuentemente cambio de planes ( ) Nunca ( ) A veces ( ) A menudo

Muchas veces pienso en hacer un viaje ( ) Nunca ( ) A veces ( ) A menudo

Tiene *excesiva* inquietud motora ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

Esto hace, en ocasiones, imposible responder adecuadamente a la pregunta formulada, afectado a la confiabilidad de la medida. Curiosamente, instrumentos de amplio uso entre los profesionales adolecen de esta deficiencia; tal es el caso de las Escalas de Conners y la EDAH para la detección de Déficit de Atención con Hiperactividad.

4. La presencia de **enunciados complejos, confusos, difíciles de entender** por todos los sujetos:

Sus esfuerzos se frustran fácilmente, es inconstante…

Su actitud es impropia y escasamente apreciable desde un punto de vista objetivo…

Si el enunciado no es claro y concreto, unos sujetos lo contestarán entendiéndolo de una manera y otros de otra diferente, con lo que sus resultados no serán comparables entre sí

C. TIPIFICACIÓN: la tipificación de un test es el aspecto final de su valoración psicométrica, que va ligado a su fiabilidad.

En la medida en que la tipificación-baremación es una característica del test de todo punto fundamental, afecta seriamente a la calidad psicométrica del mismo.

Todo test que intenta situar a un sujeto con relación a un grupo de referencia debe acreditar que ha sido aplicado a una muestra amplia y significativa de la población con la cual se utilizará posteriormente.

Obtener estas medidas es el proceso de baremación del test. Las muestras de población que participan en el estudio de baremación deben cumplir los siguientes requisitos:

a) ser suficientemente amplias,
b) estar diversificadas y no proceder del mismo lugar, y
c) representar adecuadamente al grupo al que pertenece la persona que será evaluada con el test.

Además de que las muestras deben estar bien definidas y ser amplias, el proceso de tipificación debe ser realizado con las máximas garantías; esto es, llevado a cabo por profesionales con experiencia en la administración de tests (evitando el empleo de estudiantes, sin experiencia previa) y siguiendo estrictamente las instrucciones de selección de sujetos y de aplicación y valoración del instrumento.

Algunos tests evalúan características que se distribuyen normalmente en la población, pero en otras ocasiones esto no es así. Por ejemplo, la distribución de los niveles de ansiedad no suele ser normal, sino asimétrica, mientras que la de la inteligencia lógica es normal. Por ello, todo test debería proporcionar como información relevante para su valoración las curvas de distribución en la población con la que se obtuvieron los baremos.

Además, es muy importante conocer si el test tiene "efecto techo". Es decir, si a partir de cierta puntuación todos los sujetos tienen la misma.

Resulta fundamental conocer con detalle las características de la muestra de tipificación, ya que eso permitirá al evaluador determinar la fiabilidad del test en unos u otros grupos de población.

Por ejemplo: no se puede utilizar un test de problemas emocionales, tipificado con muestra de población general, con sujetos de población clínica o al revés.

D. UTILIDAD PRÁCTICA: Todo test se diseña o desarrolla con cierta finalidad. Existen tests que se diseñan con fines exclusivamente de investigación, mientras que otros se diseñan para su aplicación en la elaboración de programas de intervención educativa o terapéutica.

Cada test puede considerarse de mayor o menor utilidad, en función de su finalidad. Así, un test puede tener una calidad indiscutible pero, no servir al profesional para determinada finalidad, en cuyo caso, su empleo es, sencillamente, inútil.

Los tests psicométricos se pueden emplear para conocer datos sobre una población o sobre un sujeto, siempre con una finalidad concreta:

D.1. Descriptiva: lo que se desea es conocer la distribución de puntuaciones en una o más variables.
Ejemplo: conocer los miedos o temores de los escolares de secundaria ante los exámenes

D.2. Evaluativa: lo que se desea es integrar las medidas obtenidas con el test con otras informaciones, para elaborar hipótesis explicativas de un fenómeno conductual o bien, conocer la influencia de un proceso de intervención en una o más variables.
Ejemplo: encontrar una posible explicación al bajo rendimiento escolar

D.3. Terapéutica: se desea conocer algunas características de un sujeto para diseñar una estrategia de tratamiento en función de las medidas obtenidas.

Ejemplo: conocer el nivel de intensidad de un trastorno depresivo o conocer factores predisponentes de un sujeto para padecer un trastorno por estrés

D.4. Investigadora: llevar adelante una investigación básica o empírica.

El investigador puede emplear los mismos tests que el profesional, dependiendo de la finalidad de la investigación.

EN RESUMEN...

Si el lector desea un esquema-resumen de estos criterios para valorar mínimamente un test psicotécnico, con vistas a su empleo profesional, puede utilizar los cuadros de las páginas siguientes.

VALORACIÓN DE UN TEST PSICOMÉTRICO

VALIDEZ DE CONSTRUCTO

¿Se describen operativamente las variables a medir?

¿Se informa del modelo conceptual en el que está basado el test?

VALIDEZ DE CONTENIDO

¿Los elementos de cada escala corresponden a la definición operativa de cada variable?

FIABILIDAD

¿Consistencia interna alrededor de 0.80 (para test con menos de 10 elementos)?

¿Índice de homogeneidad de cada elemento superior a 0.3, para todos, especialmente en los tests de más de 10 elementos?

¿Test-retest superior a 0.80?

TIPIFICACIÓN

¿Se describe con detalle la procedencia de la muestra?

¿Muestra superior a 150 sujetos por grupo de edad/sexo?

¿Muestra aleatorizada?

¿Muestra geográficamente amplia?

¿Muestra socio-culturalmente amplia y representativa?

¿Se proporciona descripción de la distribución de las puntuaciones en la muestra de baremación: normal/asimétrica?

UTILIDAD

¿Sirve para diseñar una estrategia de actuación?

¿Sirve para valorar la eficacia de una estrategia de actuación?

¿Comparte el test el modelo conceptual de algún Programa de Intervención?

Para los test de habilidad: ¿incluye un registro de observación de ejecución que permite explicar resultados incongruentes con otras fuentes de información?

VALIDEZ PREDICTIVA

¿Está documentada de manera fiable la correlación entre la medida del test y un criterio externo?

VALIDEZ CONCURRENTE / DIVERGENTE

¿Está documentada de manera fiable la correlación entre la medida del test y otros tests que midan la "misma variable"?

¿Está documentada de manera fiable la independencia entre la medida del test y otros tests que midan "distintas variables conceptualmente próximas"?

VALIDEZ DISCRIMINANTE

¿Está documentada de manera fiable la capacidad del test para discriminar a los sujetos con altas y bajas "puntuaciones"?

UN EJEMPLO...
de valoración de la Validez de Constructo y de Contenido

Las Escalas Magallanes de Evaluación de Hábitos Asertivos

Las Escalas Magallanes de Hábitos Asertivos se han diseñado para evaluar el grado de asertividad con el que se suele relacionar una persona a partir de los 12 años de edad. Al tratar de evaluar "comportamiento asertivo", inicialmente hay que definir el constructo "asertividad" y constatar la validez de constructo del instrumento.

Este instrumento consta de tres versiones, una que corresponde a los padres de un adolescente, en la cual, se solicita que informen de la frecuencia de diversos comportamientos de su hijo en estudio. Otra escala, correspondiente al profesorado tutor, a quien se solicita que informe de diversos comportamientos de este escolar en estudio. La última escala es un auto-informe del escolar sobre la frecuencia de diversos comportamientos suyos.

En los tres casos, al tratar de evaluar asertividad, como "aquella clase de comportamientos que constituyen un acto de respeto a uno mismo y a los demás", se ha decidido elaborar dos escalas conceptualmente diferentes entre sí: una escala "auto-asertividad" permite evaluar el grado en que el individuo actúa con respeto a sí mismo; otra escala "hetero-asertividad", permite evaluar el grado en que el individuo actúa con respecto hacia los demás.

Al disponer de dos escalas, es posible valorar el estilo de comportamiento del individuo, por comparación entre las puntuaciones obtenidas en cada escala: Pasivo, Agresivo, Pasivo-Agresivo o Asertivo[3]. Así diseñado, el instrumento tiene acreditada su "validez de constructo", ya que "está construido" con las dos escalas que permiten describir o identificar su estilo de conducta social, según el constructo "asertividad". Sin embargo, esta estructura no asegura su "validez de contenido". Para asegurar la validez de contenido se debe proceder a analizar cada uno de los elementos que constituyen cada escala.

En el caso de la escala de "auto-asertividad", todos y cada uno de los elementos deben representar una interacción que manifieste respeto a sí mismo. De manera análoga, la escala de "hetero-asertividad" debe estar compuesta por elementos que representen interacciones de respeto a los demás.

Una vez constatadas ambas condiciones, podemos asegurar la validez de contenido.

Un posterior análisis de fiabilidad de ambas escalas nos permitirá constatar la consistencia interna de cada una de ellas y el índice de homogeneidad de cada uno de los elementos, asegurando así que los elementos están adecuadamente formulados y resultan comprensibles para las personas que responderán al instrumento.

No se propone un índice general porque el constructo tiene dos componentes independientes.

[3] Para más detalles puede consultarse el manual del test, EMHAS

Escala de Auto-asertividad:

Expone y defiende sus ideas de manera respetuosa con los demás

Cuando se enfada, expresa su malestar de manera adecuada

Expresa sus sentimientos sin reparos, pero de manera adecuada

Se muestra enfadado/a o disgustado/a cuando no hace las tareas perfectamente

Cuando quiere algo, lo pide educadamente, de manera clara y directa

Le cuesta mucho cambiar de opinión

Cuando le preguntan algo que no sabe, lo reconoce sin buscar excusas o justificaciones

Felicita a sus compañeros/as cuando hacen algo que le gusta

Cuando no entiende algo, se enfada bastante

Cuando se equivoca busca excusas para justificarse

Cuando tiene alguna duda pregunta de manera correcta

Si necesita ayuda la pide con buenos modales

Cuando hace algo que perjudica a algún/a compañero/a lo reconoce y pide disculpas

Actúa de manera independiente, sin dejarse llevar por lo que hagan o digan los demás

Cuando cree que le critican injustamente, se defiende de manera adecuada

Sabe negarse, cortésmente, cuando los/as compañeros/as le piden que haga algo que no quiere hacer

Escala de Hetero-asertividad:

Suele criticar a los/as demás cuando se equivocan, olvidan algo o hacen algo mal

Admite, sin enfadarse, que los demás le nieguen algo que les pide

Protesta o se enfada cuando le critican

Cuando dice algo que los demás no entienden, vuelve a explicarlo sin enfadarse

Se enfada, cuando los demás le llevan la contraria

Admite, sin enfadarse, que los demás cambien de opinión

Cuando le piden cosas, responde con buenos modales

Se molesta cuando se le hacen preguntas personales

Extracto de los criterios propuestos por la Comisión de Tests:

UN MODELO PARA EVALUAR LA CALIDAD DE LOS TESTS UTILIZADOS EN ESPAÑA

Gerardo Prieto* y José Muñiz**

** Universidad de Salamanca. ** Universidad de Oviedo*

Papeles del Psicólogo: Revista del Colegio Oficial de Psicólogos. Año 2000. Número 77, pp. 65-75

Resumen

En el ejercicio de su profesión los psicólogos utilizan con frecuencia los tests para la obtención de datos. Para que los datos así obtenidos sean de calidad y puedan ayudar al psicólogo a tomar las decisiones adecuadas, los tests utilizados han de reunir las propiedades técnicas oportunas. Un medio eficaz para mejorar los tests y su práctica es ofrecer a los usuarios toda la información posible acerca de su calidad y características. El Colegio Oficial de Psicólogos tiene la intención de promover un proceso de evaluación de los tests más utilizados en España para proporcionar así una información técnica precisa a los usuarios de las pruebas. En este artículo se describen las características del procedimiento de evaluación y se presenta un modelo de cuestionario para llevar a cabo la evaluación.

Los tests psicológicos constituyen una de las herramientas más importantes al servicio de la práctica profesional y de la investigación de los psicólogos. Como ocurre con cualquier otra tecnología científica, los tests pueden utilizarse de forma correcta o incorrecta. La Comisión de Tests se creó para promocionar y potenciar el uso adecuado de los tests en nuestro país. Para llevar a cabo esta tarea, la Comisión funciona de forma coordinada con otros grupos de trabajo internacionales con fines similares, tales como la Federación Europea de Asociaciones de Psicólogos Profesionales (EFPPA), o la Comisión Internacional de Tests (ITC).

En nuestro país las informaciones acerca de las características de los tests no son exhaustivas, están muy dispersas, no son fácilmente accesibles o, al ser facilitadas por el editor del test, no pueden ser consideradas independientes e imparciales. Por este motivo, los profesionales carecen en muchas ocasiones de orientaciones científicas que les permitan seleccionar el instrumento más apropiado para sus objetivos. Consciente de este problema, la Comisión de Tests del COP ha diseñado un modelo estandarizado de evaluación de tests (CET) con la finalidad de revisar los tests empleados en nuestro país, al objeto de informar a los usuarios sobre su calidad técnica. Puesto que se dispone de datos recientes bastante fiables sobre los tests más usados en los distintos campos profesionales de la psicología (Muñiz y Fernández-Hermida, 2000), es intención de la Comisión de Tests focalizar inicialmente en estas pruebas las revisiones, aunque también sería posible incluir otros tests en la primera fase de evaluación a instancias de editores y autores.

El objetivo de este artículo es la descripción de un posible proceso de revisión a seguir, y la presentación del modelo de cuestionario que se utilizaría en las revisiones de los tests. Consideramos que la publicación del sistema de evaluación y de los criterios de evaluación generará un beneficio colateral añadido a la información técnica facilitada, a saber, informar a los autores de tests de los estándares de calidad que han de poseer las nuevas pruebas que aparezcan en el mercado. Tanto el sistema de evaluación que se describe a continuación, como el modelo de cuestionario presentado, se han inspirado principalmente en los modelos inglés y holandés actualmente en funcionamiento.

./…

Conclusiones

La mayoría de los expertos, así como las organizaciones profesionales nacionales e internacionales coinciden en señalar que **una de las medidas más eficaces para mejorar la utilización que se hace de los tests** es la de **proporcionar una buena información y formación a los usuarios**.

Un profesional con una buena información sobre las pruebas y una formación adecuada difícilmente utilizará de forma incorrecta los tests.

Enmarcado en esa filosofía, en 1995 [4] el Colegio Oficial de Psicólogos (COP) creó una Comisión de Tests (http://www.cop.es/tests/) con el fin de analizar los problemas implicados en el uso de los tests.

./…

[4] En el mes de octubre de 2009 (14 años más tarde) está Comisión todavía no ha valorado ni un solo test.

CUESTIONARIO PARA LA EVALUACIÓN DE LOS TESTS (CET)

1. Descripción general del test

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 en su adaptación española:
1.10. Clasifique el área general de la o las variables que pretende medir el test

( ) Inteligencia
( ) Aptitudes
( ) Habilidades y Rendimiento académico
( ) Psicomotricidad
( ) Alteraciones neuropsicológicas
( ) Personalidad
( ) Motivación
( ) Actitudes
( ) Intereses
( ) Valores
( ) Otros (Indique cuál:.....)

1.11. Breve descripción de la variable o variables que pretende medir el test:
1.12. Área de aplicación

( ) Psicología clínica
( ) Psicología educativa
( ) Neuropsicología
( ) Psicología forense
( ) Psicología del trabajo y las organizaciones
( ) Psicología del deporte
( ) Servicios sociales
( ) Psicología del Tráfico
( ) Otros (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, discapacitados, grupos clínicos, etc.):

1.18. Indique si existen diferentes formas del test y sus características (formas paralelas, versiones abreviadas, versiones informatizadas o impresas, etc). En el caso de que existan versiones informatizadas, describa los requisitos mínimos del *hardware* y *software*.

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 gran calidad, software muy atractivo y eficiente)

2.2. Calidad de la documentación aportada

2.3. Fundamentación teórica

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

2.5. Calidad de las instrucciones

2.10. Validez

2.10.1. **Validez de contenido**:

2.10.1.1. Calidad de la representación del contenido o dominio:

2.10.1.2. Consultas a expertos:

2.10.2. **Validez de constructo**:

2.10.2.2. Tamaño de las muestras en la validación de constructo:

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<200)
** ( ) Un estudio con una muestra moderada (200≤N<500)
*** ( ) Un estudio con una muestra grande (N≥500.)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.10.2.3. Procedimiento de selección de las muestras*:

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio

2.10.3. **Validez predictiva**

2.10.4. Comentarios sobre la validez en general:

2.11. Fiabilidad

2.11.3. **Consistencia interna**

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

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<200)
** ( ) Un estudio con una muestra moderada (200≤N<500)
*** ( ) Un estudio con una muestra grande (N≥500)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.11.4. **Estabilidad (Test-Retest)**

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

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña (N<100)
** ( ) Un estudio con una muestra moderada (100≤N<200)
*** ( ) Un estudio con una muestra grande (N≥200)
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.11. 5 Comentarios sobre la fiabilidad en general:

2.12. Normas

2.12.1. Calidad de las normas:

( ) 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
*** ( ) 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

2.12.2. Tamaño de las muestras:

( ) No se aporta información en la documentación
* ( ) Pequeño (N<150)
** ( ) Suficiente (150≤N<300)
*** ( ) Moderado (300≤N<600)
**** ( ) Grande (600≤N<1000)
***** ( ) Muy grande (N≥1000)

2.12.3. Procedimiento de selección de las muestras*:

( ) No se aporta información en la documentación
( ) Incidental
( ) Aleatorio

ANEXOS

Por su posible interés, a continuación incluimos diversas valoraciones de tests de mayor o menor uso en evaluación psicoeducativa. Las valoraciones se han efectuado por profesionales con experiencia en psicometría y en evaluación, pero las opiniones son responsabilidad exclusiva de los autores de la presente obra.

El sitio web www.preocupados.es está dedicado a la problemática del Fracaso Escolar, cuestión ésta que preocupa a padres, profesionales, interesados y autoridades educativas de todos los niveles.

En este sitio web puede acceder al Modelo Explicativo del Éxito-Fracaso Escolar (modelo D.S.I.) y a más valoraciones de tests psicométricos. Con los documentos en formato pdf que se encuentran en la actualidad y que se puedan añadir en el futuro, podrá ir ampliando y actualizando los contenidos de la presente publicación.

Con esta publicación confiamos en haber contribuido a la mejor información de los profesionales de la evaluación psicoeducativa, tanto expertos, como noveles.

Cualesquiera dudas o comentarios al respecto pueden dirigirlos a la División de Publicaciones del **Grupo ALBOR-COHS**. Las respuestas podrán ser incorporadas en el sitio web anteriormente citado.

Escala de Ansiedad Infantil revisada

Revised Children´s Manifest Scale: RCMAS

(Reynolds, C.R. y Richmond, B.O., 1978)

Adaptación al castellano de C.D. Sosa, J.I. Capafóns, P. Flores, A.M. Navarro y F. Silva. Madrid: MEPSA

Los autores de la adaptación española de esta Escala (revisada en 1978) informan de que está compuesta de "28 ítems en forma de afirmaciones, expresados en la dirección de la ansiedad". *En la escala se intenta recoger tanto aspectos motores como fisiológicos y cognitivos que tengan relación con el constructo ansiedad*".

A este respecto cabe indicar lo siguiente:

a) Para hacer tales afirmaciones los autores deberían definir previamente la ansiedad, cosa que no hacen en el manual de la adaptación española.

b) La ansiedad se considera universalmente una respuesta de tipo fisiológico exclusivamente, lo cual no quiere decir que no esté presente simultáneamente con cogniciones y acciones motrices.

El hecho de haya algunas cogniciones frecuentemente asociadas a la ansiedad no justifica, desde el punto de vista de la validez de constructo, que se puntúe una escala de Ansiedad con afirmaciones de tipo cognitivo tales como:

Me cuesta decidirme (más propio de un estado depresivo que de un estado ansioso)

Parece que los demás hacen las cosas con más facilidad que yo (¿qué tiene que ver esto con la ansiedad?)

Me preocupo mucho (las preocupaciones se dan en las personas sin que sea necesario que tengan ansiedad)

Me enfado con mucha facilidad (un indicador de irritabiliad, no de ansiedad)

Me preocupa lo que los demás piensen de mi (de nuevo preocupaciones que pueden aparecer sin estados de ansiedad)

Otros niños son más felices que yo (eso no quiere decir que el niño tiene ansiedad)

Muchas personas están en contra mía (no es un indicador de ansiedad)

Bien, no vamos a continuar: la totalidad de los ítems se puede consultar en la web www.preocupados.es

Como puede comprobarse con facilidad, esta "supuesta escala para evaluar ansiedad" contiene poquísimos indicadores fisiológicos de ansiedad, como máximo los siguientes: ítem 2 (ansiedad situacional y no

permanente), ítems 4 y 6 (también situacional e incluyendo el miedo), ítem 10 (que puede ser por preocupaciones no por ansiedad), ítem15 (claro indicador de ansiedad), ítems 25 y 26. En total 7 indicadores fisiológicos frente a 21 indicadores que no son manifestaciones unívocas de ansiedad.

Véase pues, cómo esta escala de ansiedad carece absolutamente de validez de contenido y, por lo tanto, de también de constructo. Toda vez que, incluso, suma indicadores fisiológicos con indicadores cognitivos.

De acuerdo a la propuesta del Prof. Pasquali [5], no corresponde seguir analizando un test psicométrico que carece de validez de contenido/constructo. Resulta irrelevante considerar su posible fiabilidad (¿de qué nos vamos a "fiar"?)

En cuanto a los baremos, ¿qué sentido tiene comparar los pensamientos y sentimientos de un sujeto con la media de un grupo poblacional?. Si la población tiene un nivel de ansiedad moderado, quien lo tenga menor, ¿qué representa? y quien tiene un valor medio, ¿quiere decir que no tiene ansiedad, o que tiene la ansiedad de la mayoría?

Aún más, la edición a la venta en España (Madrid: TEA Ediciones, 2004) es una edición mexicana con baremos americanos y mexicanos.

Por todas estas razones, en nuestra opinión, la CMAS-R **carece de valor alguno para considerar sus datos con valor pericial** o **para tenerlos en cuenta como variable interviniente en el análisis de casos de bajo rendimiento o fracaso escolar**.

En conclusión: ***la escala CMAS-R puede muy bien formar parte de la historia de la evaluación psicológica y archivarla en el lugar correspondiente***.

Finalidad	**Valorar el nivel de ansiedad en Niños de 6 a 19 años**
Validez de contenido	**Nula**
Validez de constructo	**Nula**
Fiabilidad	**Irrelevante**
Baremos	**Irrelevantes**
Facilidad de Aplicación	**Alta**

[5] Profesor Luiz Pasquali del Instituto de Psicología de la Universidad de Brasilia

BENDER

Test guestáltico visomotor de Lauretta Bender

(Bender, L. 1938)

Lauretta Bender es la autora de un test muy conocido y empleado en Educación: el test guestáltico viso-motor, cuya primera publicación se llevó a cabo **en 1938.**

La autora se describe a sí misma como psiquiatra y neuróloga, especialista en patología cerebral con adultos. Se reconoce sin formación en psicología clínica ni otra clase se estudios de psicología. Sin embargo, en una época de gran desarrollo de la Psicología de la Guestalt y en la investigación psicopatológica con adultos creó este test, formado por una serie de tarjetas en las que aparecen unos dibujos que el paciente debía reproducir con la máxima exactitud posible. Los resultados que obtuvo la autora (que nunca hizo ningún otro test o instrumento de evaluación, es decir, que no era experta en el diseño o elaboración de instrumentos de evaluación) se referían a qué clase de errores en la reproducción de los dibujos cometían pacientes adultos con lesiones cerebrales.

El error que subyace al empleo de este test en educación (cuyo uso comenzó en los 60) es que considera que si un error en la reproducción de dibujos por parte de un adulto, se encuentra asociado a un daño cerebral concreto, entonces... si un niño entre los 5 y los 11 años de edad (época en la que se encuentra en desarrollo su cerebro) comete el mismo tipo de errores, eso significa que sufre algún tipo de anomalía en su funcionamiento cerebral. Los autores más atrevidos lo atribuyen a lesión cerebral y otros, más prudentes, lo explican como un indicador de retraso en el desarrollo madurativo del cerebro.

Lo más curioso del caso es que en varias decenas de años en que se lleva empleando este test **NADIE HA COMPROBADO que sean ciertas las hipótesis de partida;** habiéndose comprobado, todo lo contrario: los niños con indicadores de daño o anomalía cerebral según el test de Bender que son remitidos al Servicio de Neurología Pediátrica no reciben una diagnóstico confirmatorio de tales daños o desajustes.

En conclusión: **Nunca se ha puesto de manifiesto la validez criterial o predictiva del test de Bender con niños, por lo que consideramos absolutamente desaconsejado su uso en evaluaciones psicopedagógicas.**

En la actualidad, la mejor evaluación de daño o mal funcionamiento cerebral debe llevarla a cabo un Profesional de Neurología Pediátrica.

BOEHM

Test de Conceptos Básicos de BOEHM

(Boehm Ann, 1971; Madrid: TEA)

Para valorar este instrumento hemos optado por citar el resumen de un trabajo de investigación del Profesor Dr. D. Miguel Galeote Moreno (Departamento de Psicología Evolutiva y de la Educación. Facultad de Psicología. Universidad de Málaga. Campus de Teatinos s/n. 29071-Málaga. España)

"*El Test Boehm de Conceptos Básicos (TBCB) (Boehm, 1971) es una prueba bastante utilizada en el ámbito psicopedagógico y logopédico. Esta prueba parte del supuesto de que gran parte del fracaso escolar se debe a que los alumnos desconocen una serie de conceptos / términos básicos imprescindibles para seguir con éxito las enseñanzas curriculares. Tras varias etapas de investigación, en las que se analizaron las instrucciones que acompañaban los materiales didácticos en los centros de preescolar, se seleccionaron los 50 conceptos / términos que componen el test.*

Aunque en la actualidad se reconoce que la prueba adolece de varios problemas, éstos se ven agudizados en su adaptación española. Uno de los más importantes es que no se analizaron los programas y materiales curriculares para el nivel de preescolar empleados en nuestro país. Este paso es, no obstante, esencial para determinar la validez de contenido de la prueba, entendiendo por tal la representatividad de los ítems del universo de conocimiento especificado (Jensen, 1980).

El objetivo del presente trabajo consiste en verificar dicha validez. Más concretamente, se trata de verificar la hipótesis de si los conceptos / términos examinados son pertinentes para el aprovechamiento / éxito escolar en nuestro país. Estudios realizados en otros países no la han confirmado.

Para este fin: (1) se seleccionó una muestra bastante amplia de los materiales curriculares más utilizados para el nivel de preescolar; (2) se analizó las instrucciones que acompañaban esos materiales y (3) se comprobó si los conceptos incluidos en el TBCB (así como sus derivados, sinónimos y antónimos) aparecían en dichas instrucciones.

Los resultados muestran que los conceptos / términos que componen el TBCB no son representativos de los materiales curriculares empleados en nuestro país, en otras palabras, el test carece de validez de contenido."

Tal y como mencionamos en la página principal de esta Sección, si un test carece de validez de contenido resulta poco útil seguir valorando otras propiedades del mismo, aunque sería bueno destacar que los elementos del test (50 se consideran escasos) solamente incluyen tres opciones de respuesta con lo que la probabilidad de acertar por azar es muy alta.

En conclusión: ***en nuestra opinión, y de acuerdo a las conclusiones del Prof. Galeote, consideramos que el test de Boehm, cumplió una función en un momento dado, pero en la actualidad puede muy bien formar parte de la historia de la evaluación psicológica.***

Finalidad	**Valorar el conocimiento de palabras que designen conceptos básico** **(incluidos en el currículum académico)**
Validez de contenido	**Insuficiente**
Validez de constructo	**Regular**
Fiabilidad	**Regular**
Baremos	**Insuficientes: muestras escasas**
Facilidad de Aplicación	**Alta**

CARAS

Test de Percepción de Diferencias

(Thurstone, L.L., 1944)

(Versión española preparada por el Dr. Mariano Yela en 1979; Madrid: TEA)

Se afirma en la introducción del Manual de este test que... "*este instrumento psicométrico fue preparado inicialmente con el propósito de apreciar la rapidez para percibir detalles y discriminar objetos, esto es, la capacidad del sujeto para detectar semejanzas y diferencias*".

Así entendido, constituye un grave error conceptual y metodológico considerar, como se ha venido haciendo desde bastantes años, el Test de Caras como un test para medir la atención. Atender a un estímulo es una condición imprescindible para poder realizar algún tipo de operación con él, por lo tanto "todos los test serían test de atención", si consideramos que para realizar la tarea que se nos solicita en cada uno de ellos, hace falta "poner atención".

Los doctores Thurstone y Yela nunca consideraron a este test como una prueba para medir la capacidad de atención, sino para valorar la "RAPIDEZ para APRECIAR DETALLES y DISCRIMINAR OBJETOS".

Con esta breve aclaración introductoria, deseamos salir al paso de quienes, ahora que está "de moda" diagnosticar a muchos escolares como Hiperactivos o Inatentos, en todo caso con TDAH, deciden utilizar este test para la valoración de sus capacidades atencionales. A nuestro entender, esto constituye un ERROR METODOLÓGICO GRAVE en los procesos de evaluación psicopedagógica de los escolares con sospechas de algún tipo de déficit de atención. Este test nunca se diseñó con esa finalidad.

Por otra parte, en la misma introducción se destaca que los estudios factoriales llevados a cabo con esta prueba por el Dr. Yela han mostrado que tiene una composición factorial compleja que abarca, principalmente, aspectos perceptivos y espaciales.

¿Cuál es el significado de esta afirmación introductoria?

Pues la más importante es la siguiente: si la prueba tiene una composición factorial, ¿por qué no se proporcionan los valores correspondientes a cada factor de manera independiente?

Resulta una incongruencia que el test proporcione una puntuación única y sin embargo se presente como de componentes factoriales.

¿Qué es lo que representa la medida del test y, por tanto, los baremos que proporciona? y, ¿cómo es posible que solamente se contabilicen los aciertos sin ponerlos en relación con el total de figuras vistas, los errores y las omisiones?

¿Significa lo mismo una puntuación de 30 con diez errores que otra de 30 sin errores, que otra de 30 con 15 errores? Y ¿es igual haber acertado 30 habiendo revisado sólo 30 figuras, que haber acertado 30 con 20 errores, o sea haber revisado 50 figuras?

Pues según este test sí: es lo mismo.

Así que, si el test mide... VELOCIDAD DE EJECUCIÓN + EFICACIA PERCEPTIVA DE SEMEJANZAS y DIFERENCIAS…

El resultado, ¿qué dice sobre la velocidad y la eficacia?

Por otra parte, si tiene componentes factoriales de ASPECTOS PERCEPTIVOS y ESPACIALES, una buena puntuación quiere decir que tiene buenas capacidades perceptivas y espaciales, pero una mala puntuación, ¿qué implica?: ¿mala percepción o mala orientación espacial?

Desacreditada la validez de constructo de este test, carece de interés proceder a valorar su fiabilidad o su baremación. No obstante, destaca que en el manual no se indica nada de la procedencia de las muestras, claramente insuficientes.

En conclusión: Con todo nuestro respeto y consideración hacia el Dr. Thurstone, en nuestra opinión, el test de Percepción de Diferencias (Caras) , igual que el Test de Formas Idénticas, puede muy bien formar parte de la historia de la evaluación psicológica y archivarlo en el lugar correspondiente.

Finalidad	**Valorar la rapidez y eficacia perceptiva de semejanzas y diferencias en Niños de 6 a 15 años**
Validez de contenido	**Nula**
Validez de constructo	**Nula**
Fiabilidad	**Irrelevante**
Baremos	**Irrelevantes**
Facilidad de Aplicación	**Alta**

COLUMBIA

Escala de Madurez Mental de Columbia

(Blum, L. H., Burgemeister, B. B. y Lorge, I., 1954; Madrid: TEA)

Sobre la Validez de Constructo

La escala de Madurez Mental de Columbia, ha sido durante los pasados 50 años, pese a su notable antigüedad, uno de los mejores instrumentos para identificar niños con retraso en el desarrollo intelectual.

Su construcción se describió en detalle por los autores en 1951, en la Revista "School and Society", 73 (1895), 232-233.

De una moderada complicación en su aplicación, su contenido es claramente un método poco discutible de evaluación de la capacidad de razonamiento del niño.

Para realizarla, el niño no requiere ni dominio del lenguaje oral, ni motricidad fina, ni memoria. Solamente requiere poner en funcionamiento sus recursos de razonamiento.

Asegurada la validez de constructo del instrumento, solo cabe considerar la fiabilidad del mismo.

Sobre la Fiabilidad

La consistencia interna y el test-retest presentan buenos valores (página 33 del manual)

En cuanto a los baremos con muestras poblacionales la cuestión es diferente. Los estudios de comparación con la población estadounidense concluyeron con la imposibilidad de utilizar los baremos originales; por lo cual se llevó a cabo por el editor español un estudio nacional cuya muestra resultó muy reducida (N= 571 sujetos).

Evidentemente, esta muestra de tipificación resulta insuficiente para una valoración objetiva y consistente de la inteligencia de los niños, no obstante, sí ha servido durante un tiempo para poner de manifiesto niños con retraso en el desarrollo individual de manera relevante. Ese y no otro ha sido el uso que le hemos dado nosotros, resultando mucho mejor este instrumento que el WIPPSI o el Terman-Merrill.

No obstante, la necesidad de disponer de un instrumento válido y fiable, tipificado con muestras amplias, diversificadas y aleatorizadas, nos llevó a desarrollar las investigaciones que culminaron con la Escala Magallanes de Inteligencia para Niños, **EMIN-6**, inspirada en parte en la metodología de la Escala Columbia y en parte en la del test SON-R.

En conclusión: Con todo nuestro respeto y consideración hacia los autores que en la primera mitad del siglo pasado lo elaboraron, en nuestra opinión la Escala de Madurez Mental de Columbia, puede muy bien formar parte de la historia de la evaluación psicológica y archivarlo en el lugar correspondiente.

Finalidad	**Valorar el nivel de desarrollo de la capacidad de razonamiento en niños preescolares**
Validez de contenido	**Alta**
Validez de constructo	**Alta**
Fiabilidad	**Escasa**
Baremos	**Anticuados e insuficientes**
Facilidad de Aplicación	**Moderada**

CONCEBAS

Test de Conceptos Básicos para Educación Infantil y Primaria

(Galve Manzano, J.L., García Pérez, E. M. y Yuste, J., 1993; Madrid: CEPE)

La construcción de este test adolece de un error de diseño de sus elementos (explicable por la falta de experiencia de sus autores en el diseño de este tipo de tests), por lo que resultaba afectada moderadamente su validez de constructo, pero gravemente su fiabilidad. Los distintos elementos del test son dibujos en un cuadrado de entre los cuales el sujeto debe señalar aquél que cumple un criterio: "*el que está arriba*", "*el que está debajo*", "*el mayor*",... El error de diseño consiste en que en cada elemento debería haber el mismo número de opciones de respuesta y sin embargo, no es así. Existen elementos en los que hay más y otros en los que hay menos opciones de respuesta, lo que afecta a la probabilidad de acertar por azar.

En el momento actual el instrumento se encuentra agotado en el mercado y desautorizada cualquier reedición del mismo.

Los usuarios de este instrumento pueden sustituirlo cuando lo deseen por CONCEBAS-2000, similar conceptual y metodológicamente al CONCEBAS pero con las modificaciones pertinentes para mejorar su fiabilidad y, consecuentemente con una nueva baremación.

Finalidad	**Valorar el conocimiento de palabras que designen conceptos básicos**
Validez de contenido	**Buena**
Validez de constructo	**Regular**
Fiabilidad	**Regular**
Baremos	**Buenos: muestras amplias**
Facilidad de Aplicación	**Alta**

CUMANIN

Cuestionario de Madurez Neuropsicológica Infantil

(Portellano, J.A. y otros, 2002. Madrid: TEA Ediciones)

En primer lugar correspondería destacar que el instrumento se denomina "Cuestionario", lo que automáticamente lo descartaría como un test psicométrico. No obstante su manual corresponde en todos sus apartados a un test psicométrico, ya que el índice explicita los siguientes apartados:

- *Fundamentos teóricos de la "prueba"*
- *Normas de corrección y puntuación de las diferentes "escalas"*
- *Justificación estadística*
- *Normas de interpretación de los resultados*

Así pues, debemos considerar el instrumento como un test psicométrico y proceder a valorarlo como tal, pese a su extraña denominación de "Cuestionario".

En cuanto a sus fundamentos conceptuales, los autores introducen el test exponiendo el marco de referencia de la neuropsicología infantil. Obviamente la neuropsicología infantil es, sencillamente, neuropsicología, y a tal fin convendría destacar los primeros errores conceptuales que se presentan en el texto.

Los autores, en la página 10, indican que "*la neuropsicología infantil, también llamada neuropsicología del desarrollo,..., estudia las relaciones que existen entre la conducta y el cerebro en fase de desarrollo, desde el embarazo hasta el comienzo de la escolaridad obligatoria en torno a los 6 años*."

Cabe hacer aquí una primera precisión que resultará fundamental para comprender los siguientes análisis: los autores no diferencian "comportamiento" de "conducta". Ya se encuentra claramente establecido en Psicología que el objeto de estudio de esta Ciencia es la Conducta y no el Comportamiento.

La Conducta es la interacción de un individuo con el medio en que se encuentra en un momento dado.

El Comportamiento es la ejecución cognitiva (una idea), emocional (una emoción) o motora (una acción) de un sujeto, ante un Contexto Estimular Determinado (Antecedente), cuya manifestación puede ir seguida de un Cambio Estimular Consecuente (Consecuencias)

El Comportamiento es el producto de la fisiología, la cual NO SOLAMENTE INCLUYE A LAS NEURONAS, sino también a los MÚSCULOS y a las GLÁNDULAS.

Por ello, el objeto de estudio de la Neuropsicología, NO ES LA CONDUCTA, sino el COMPORTAMIENTO. En tanto que el Comportamiento es un elemento constitutivo de la Conducta, resulta de cierto interés al Psicólogo conocer qué factores de tipo biológico (neurológico, glandular o muscular) intervienen como Factores Moduladores de la Interacción Conductual.

Aclarado este primer punto, una prueba de evaluación neuropsicológica -como pudiera ser ésta- solamente pretende explicar la relación existente entre el estado fisiológico del sujeto (neurológico, muscular y glandular) y sus ejecuciones (cognitivas, emocionales o motoras).

Por ello, se convierte en una necesidad perentoria asegurar que los resultados de esta prueba (y otras similares) se explican de manera unívoca por el estado de desarrollo fisiológico y no por la ausencia de procesos de enseñanza-aprendizaje o por la presencia de procesos insuficientes o erróneos, aún cuando éstos pudieran afectar al primero.

A este respecto, la aspiración del CUMANIN es informar, a través de diversas ejecuciones motrices de la situación de desarrollo neurofisiológico de los niños. Los cual pretende llevarse a cabo con diversas pruebas: de **Motricidad** (11 elementos), **Lenguaje articulatorio** (15 elementos), **Lenguaje expresivo** (4 elementos), **Lenguaje Comprensivo** (9 elementos), **Orientación en el plano gráfico** (15 elementos), **Reproducción gráfica de figuras** (15 elementos), **Memoria icónica** (10 elementos), **Reproducción de ritmos** (7 elementos), **Fluidez verbal** (4 elementos), **Atención selectiva visual** (20 elementos), **Lectura** (12 elementos; solo para los que saben leer), **Dictado** (10 palabras y dos frases, solo para quienes saben leer y escribir) y **Lateralidad** de manos, ojos y pies.

Como bien ponen de manifiesto los baremos que se proporcionan, los niños no nacen sabiendo hacer nada de esto, sino que lo tienen que APRENDER. El aprendizaje se lleva a cabo mediante procesos de mayor o menor calidad, frecuencia e intensidad y sus resultados varían de acuerdo a los métodos y ritmos de aprendizaje en interacción con las capacidades básicas de los sujetos (las que se pretende evaluar con este instrumento)

Pues bien, los autores del CUMANIN atribuyen unívocamente el resultado insatisfactorio en estas pruebas al grado de desarrollo neurológico (neuropsicológico afirman en el manual) cuando muy bien podrían explicarse de una manera alternativa por una deficiencia en los procesos de enseñanza-aprendizaje a los que se han visto sometidos con anterioridad. La duda se resolvería de un modo claro si el CUMANIN ofreciera información sobre su VALIDEZ CRITERIAL, es decir, si los sujetos con resultados deficitarios en alguna o varias de estas pruebas hubieran sido sometidos a evaluación neurológica y los Neurólogos hubieran confirmado en los casos de bajas puntuaciones, o descartado en los casos de altas puntuaciones, la presencia de anomalías en el desarrollo neurofisiológico.

Habida cuenta de que estos estudios no aparecen por ninguna parte del Manual, hemos de considerar que no se han llevado a cabo, por lo que **el instrumento CARECE ABSOLUTAMENTE DE VALIDEZ DE CONTENIDO/CONSTRUCTO para las variables que pretende medir.**

En conclusión: en nuestra opinión, la ausencia de validez intrínseca del CUMANIN lo hace descartable como medio de evaluación clínica o educativa en casos de niños con Problemas de Aprendizaje o de Conducta.

Finalidad	**Valorar el grado de madurez neuropsicológica de niños de 3 a 6 años** **Detectar la presencia de signos de disfunción cerebral**
Validez de contenido	**Nula**
Validez de constructo	**Nula**
Fiabilidad	**Irrelevante**
Baremos	**Irrelevantes**
Facilidad de Aplicación	**Irrelevante**

d2

Test de Atención

(R. Brickenkamp y E. Zillmer, 1962; Edición en España en el año 2002, Madrid: TEA)

La atención fue una variable menospreciada durante los dos primeros tercios del pasado siglo en nuestro país. Buena prueba de ello ha sido la ausencia de nuevos tests sobre la misma, desarrollados en el último tercio del siglo. Esta situación no es específica de España ya que toda Europa ha adolecido del mismo déficit. Probablemente, al iniciarse un interés repentino por los Trastornos por Déficit de Atención, Psicólogos, Pedagogos, Neurólogos y Psiquiatras, han vuelto los ojos hacia los tests psicométricos con los que poder contrastar los criterios conductuales del diagnóstico clínico de TDAH.

Quizás por ello, algunas editoriales de tests han querido sacar provecho de este repentino interés y, en vez de desarrollar nuevos instrumentos, basados en los nuevos modelos conceptuales sobre la atención (por ejemplo, el modelo de los cinco factores de Mirsky y Tatman, 1995), decidieron que lo más cómodo, rápido y económico resultaría lanzar una campaña publicitaria de un test de antigüedad notable: el D2.

Así, hace algunos años, la Revista de la Asociación Europea de Evaluación Psicológica vio aparecer en sus páginas el anuncio de este test como "nuevo", "válido" y "fiable". Nada más alejado de la realidad. Cuando, motivados por esta publicidad, adquirimos este test para su análisis, nos llevamos una gran desilusión: el test carecía de fundamento conceptual (los autores no mencionan en qué modelo explicativo de la atención basan su diseño), se atreven a afirmar que el test permite evaluar la velocidad de procesamiento (the test measures processing speed) y solamente permite su aplicación **a partir de los 9 años** de edad.

Obviamente **ES TÉCNICAMENTE IMPOSIBLE evaluar la velocidad de procesamiento con este test.** No hay ningún método en el mismo que permita diferenciar el tiempo que lleva al sujeto el procesamiento visual de las imágenes, del tiempo que le ocupa la ejecución de la respuesta motriz. De este modo, un sujeto que realice el test con cierta lentitud motora resultaría asignado como de "velocidad de procesamiento lenta", frente a un sujeto de ejecución motora rápida, quien sería valorado como de velocidad de procesamiento rápida (problema de la misma naturaleza que el presentado por los tests WISC-III y WISC-IV).

Además, dada la brevedad del test, la evaluación de la atención sostenida probablemente nos dejaría fuera de rango a los sujetos con este déficit (TDAH) al igual que detectamos que ocurría con los tests de **Caras** y de **Percepción de Diferencias**. Evidentemente, tal y como reconoce el manual, este test ha sido diseñado inicialmente para evaluar la capacidad de atención selectiva en sujetos aspirantes a obtener el carnet de conducir y en selección de personal (o sea que su uso ha sido durante los pasados 40 años, predominantemente en adultos).

Por otra parte, resulta discutible metodológicamente que la valoración de la capacidad atencional se lleve a cabo considerando solamente el resultado de Aciertos (TR) menos las Omisiones (O) y los Errores o Comisiones (C). Debe tenerse en cuenta que un mismo resultado, por ejemplo "43", puede obtenerse con un número elevado de combinaciones posibles de sus componentes: TR, O, y C; Caso 1: 60 -(10+7); Caso 2: 60 - (7 +10); Caso 3: 60 - (12 + 5),....; Caso 20: 43 - (0+0);....

La **fiabilidad de este parámetro es nula** para informar sobre la relación existente entre el total de figuras acertadas, las omitidas y las erróneamente marcadas.

Por lo tanto, a vista de los datos anteriores, **consideramos el D2 de uso inaceptable en evaluación de dificultades de aprendizaje y en valoración de trastornos por déficit de atención.**

Sin embargo y pese a las limitaciones antes expuestas, de las que TEA Ediciones no parecía estar al tanto, hace unos años nos sorprendió con la edición de este test en España. Y lo llevó a cabo incumpliendo un protocolo universalmente aceptado en la adaptación de un test psicométrico a poblaciones nuevas, en este caso a población española. El editor español, decidió aplicar el test a sujetos de 8 años de edad, siendo así que el instrumento original se había diseñado (y supuestamente validado) con sujetos a partir de 9 años de edad.

No sólo incluye a los sujetos de 8 años de edad en la realización de una prueba que tiene unos requerimientos pensados primordialmente para adultos, sino que nos sorprende al ponerse en evidencia notables contradicciones en los baremos que proporciona el manual técnico del test. Así, en la página 55 informa que el grupo de sujetos de 8 a 10 años de edad es de 127 en total (repartidos muy concretamente en 65 varones y 62 mujeres), pero en la tabla de baremos de la página 74 este grupo se ha convertido en 217. Por otra parte, en la página 11 indica que "*los baremos españoles han sido construidos a partir de una muestra general de varios millares de casos*", sin embargo en la página 55 la suma total de la muestra de baremación no llega al millar y medio. **Si ambos son errores involuntarios, la editorial no ha incluido una fe de erratas.**

En cualquier caso, no parece que en el siglo XXI corresponda usar este instrumento, posiblemente ventajoso en su época, 1962, de gran utilidad durante los años 60, 70 y 80, pero muy alejado de las necesidades de evaluación actuales.

En este caso, consideramos que la validez de constructo del test es muy discutible [los autores se limitan a establecer correlaciones con otras pruebas para validar el constructo], sus baremos con población infantil y adolescente escasos y su utilidad práctica en la evaluación de las capacidades atencionales en niños muy reducida, lo que haría necesario contrastar esta evaluación con otro instrumento más moderno y mejor baremado como la EMAV.

Finalidad	**Evaluar la capacidad de atención selectiva y**

	sostenida
Validez de contenido	**Aceptable**
Validez de constructo	**Discutible**
Fiabilidad	**Muy baja**
Baremos	**Buenos en adultos, Insuficientes en niños y adolescentes**
Facilidad de Aplicación	**Alta**

EACP

Escalas de Áreas de Conductas Problema

(García Pérez, E.M. y Magaz, A., 2000; Bilbao: COHS Consultores en CC.HH.)

Las Escalas de Áreas de Conductas-Problema se diseñaron para su empleo en los Centros Educativos y en las consultas de Atención Primaria de Salud, con la finalidad de detectar bien, de modo primario, bien secundario, la existencia de problemas en el desarrollo de niños desde los 4 hasta los 12 años de edad. Estas escalas se dotaron, cada una de ellas, de suficiente validez de contenido y predictiva, así como de una gran facilidad de aplicación, al objeto de resultar de una gran utilidad a Tutores, Orientadores, Personal de Enfermería Pediátrica y otros, en la detección de posibles problemas de

- Hiperactividad (DAH)
- Agresividad
- Retraimiento Social o Depresión
- Ansiedad
- Bajo Rendimiento Escolar

Pese a sus buenas propiedades, la aparición del grupo de sujetos con la condición Déficit de Eficacia Atencional (no déficit de atención sostenida) y Lentitud Motriz y Cognitiva (denominada provisionalmente "tempo cognitivo lento") hizo necesario proceder a una sustitución de este instrumento por otro, funcionalmente equivalente, pero que incluyese la posibilidad de detectar también a los niños denominados "Inatentos".

En la actualidad las **Escalas Magallanes de Detección de Déficit de Atención** (EMA-DDA) y otros problemas del desarrollo han sustituido en el mercado profesional a las EACP. Esta sustitución no se ha llevado a cabo por errores conceptuales ni metodológicos, sino por insuficiencia de las mismas para detectar al grupo de *Inatentos*, del que los primeros estudios de prevalencia realizados indican un posible porcentaje del 13% en la población, con una gran incidencia en el bajo rendimiento escolar.

Finalidad	**Detectar la presencia de DAH y otros problemas en el desarrollo en niños**
Validez de contenido	**Adecuada, pero insuficiente para identificar inatentos**
Validez de constructo	**Adecuada**
Fiabilidad	**Alta**
Baremos	**No relevantes**
Facilidad de Aplicación	**Alta**

EDAH

Escalas para la Evaluación del Trastorno por Déficit de Atención con Hiperactividad

(Farré i Riba, Anna y Narbona, Juan, 1998; Madrid: TEA)

Pocos profesionales de la Educación y la Salud Infanto-Juvenil ignoran el fenómeno mediático internacional en que se ha convertido el TDAH en los últimos años. De ser un problema ignorado por más del 75% de los profesionales de la Enseñanza y la Pediatría, se ha convertido en poco tiempo en el denominado "trastorno de conducta más frecuente en la infancia" o en "una de las causas más frecuentes de fracaso escolar y de problemas sociales en la edad infantil". Aunque estas afirmaciones no son ciertas en absoluto, sino basadas en observaciones sesgadas y, en muchos casos interesadas, la realidad es que el denominado "trastorno por déficit de atención con hiperactividad" es una categoría diagnóstica que se viene atribuyendo con excesiva frecuencia a muchos niños y adolescentes. En realidad, si no se detiene este proceso, puede que llegue a darse el caso de que las afirmaciones anteriores resulten ciertas. En realidad, se van convirtiendo en "verdades de mentira" (algo que es verdad que sucede: *lo que más se diagnostica es TDAH*, pero que conlleva una falsedad intrínseca: *se diagnostica TDAH por interés académico o económico, por desconocimiento, o por otros motivos*, pero es un diagnóstico erróneo y falso). Cada vez son más las voces que se levantan en todo el mundo para denunciar un sobre-diagnóstico de TDAH (falso, erróneo). Un ejemplo excelente de esto lo constituye el trabajo encargado por el Parlamento del Gobierno Autónomo del Oeste de Australia a una Comisión de Expertos y su Informe y Recomendaciones al mismo [6].

Sin embargo este fenómeno del sobre-diagnóstico puede explicarse muy bien, en parte, por el empleo que realizan algunos profesionales de instrumentos como el presente.

Aprovechando algunas afirmaciones de la introducción de los autores, deseamos destacar lo siguiente:

1. Afirman ambos que *medición y evaluación son complementarios: para evaluar debemos medir*.... y la medida se debe interpretar en función de los conocimientos del evaluador. Como analizaré a continuación, ambas condiciones NO se cumplen en el instrumento del que son coautores. Si la medida es errónea, equivocada, la evaluación ya queda afectada en origen por el error de medida. Y si los criterios de la evaluación son erróneos o insuficientes, el resultado no puede ser acertado. Fíjese el lector que los baremos empleados en este test corresponden a un total de 33 casos (¡!) según consta en la página 30, que, además, tienen entre 5 y 12 años de edad. Menos mal que reconocen que..."*el pequeño tamaño de la muestra puede ser un inconveniente a la hora de analizar los resultados*." A nuestro juicio no es un inconveniente, es sencillamente algo inaceptable para poder generalizarlos.

[6] Puede obtener acceso a este informe visitando las páginas web www.tda-h.com

2. *Con este trabajo se intenta profundizar en el conocimiento teórico del TDAH*. Difícilmente se puede profundizar teóricamente, si se parte de los supuestos teóricos del DSM-III (1980), como es el caso de la EDAH, basada en las escalas de la citada época -sustituido en 1992 por el DSM-IV- y no se lleva a cabo ninguna innovación ni conceptual ni metodológica en este instrumento (téngase en cuenta que los autores reconocen que los ítems se han tomado de las Escalas de Conners)

3. Aseguran que la EDAH se trata de una "nueva propuesta de escalas para la valoración del TDAH con baremación en nuestra población" (en la primera edición 33 sujetos)

Sobre la Validez de Constructo

Resulta evidente que todo instrumento de medida debe, en primer lugar, acreditar su validez de constructo ya que, en caso contrario, sus restantes características, por muy buenas que sean, resultan irrelevantes. Pues bien, la EDAH, instrumento concebido para "valorar la existencia de TDAH en un sujeto", debería recoger evidencia empírica indiscutible de que el sujeto en cuestión cumple los criterios aceptados internacionalmente para asignarle esta categoría clínica.

¿Lo hace la EDAH? Pues de ningún modo. Comprobémoslo:

En primer lugar debemos partir de la Clasificación Internacional de Enfermedades (CIE-10) aprobada por la O.M.S. o bien del Manual de Diagnóstico Estadístico (DSM-IVTR) aprobado por la A.P.A.. Por razones difíciles de admitir y que solo se pueden presuponer, la Clasificación CIE-10, de la Organización Mundial de la Salud, suele ser ignorada sistemáticamente en el campo de la Salud Mental por los profesionales de casi todo el mundo.

El Profesor Narbona es conocido por su adhesión a los criterios DSM-IVTR, posiblemente porque los considera más acertados que los de la CIE-10 (a pesar de que ya están previstos sus cambios en 2011, en el DSM-V) Así pues, revisemos los criterios propuestos por el DSM-IV para el diagnóstico de TDAH...

Criterios DSM-IVTR sobre el TDAH

I. A o B:

A. A. Seis o más de los siguientes síntomas de inatención han estado presentes en la persona por lo menos durante 6 meses, al punto de que son inadecuados y tienen un efecto perturbador del nivel de desarrollo:

Inatención

1. A menudo no presta la debida atención a los detalles o, por descuido, comete errores en las tareas de la escuela, el trabajo y otras actividades.
2. A menudo tiene problemas para concentrarse en las tareas o en los juegos.
3. A menudo parece que no escucha cuando se le habla directamente.

4. A menudo no sigue las instrucciones y no termina las tareas de la escuela, los quehaceres o cualquier otra responsabilidad en el trabajo (no por conducta oposicional o por no entender las instrucciones).

5. A menudo le cuesta organizar actividades.

6. A menudo evita, rechaza o se niega a hacer cosas que requieren mucho esfuerzo mental por mucho tiempo (como tareas escolares o quehaceres de la casa).

7. A menudo pierde las cosas que necesita para hacer ciertas tareas o actividades (p. ej. juguetes, trabajos escolares, lápices, libros, o herramientas).

8. Se distrae con frecuencia.

9. Tiende a ser olvidadizo en la vida diaria.

B. Seis o más de los siguientes síntomas de hiperactividad-impulsividad han estado presentes en la persona por lo menos durante 6 meses, al punto de que son inadecuados y tienen un efecto perturbador del nivel de desarrollo:

Hiperactividad

1. A menudo no deja de mover las manos ni los pies mientras está sentado.

2. A menudo se levanta de la silla cuando se quiere que permanezca sentado.

3. A menudo corre o trepa en lugares y en momentos inoportunos (es posible que los adultos y adolescentes se sientan muy inquietos).

4. A menudo, tiene problemas para jugar o disfrutar tranquilamente de las actividades de recreación.

5. A menudo, "está en constante movimiento" o parece que tuviera "un motor en los pies".

6. A menudo habla demasiado.

Impulsividad

1. A menudo suelta una respuesta sin haber oído antes toda la pregunta.

2. A menudo le cuesta esperar su turno.

3. A menudo interrumpe al que esté hablando o se entromete, por ejemplo, en una conversación o juego.

II. Algunos de los síntomas que causan alteraciones están presentes **desde antes de los 7 años de edad.**

III. Alguna alteración provocada por los síntomas está presente en dos o más situaciones (p. ej., en la escuela o el trabajo y en la casa).

IV. Debe haber **clara evidencia de una alteración considerable en el funcionamiento social, escolar o laboral.**

V. Los síntomas no ocurren únicamente mientras la persona sufre de trastorno generalizado del desarrollo, esquizofrenia u otro trastorno sicótico. Los síntomas no indican la presencia de otro trastorno mental (p. ej. trastorno del humor, trastorno de ansiedad, trastorno disociativo o trastorno de la personalidad)

Con base en estos criterios, se identifican tres tipos de TDAH:

1. TDAH, tipo combinado: si en los últimos 6 meses se ha cumplido tanto el criterio 1A como el 1B.

2. TDAH, tipo predominantemente inatento: si en los últimos seis meses se ha cumplido el criterio 1A, pero no se ha cumplido el 1B.

3. TDAH, tipo predominantemente hiperactivo-impulsivo): si en los últimos seis meses se ha cumplido el criterio 1B, pero no se ha cumplido el 1A.

American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision. Washington, DC, American Psychiatric Association, 2000

Respecto a la validez de contenido, sin la cual el resto de parámetros de la valoración del test quedarían invalidados:

A. ¿Incluye la EDAH los 9 indicadores de Inatención propuestos por el DSM-IV? ..... No.

De hecho ni siquiera los que la hoja de respuestas asigna como indicadores de déficit de atención lo son. Véalo:

Item 2. Tiene dificultades de Aprendizaje Escolar (un punto inaceptablemente asignado a déficit de atención, ya que las dificultades de aprendizaje pueden ser diversas)

Item 4. Se distrae fácilmente. Muestra escasa atención (un punto -razonable- a déficit de atención)

Item 7. Está en las nubes, ensimismado (un punto -razonable- a déficit de atención)

Item 8. Deja por terminar las tareas que empieza (un punto, discutible pero incluso aceptable, a déficit de atención; puede dejarlas por fatiga, por que no sabe hacerlas u otros motivos).

Item 19. Sus esfuerzos se frustran fácilmente, es inconstante (un punto, inaceptable, a déficit de atención)

Como puede verse la escala DA solo puede puntuar un máximo de 5 puntos; menos que el mínimo de entre 9 que propone el DSM-IV y que, además, incluye dos/tres ítems que no se pueden atribuir de manera unívoca a déficit de atención. Esto convierte la escala de Déficit de Atención en una escala de dos-tres puntos.

B. ¿Incluye la EDAH los 6 indicadores de Hiperactividad propuestos por el DSM-IV? ..... No.

De hecho ni siquiera los que la hoja de respuestas asigna como indicadores de hiperactividad lo son. Véalo:

Item 1. Tiene excesiva inquietud motora (un punto, razonable, a hiperactividad-hiperkinesia)

Item 3. Molesta frecuentemente a otros niños (un punto -inaceptable- a hiperactividad-hiperkinesia; molestar son los efectos de un comportamiento, pero no es un indicador de comportamiento y, además, no necesariamente de un comportamiento hiperkinético)

Item 5. Exige inmediata satisfacción a sus demandas (un punto -inaceptable- a hiperactividad-hiperkinesia; que quiera las cosas rápidamente, no implica necesariamente hiperactividad, se presenta en casi todos los niños maleducados)

Item 13. Se mueve constantemente, intranquilo (un punto, razonable pero "tramposo", a hiperactividad-hiperkinesia, puesto que es preguntar de nuevo lo mismo que en el item 1)

Item 17. Es impulsivo e irritable (un punto -inaceptable- a hiperactividad-hiperkinesia; aparte de difícil de responder porque incluye dos preguntas)

Como puede verse la escala H solo puede puntuar un máximo de 5 puntos; menos que el mínimo de 6 que propone el DSM-IV que, además, incluye tres ítems que no se pueden atribuir de manera unívoca a hiperactividad y un ítem que evalúa de manera inadecuada impulsividad. Esto convierte la escala de Hiperactividad en una escala de dos puntos. En el caso de querer asimilar esta escala H a la combinación de Indicadores de Hiperactividad-Impulsividad, véase cómo el total de indicadores del DSM-IVTR son 6 + 3 = 9, mientras que en esta escala HI tiene 2 + 1 = 3

Llegados aquí, se puede comprobar el primer efecto diagnóstico de la EDAH, incumpliendo los criterios internacionales, se puede llegar a considerar que el sujeto tiene TDAH con solamente unos 4-5 indicadores de déficit de atención e hiperactividad. Algo sencillamente inaceptable técnicamente.

Pero, aún más. Las condiciones para aceptar la presencia de trastorno por déficit de atención incluyen las siguientes:

II. Algunos de los síntomas que causan alteraciones están presentes desde antes de los 7 años de edad.

III. Alguna alteración provocada por los síntomas está presente en dos o más situaciones (p. ej., en la escuela o el trabajo y en la casa).

IV. Debe haber clara evidencia de una alteración considerable en el funcionamiento social, escolar o laboral.

¿dónde se contemplan en la EDAH estos condicionantes fundamentales? En ningún sitio.

Si se desea objetar que los indicadores de la escala de Trastornos de Conducta son las "alteraciones provocadas por los síntomas", en modo alguno se pueden atribuir de manera unívoca a los "síntomas" que aparecen en las otras escalas; aunque claro está: tener bajo rendimiento escolar debe estar provocado por el "síntoma-tiene dificultades de aprendizaje", y tener problemas de conducta en casa y en el aula también debe estar provocado por los "síntomas-molesta frecuentemente a otros niños-exige inmediata satisfacción a sus demandas-es irritable"

Por ello, la EDAH, copia fiel de las Escalas de Conners y en nada originales o diferentes de ellas, nunca podrían considerarse un instrumento para evaluar el TDAH, sino, tal y como ha propuesto siempre el Dr. Keith Conners y otros muchos profesionales (Taylor, 2003) como un instrumento de DETECCIÓN DE POSIBLES SUJETOS CON TDAH. Por lo tanto, si el profesional ya tiene en su consulta médica o psicopedagógica el "posible caso de TDAH" ¿para qué utilizar este instrumento si dispone de otros para confirmar o rechazar este diagnóstico?

Por otra parte, ¿cuál es el fundamento conceptual de sumar características de diferente naturaleza?

¿Cómo se explica que los autores decidan sumar puntuaciones de Déficit de Atención e Hiperactividad (puntuación H +DA) y, para colmo de barbaridad psicométrica, sumar a éstas los indicadores de Trastornos de Conducta (puntuación H + DA + TC).

Finalmente, si los autores se adhieren al DSM-IVTR, ¿cómo se explica que la EDAH no permita discriminar entre los subtipos propuestos por esta clasificación?. Subtipos por otra parte discutibles y discutidos desde su aprobación en 1992, encontrándose previsto su cambio en el DSM-V (Barkley y otros, 1997, 2005, 2008)

En resumen: la EDAH es un instrumento reproducción parcial de las Escalas de Conners, con una lejana referencia en los criterios diagnósticos del DSM-III, carente en absoluto de validez de constructo y con muy deficiente validez de contenido cada una de sus escalas.

Si bien las consideraciones anteriores resultarían motivo suficiente para descartar el empleo de este instrumento con la finalidad propuesta, los problemas de su fiabilidad son de tan grave naturaleza que requieren un análisis detallado para decidir descartar de manera absoluta su empleo con fines profesionales.

Sobre la fiabilidad

A este respecto quizás viene bien recordar el antiguo dicho, que posiblemente provocará sonrisas en el autor original de estas escalas (ya modificadas y actualizadas), el Dr. Conners:

"Bienaventurados quienes nos copian porque ellos heredarán nuestros defectos"

La fiabilidad de un instrumento pretendidamente psicométrico es un aspecto fundamental del mismo que se ve afectado por muchos factores, destacando entre ellos:

f.1. Preguntas que contienen en su enunciado frecuencia o intensidad de la respuesta y para las que se solicita respuesta sobre intensidad o frecuencia.

*Tiene **excesiva** inquietud motora* ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

***A menudo** grita en situaciones inadecuadas* ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

***Se mueve constantemente**, intranquilo* ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

¿cómo contestar a esto?

f.2. Formulación confusa o poco clara de las preguntas, evitando la posibilidad de interpretaciones diferentes por sujetos diferentes

***Sus esfuerzos se frustran fácilmente**, es inconstante* ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

f.3. También, las preguntas que contienen en su enunciado dos componentes, de los cuales uno puede estar presente pero el otro no.

Es impulsivo e irritable ( ) Nada ( ) Poco ( ) Bastante ( ) Mucho

Estos déficits en la construcción de test son frecuentes en las etapas iniciales de quienes los elaboran, poniendo de manifiesto su escasa experiencia en este tipo de trabajos. El paso del tiempo ayuda a evitarlos y, de ese modo, mejorar la fiabilidad de los instrumentos aumentando su calidad psicotécnica.

Con respecto a los baremos ya hemos analizado con anterioridad la escasez de las muestras empleadas y su escasa diversificación y aleatorización, así como lo absurdo de los índices que se proponen al sumar puntuaciones que corresponden a variables de naturaleza diferente.

Pues bien, de manera incomprensible este instrumento recibió de la empresa editora el primer premio de investigación y así lo promocionan, confiando, suponemos, que el hecho de haber sido premiado (por la misma empresa editora) le otorgue un prestigio técnico del que a todas luces carece.

Como comentamos al principio, habida cuenta de los graves errores en que se incurre al emplearlo (exactamente los mismos en que incurrían las clásicas escalas de Conners) no es de extrañar que se incremente notablemente el número de escolares que reciben el diagnóstico de TDAH.

En conclusión: en nuestra opinión, la Escala EDAH carece de validez de constructo, de contenido y predictiva; su fiabilidad es irrelevante y muy reducida por los defectos de construcción de sus elementos y los baremos totalmente insuficientes y escasamente diversificados. Por todo ello lo consideramos un instrumento totalmente inadecuado, dados los conocimientos actuales sobre la condición DAH y la situación de Trastorno por DAH, tanto para la detección de TDAH como, por supuesto, para el diagnóstico clínico de TDAH.

Finalidad	**Valorar la presencia de TDAH en niños**
Validez de contenido	**Deficiente**
Validez de constructo	**Deficiente**
Fiabilidad	**Muy baja**
Baremos	**Insuficientes: muestras escasas**
Facilidad de Aplicación	**Alta**

EMTDAH

Escalas Magallanes de Identificación de Trastorno por Déficit de Atención con Hiperactividad

(García Pérez, E.M. y Magaz, A., 2000; Bilbao: COHS Consultores en CC.HH.)

Estas Escalas se diseñaron para su empleo en los Centros Educativos, en las consultas de Atención Primaria de Salud o en Consultas de Psicología Clínica, Psiquiatría o Neurología Infanto-Juvenil, con la finalidad de identificar o descartar, la presencia de "trastorno por déficit de atención con hiperactividad" en escolares.

Estas escalas se dotaron, cada una de ellas, de suficiente validez de contenido y predictiva, así como de una gran facilidad de aplicación, al objeto de resultar de una gran utilidad a los profesionales cualificados para el diagnóstico diferencial de TDAH.

Pese a sus buenas propiedades, la aparición del grupo de sujetos con la condición Déficit de Eficacia Atencional (no déficit de atención sostenida) y Lentitud Motriz y Cognitiva (denominada provisionalmente "tempo cognitivo lento") hizo necesario proceder a una sustitución de este instrumento por otro, funcionalmente equivalente, pero que incluyese la posibilidad de identificar también a los niños denominados "Inatentos".

Por otra parte, las condiciones "Hiperactivo" e "Inatento", son permanentes en los sujetos, pero sus manifestaciones conductuales se modifican con la edad. Por ello, se hacía necesario construir otras escalas con nuevos indicadores conductuales para identificar correctamente adolescentes y adultos.

Además, la situación de trastorno es cuantitativa, al modo del trastorno depresivo, lo que hace necesario valorar su "intensidad" y su "amplitud".

En la actualidad las **Escalas Magallanes de Identificación de Déficits de Atención** (ESMIDAs) han sustituido en el mercado profesional a la EMTDAH. Esta sustitución no se ha llevado a cabo por errores conceptuales ni metodológicos, sino por insuficiencia de las mismas para identificar al grupo de *Inatentos*, del que los primeros estudios de prevalencia realizados indican un posible porcentaje del 13% en la población, con una gran incidencia en el bajo rendimiento escolar.

Además, se incluye la posibilidad de estudiar los casos de adolescentes y adultos, así como el impacto de la condición hiperactivo/a o inatento/a en la población.

Finalidad	**Identificar la presencia de DAH en niños**
Validez de contenido	**Adecuada, pero insuficiente para identificar inatentos y, en general, para adolescentes**
Validez de constructo	**Adecuada**
Fiabilidad	**Alta**
Baremos	**Adecuados para la finalidad del test**
Facilidad de Aplicación	**Alta**

FORMAS IDÉNTICAS

(Thurstone, L.L., 1944)

Nacido el 29 de mayo de 1887 y fallecido el 29 de setiembre de 1955, el extraordinario Dr. Louis Leon Thurstone ingeniero notable, asistente y discípulo avezado de Thomas Edison, se doctoró en Psicología en 1917. En 1930 creó el Laboratorio Psicométrico, lugar donde se desarrollaron múltiples tests. Muchos de sus tests constituyeron la indiscutible vanguardia de los actuales, pero, con el paso del tiempo, Thurstone, hombre riguroso y honesto, siguió poniendo a prueba los resultados de sus estudios factoriales, llegando a concluir, tras muchos ensayos con las pruebas factoriales de inteligencia (P.M.A.) que era Spearman quién tenía razón y no él: la inteligencia es unifactorial y el parámetro para medirla debía ser el factor "g" propuesto por Spearman. Lamentablemente, a su fallecimiento, los herederos no quisieron tomar en consideración esta conclusión de su padre y permitieron reediciones sucesivas y adaptaciones en diversos países del P.M.A.

Con estos antecedentes, debe considerarse que el test de Formas Idénticas, editado por primera vez en España en 1975, no fue desarrollado en ese año, sino, evidentemente varios antes. Sobre todo porque el autor había fallecido en el año 1955. Este test se desarrolló en los años 40 en el Laboratorio de la Universidad de Carolina del Norte con la finalidad de estudiar de manera cuantitativa los procesos de percepción y atención, poniéndolos en relación con la inteligencia (idea, excepcional, sobre la que consideramos que debemos seguir trabajando en la actualidad)

Obviamente, dada su antigüedad, el manual del test no ofrece ni los estudios psicométricos originales, ni cualesquiera otros realizados por la empresa editora en España, lo cual imposibilita una valoración crítica del mismo.

Este test no se diseñó de acuerdo a ningún modelo concreto de atención y, por lo tanto, carece de sentido hablar de su validez de contenido/constructo.

Tampoco se ponen de manifiesto validez concurrente con alguna prueba de atención, aunque sí hay estudios de validez divergente con otras pruebas.

Por su naturaleza, la situación estimular que constituye el test es de una naturaleza muy similar al de Percepción de Diferencias, por lo que puede trasladarse aquí los mismos comentarios críticos que expusimos en el Test de Caras.

Anexo: el editor incluye en su página web lo siguiente sobre este test:

Informe narrativo

"*La apreciación de la capacidad para percibir y observar con atención es una de las medidas que tiene más utilidad en diversos campos pues el rendimiento en la mayoría de las tareas se ve influido por la mayor o menor dotación en estas aptitudes. Se trata de la capacidad para captar o aprehender visualmente, con rapidez y exactitud, una cierta configuración dada (palabra, letra, número o dibujo) que se repite una y otra vez dentro de un contexto que contiene otros materiales distractivos. Esta tarea supone el seguimiento de unas instrucciones, concentración, resistencia a la monotonía y perseverancia en tareas repetitivas. Estas aptitudes tienen una importancia relevante en la adquisición de experiencias, en el reconocimiento de nuevas situaciones y en la concepción clara de los problemas.*

A partir de las respuestas en la prueba, el examinado parece presentar alguna dificultad para realizar ciertas tareas repetitivas que precisan rapidez de percepción, atención y concentración. Derivado de ello, ***podría presentar dificultades de aprendizaje y rendimiento en las tareas que requieran de esta aptitud***".

No nos cabe ninguna duda de la exactitud de la afirmación destacada en negrita en el párrafo anterior. Nuestra duda se refiere a que en ningún apartado del manual técnico del test se pone de manifiesto tal riesgo, mediante los correspondientes estudios de validez predictiva. Es decir que tal afirmación es muy similar a la que nos comenta frecuentemente la Dirección General de Tráfico: "conducir a velocidad elevada podría tener relación con un posible accidente".

Si la capacidad perceptiva y de atención interviene en todo tipo de actividades, el rendimiento puede verse disminuido si alguna de estas capacidades se encontraran disminuidas. Ello conlleva la necesidad de evaluar, de manera válida y fiable las mismas, algo que no se consigue con el uso de este test (al menos tras el análisis de la justificación conceptual y estadística del mismo)

En conclusión: Con todo nuestro respeto y consideración hacia el Dr. Thurstone, en nuestra opinión, el Test de Formas Idénticas, igual que el de Percepción de Diferencias (Caras) puede muy bien formar parte de la historia de la evaluación psicológica y archivarse en el lugar correspondiente.

Finalidad	**Valorar de manera general las aptitudes perceptivas y de atención desde los 10 años**
Validez de contenido	**Nula**
Validez de constructo	**Nula**
Fiabilidad	**Irrelevante**
Baremos	**Irrelevantes**
Facilidad de Aplicación	**Alta**

FROSTIG

Test de Desarrollo de la Percepción Visual

(Frostig M.1964; Madrid: TEA)

Sobre la Validez de Constructo

El test de desarrollo de la percepción visual es un instrumento diseñado para comprobar que el escolar que va a iniciarse en el aprendizaje de la lectura ha adquirido las habilidades perceptivo visuales que se requieren para el mismo. Aunque el manual de la versión española no lo indica, consideramos que este test no es solamente de utilidad para la lectura, sino que lo es asimismo, para el aprendizaje de la escritura. En caso contrario no se entendería la razón de incluir una prueba de control motriz fino (coordinación viso-motora) y otra de relaciones espaciales. Así pues, considerando, de acuerdo a los Principios de la Psicología del Aprendizaje, que para adquirir una nueva habilidad (leer, escribir) se deben haber adquirido algunas habilidades senso-motrices, atencionales y de razonamiento, se comprende la buena estructuración de los contenidos de este instrumento.

Desde este punto de vista conceptual, la validez de contenido del TEST DE FROSTIG resulta poco discutible, aunque pudiera mejorarse.

Otra cuestión muy diferente es la cuantificación del constructo que trata de medir el test. El manual no aporta estudio estadístico alguno con el que se pueda justificar la consideración de un "COCIENTE PERCEPTIVO", tal y como propone la autora. ¿Qué significado puede tener un valor determinado de este índice si se construye con la suma de puntuaciones que corresponden a habilidades diferentes?

Reflexionemos sobre lo siguiente:

Un niño tiene un buen dominio de la coordinación visomotora y de relaciones espaciales, obteniendo una puntuación alta. Sin embargo tiene una escasa habilidad de discriminación figura-fondo. ¿Qué significado puede tener una puntuación en la que se suman unas y otras habilidades? Téngase en cuenta que también podría ocurrir al contrario: el niño tiene buenas habilidades perceptivas, pero escasa destreza visomotora.

Bien, algún profesional podrá afirmar que es posible llevar a cabo una valoración independiente de cada habilidad y estaríamos totalmente de acuerdo. Sin embargo, la cuestión a debatir es la existencia del parámetro COCIENTE PERCEPTIVO, objetivo conceptual de este test. La conclusión no puede ser otra que asegurar que tal parámetro no tiene justificación alguna: ni conceptual, ni estadísticamente ha sido establecido como válido.

Aún más, en el manual del test se encuentra ausente cualquier estudio sobre la validez predictiva del mismo. Por lo tanto, ¿en qué se fundamentan los criterios para predecir que el escolar no puede iniciar el

aprendizaje de la lectura (o la escritura)?. En conclusión, a pesar de tener una buena validez de contenido, **el Test de Frostig carece de Validez de Constructo y Predictiva.** Por ello, tal índice no puede incluirse en modo alguno ni en informes con valor pericial, ni en informes de evaluación psicopedagógica.

Sobre la Fiabilidad

En la medida en que el manual de este test no informa de los estudios que se hayan podido llevar a cabo sobre consistencia interna (homogeneidad de cada ítem con el total de cada prueba), consistencia temporal (test-retest) y que la baremación del mismo, con la que se han establecido las tablas de corrección que se ofrecen, se llevó a cabo con un grupo de niños franceses (en el año 1964!!!), menor de 300 sujetos en el rango de edad de 4 a 7 años; es decir, un promedio de 75 niños por año de edad. Tal grupo de baremación resulta sencillamente inadmisible. Si a tan reducido número de sujetos (de procedencia desconocida en cuanto a nivel sociocultural) le añadimos la antigüedad del mismo, podemos asegurar que este test no tiene, en modo alguno, baremos fiables.

Resumiendo: el test de Marianne Frostig, carece de Validez de Constructo y Predictiva, su Fiabilidad nunca ha quedado acreditada y sus baremos son insuficientes, inadecuados y anticuados.

En conclusión: apreciando en lo que vale la aportación que supuso este instrumento en el momento de su edición original, proporcionando un punto de vista conductual a la evaluación de las dificultades de aprendizaje, en nuestra opinión, el test de FROSTIG, puede muy bien formar parte de la historia de la evaluación psicológica.

Finalidad	**Valorar el nivel algunas habilidades perceptivo visuales y motrices en niños de 4 a 7 años**
Validez de contenido	**Aceptable en cada escala**
Validez de constructo	**Inaceptable**
Fiabilidad	**Escasa**
Baremos	**Escasos, Impropios y Anticuados**
Facilidad de Aplicación	**Media**

En la actualidad, los profesionales de la evaluación psicopedagógica cuentan con un instrumento alternativo para evaluar de manera concreta, operativa e independiente las diversas habilidades perceptivo-visuales y motrices, necesarias para la escritura y la lectura: la **Batería BAMADI**, con amplios y recientes baremos españoles.

ITPA

Test Illinois de Aptitudes Psicolingüísticas

(Kirk, S.A. McCarthy, J.J. y Kirk, W.D., 1968; Edición en España en el año 1989, Madrid: TEA)

(Edición revisada de la prueba, 2004, Madrid: TEA)

Cuando, en el año 89, la Dra. Soledad Ballesteros nos ofreció la posibilidad de colaborar en los estudios de adaptación del ITPA en España no lo dudamos un momento. El ITPA se encontraba precedido de un gran prestigio en el caso de la evaluación psicopedagógica como ejemplo de un instrumento de evaluación que permitía generar planes de intervención. Desde entonces las cosas han ido por ese camino. Abandonados los instrumentos que ofrecen datos con los cuales es difícil hacer algo, el desarrollo de los siguientes siempre ha llevado aparejada la posibilidad de intervenir con los sujetos que presentaban déficits en las habilidades evaluadas.

Tres problemas iniciales planteó el ITPA:

1. Su antigüedad. Elaborado en los años 60, resultaba con 20 años de antigüedad al llegar a España, lo que lo hacía candidato a ser "remozado". Pero no fue así, el editor lo publicó tal cual.

2. El modelo conceptual en el que se basaba el test y, por lo tanto, el fundamento científico del mismo. Este modelo teórico, se elaboró en el año 1957 por Osgood y se enmarca en el ámbito de la Psicología Cognitiva; lo cual lo sitúa fuera del estudio del comportamiento, centrándose en el análisis de los procesos (cognitivos) que se suponen juegan un papel en la comunicación individuo-medio.

3. Su denominación: el nombre original especifica que el test mide "habilidades" (Illinois Test of Psycholinguistic Abilities") pero se tradujo por test de Aptitudes; error que se mantiene en la edición revisada en el año 2004, obviando el consenso profesional y científico respecto a la evaluación de aptitudes: los tests psicométricos no pueden medir las aptitudes, ya que las aptitudes o capacidades son las posibilidades con las que cuenta un individuo para desempeñarse. Los tests miden (mejor o peor) grados o niveles de habilidad en un campo determinado. Es el evaluador quien, lleva a cabo una "inferencia" y, a partir del nivel de dominio de una habilidad, puesto de manifiesto en una situación estandarizada, establece, como criterio personal, el nivel de capacidad o aptitud del sujeto.

Así pues, partiendo de que el test aspira a medir habilidades y que fue publicado en el año 1968, el primer aspecto del mismo es su antigüedad actual. Debe tenerse en cuenta que a principios del año 2010, este test habrá cumplido los de 42 años de antigüedad. Sólo teniendo en cuenta este dato, el profesional podría plantearse las siguientes preguntas:

a. Partiendo del supuesto de que el test se hubiera elaborado en los años 60, de acuerdo a un modelo conceptual de la máxima vigencia en esa época, ¿resulta admisible en la actualidad el modelo y la estructura conceptual del test?

b. Teniendo en cuenta las variables a evaluar: procesos cognitivos relacionados con el lenguaje, y partiendo del hecho constatado de las grandes diferencias culturales entre la cultura norteamericana y la española, ¿la versión española del test fue correctamente adaptada en su momento, cumpliendo los requerimientos necesarios para dar por equivalente la versión española a la americana?

c. Los baremos realizados con población muestral española, ¿fueron suficientes y adecuados en el momento de la publicación del test?

d. Finalmente, y de manera añadida: ¿ha cumplido este test la función para la que se empleó durante estos 20 años? Realmente ¿ha generado programas eficaces de intervención en España?

En el año 2004, el editor español decidió llevar a cabo una actualización que, según indica en su manual (página 7) *mejoren la calidad de los estímulos, faciliten el proceso de aplicación e incorpore una nueva tipificación*.

Tras veinte años de uso del instrumento con población española, los profesionales que lo utilizamos durante ese tiempo disponemos de la experiencia suficiente para comunicar a las nuevas generaciones de psicólogos nuestros resultados con el mismo. Dicho en otras palabras, ¿qué se puede afirmar en el año 2009 sobre la **calidad psicotécnica** de este test y, sobre todo, de su **validez práctica**?

Intentaremos dar respuesta a estas preguntas.

Sobre la Validez de Constructo

El ITPA es un test diseñado en base a las propuestas del Modelo de Procesamiento de la Información, según el cual, brevemente,

- existen grupos de neuronas especializadas en la realización de ciertas funciones específicas
- la información procedente de los receptores sensoriales es transmitida, serialmente, de unas estructuras neuronales a otras
- una "unidad central de procesamiento" se ocupa de regular la acción consecutiva de las neuronas específicas.

Este modelo, que generó un gran número de expectativas, fue puesto a prueba en diversos experimentos y las conclusiones finales constituyeron un proceso de auténtica "falsación del modelo". Como todo el mundo sabe, la ciencia primero recoge datos de la realidad cotidiana, luego, reflexiona sobre ellos y elabora modelos y teorías, con las cuales se podría explicar la regularidad de los datos. Hasta este punto nos movemos en el campo de la fe: *yo creo que, yo pienso que, en mi opinión, en función de mi experiencia, otros colegas y yo hemos comprobado que*.... Sin pretender restar importancia a la experiencia cotidiana de los profesionales, la adquisición del conocimiento sobre la realidad exige poner a prueba de manera rigurosa

las teorías y modelos. En el caso de que la contrastación empírica de los modelos y las teorías no confirme las predicciones de éstas, sino que las contradiga, entonces tales teorías o modelos resultan "falsados" y no se pueden seguir manteniendo en el futuro en el ámbito del conocimiento científico. Desde luego pueden seguir manteniéndose en el ámbito de la fe. Por ello no es infrecuente encontrarse con colegas que mantienen afirmaciones del tipo: pues yo estoy convencido de..., pues lo que yo veo...., pues eso es discutible (curioso que lo discutible es el proceso de falsación pero no el de aceptación de la teoría,...)

Pues bien, con relación al modelo de procesamiento de la información, los trabajos realizados por diferentes equipos de investigación llegaron a la conclusión de que no podía mantenerse ya que:

- no se han puesto en evidencia que el cerebro esté estructurado globalmente en neuronas que se agrupen para realizar específicamente un tipo de funciones
- no tenemos suficientes neuronas en el cerebro para especializarse en tan gran número de funciones como lleva a cabo el cerebro humano (el animal tampoco)
- la velocidad de procesamiento de la información es muchísimo más alta que la que tendría lugar si verdaderamente el procesamiento fuera serial: primero una función, luego otra, luego otra,...
- no se ha identificado ninguna estructura neuronal que pudiera desempeñar la función de control y regulación de todas las funciones (una unidad central de procesamiento)
- solamente presuponer la existencia de una unidad central de procesamiento supondría admitir un elevadísimo riesgo adaptativo, ya que todas las funciones dependerían de ella (recuérdese lo que ocurre cuando se bloquean nuestros PCs)

Por lo tanto, hace ya bastantes años se desechó el modelo de procesamiento de información como una teoría plausible que explicase el funcionamiento del cerebro y se sustituyó por el Modelo Conexionista (aconsejamos la descarga de la información ofrecida en el link anterior disponible en www.preocupados.es)

Como se afirma en el manual del test, el modelo clínico del ITPA es una adaptación del modelo de comunicación de Osgood (página 9 del manual original) con modificaciones "por problemas prácticos en la construcción del test", según se indicaba en la página 9 del manual original, el cual forma parte de la variedad de modelos y teorías basadas en el modelo de procesamiento de la información. En el nuevo manual, las "alteraciones en el modelo teórico se deben, también, a la observación clínica, además de los problemas anteriormente citados. Destacamos aquí, el reconocimiento de los autores de haber introducido "modificaciones en el modelo teórico original".

En la medida en que este test está basado en el "modelo de procesamiento de la información", con lo cual su finalidad es "evaluar funciones psicolingüísticas" (ficha técnica) y que tal modelo ya se ha descartado desde hace bastantes años como explicación factible del funcionamiento cerebral, podemos afirmar que las

bases conceptuales del ITPA, habiendo sido falsadas, constituyen un criterio suficiente para rechazar la Validez de Constructo del Test.

No obstante, en el nuevo manual de la prueba (6ª edición), destaca el hecho de que los análisis estadísticos llevan al equipo que los realiza a concluir: **"LO QUE APUNTA HACIA…….. LA CONFIRMACIÓN…. del modelo teórico propuesto"**. Eufemismo éste, "SE APUNTA", a nuestro parecer, inaceptable en una publicación de carácter técnico-científico, cuyo empleo necesariamente tiene una gran trascendencia para el destinatario de la misma (el escolar evaluado). A este respecto, se cita textualmente:

a) que no se pudo calcular el coeficiente de consistencia interna (alfa) de dos subpruebas: Integración Visual y Expresión Verbal (página 61)

b) que "aunque el valor Chi cuadrado estuvo por debajo del valor crítico "debido a la sensibilidad de este estadístico al tamaño de la muestra", en ambos casos los valores de ajuste AGFI y RMSEA fueron elevados… ¿Cómo explicar que si el estadístico estuvo por debajo del valor crítico, los resultados son aceptables para confirmar la validez del modelo? ¿Cómo explicar el bajo valor crítico por la sensibilidad del estadístico al tamaño de la muestra? ¿La muestra era insuficiente o era excesiva?

Una vez más se intenta demostrar la validez de un constructo, que, como elaboración teórica que es, sólo puede mantenerse como hipótesis de trabajo o falsarse mediante evidencias empíricas, mediante el uso conveniente de la estadística. El constructo del test se basa en el modelo de comunicación de Osgood, el cual se fundamenta en el modelo de procesamiento de información humana que ya ha sido falsado. Por lo tanto, no cabe intentar acreditar la validez de constructo con un parámetro estadístico, el cual, además, se encuentra "por debajo del valor crítico".

Sobre la Fiabilidad del test

En el año de publicación del test, 1989, el manual técnico de la prueba no ofrecía justificación estadística alguna del instrumento: ni datos de consistencia Interna de cada una de las pruebas, ni datos de consistencia temporal de cada una de las pruebas (test-retest), ni datos sobre la selección de sujetos de la muestra de tipificación. Los baremos que se ofrecieron a los profesionales se obtuvieron con grupos muy reducidos y, por ello, escasamente diversificados y aleatorizados. En el caso de niños de 3 y de 10 años, la muestra era de 40-43 sujetos, la de 4 años es de 93; el resto ninguna llegaba a los 200 sujetos por grupo de edad. Finalmente, no se disponía de ninguna información sobre la Validez Predictiva del Instrumento.

Durante, los años transcurridos desde 1989 hasta el 2004 (15 años) el test se ofreció a los profesionales sin tales datos. Ahora se nos proporcionan estos datos, pero referidos, lógicamente, a los nuevos componentes del test. El usuario puede analizar sin dificultad los datos que se proporcionan en las diversas tablas del manual. Por nuestra parte solo destacaremos que se omite la información referida a los índices de

homogeneidad de cada elemento con el total (tabla 5.6, página 63), dato de gran importancia que se sustituye por el "promedio por edad de esas correlaciones o índices de homogeneidad".

Una vez analizados los elementos de cada prueba, algunos de los cuales, los consideramos, "a priori", inadecuados, estimamos que no informar de los índices de cada elemento impide al evaluador conocer la realidad sobre la calidad de la consistencia interna de cada prueba.

Un solo ejemplo:

En la prueba de integración auditiva se incluyen los siguientes elementos que el niño debe identificar: *tocadiscos*, *chaqueta*, *autocar*, *madrileño*,… Estas palabras corresponden a elementos que resultan prácticamente desconocidos o muy poco empleados, para los escolares de este siglo, lo cual conllevaría una "penalización cultural al sujeto" al valorar su capacidad para llevar a cabo tareas de "integración auditiva". Ni qué decir tiene el hecho de la imposibilidad de emplear este instrumento con población escolar inmigrante, porcentaje cada vez mayor en nuestro país.

Por otra parte, no disponemos de respuesta -por falta de publicaciones o comunicaciones al respecto- de la Utilidad o Validez Práctica del test, más allá de que su empleo permita obtener información -poco fiable- del nivel de adquisición de las diversas habilidades que evalúa.

En este sentido, durante los años 90, por nuestra parte, utilizamos los resultados en este test como medidas de escalas cuantitativas de habilidades (quizás para lo que seguramente se concibió el test) de modo que diseñamos actividades de mejora de las diversas capacidades-aptitudes que se manifestaban claramente deficitarias (sin dar mucha importancia a los baremos). Como ejemplo de esta utilidad del test incluimos un Programa Específico en la publicación PROGENDA hoy en día sustituida por el PROGENDA-2000.

En conclusión: pese a ser un test de posible utilidad práctica, en nuestra opinión, carece de la validez de constructo, predictiva y la fiabilidad suficientes como para su empleo en evaluaciones clínicas o psicoeducativas, en casos de análisis de bajo rendimiento o fracaso escolar, pudiendo sustituirse por otros productos más actuales y relevantes para cada variable que se desee evaluar, obviando la evaluación de procesos cognitivos.

Finalidad	**Evaluar funciones psicolingüísticas implicadas en el proceso de comunicación**

Validez de contenido	**Aceptable, en función del constructo**
Validez de constructo	**No acreditada**
Fiabilidad	**Discutible**
Baremos	**Aceptables pero no relevantes**
Facilidad de Aplicación	**Moderada**

MSCA

Escalas McCarthy de Habilidades Infantiles

(McCarthy, D.1972; Madrid: TEA)

Sobre la Validez de Constructo

Nacida en Minneapolis el día 4 de mayo de 1906 y fallecida el día 28 de enero de 2003, Agnes Dorothea McCarthy se especializó en áreas de desarrollo del lenguaje. Sus trabajos la llevaron a elaborar las **Escalas McCarthy de Habilidades Infantiles**, conocidas como MSCA, cuya publicación en España tuvo lugar varios años más tarde.

Respecto a la validez de constructo de las Escalas MSCA, es preciso destacar que, no siendo un test en sí mismo, sino una batería de diversas pruebas, elaboradas para valorar el progreso en el dominio de diversas habilidades de los niños, no debería considerarse más que la validez de contenido de las mismas. En otras palabras: NO HAY CONSTRUCTO que valorar.

Por ello, lo que resulta sorprendente en este instrumento es la propuesta de la autora de obtener unas puntuaciones combinadas como resultado de la simple adición de puntaciones de diversas pruebas. Aquí es donde resulta factible establecer un punto de discusión:

¿Cuál es el fundamento de la construcción de estas escalas? ¿en qué se basa la autora para establecer los índices que propone?

La respuesta se bien sencilla y la proporciona el manual en su página 11 (selección de las seis escalas): "*la elección del contenido de los tests de la batería y la agrupación de éstos en unas escalas clínicamente útiles, se apoyó, fundamentalmente, en la amplia experiencia docente y clínica de la autora en el campo de la psicología del desarrollo infantil...*"

Es decir, que *las escalas MSCA no tienen fundamento conceptual alguno*, ni su estructura como tal ha sido validada por medio de metodología científica: simplemente son así porque así lo dispuso su autora en base a sus propios criterios (evidentemente, todos los autores adoptan criterios en base a experiencias).

Aún más, pese a que pudiéramos aceptar que cada Escala estuviera compuesta por varios subtests, ¿por qué la puntuación total en cada escala es la suma de cada subtest? ¿por qué se le da el mismo peso a una prueba de memoria pictórica que a una de memoria verbal y que a una de fluidez verbal?

VERBAL (V)

Memoria pictórica + Vocabulario + Memoria verbal + Fluidez verbal + Opuestos

PERCEPTIVO MANIPULATIVA (PM)

Construcción con cubos + Rompecabezas + Secuencia de golpeo + Orientación derecha-izquierda + Copia de dibujos + Dibujo de un niño + Formación de conceptos

NUMÉRICA (N)

Cálculo + Memoria numérica + Recuento y distribución

MEMORIA (MEM)

Memoria pictórica + Memoria verbal + Memoria numérica

MOTRICIDAD (MOT)

Coordinación de piernas + Coord. de brazos + Acción imitativa + Copia de dibujos + Dibujo de un niño

GENERAL COGNITIVO (IGC)

Memoria pictórica + Vocabulario + Memoria verbal + Fluidez verbal + Opuestos + Construcción con cubos + Rompecabezas + Secuencia de golpeo + Orientación derecha-izquierda + Copia de dibujos + Dibujo de un niño + Formación de conceptos + Cálculo + Memoria numérica + Recuento y distribución

La trascendencia de esto se percibe cuando hay desajustes en alguno de los subtests. Al hacer una suma total, sin independizar cada subprueba, **un determinado valor de cada escala:** V, PM, N, MEM, MOT y el INDICE GENERAL COGNITIVO **puede ser el resultado de múltiples combinaciones de resultados sin constituir ninguna información concreta, válida y fiable.**

Además el error es acumulativo, ya que, en primer lugar se comete el error de dar significado a la puntuación combinada verbal, la cual enmascara niveles altos con bajos y medios de Memoria (pictórica y verbal) con Vocabulario, Fluidez verbal y Opuestos. Tras repetir este error en cada una de las Escalas, el mismo se multiplica al calcular el Índice General Cognitivo, puntuación que se suele utilizar en Informes Diagnósticos y, que a la vista de lo anterior, se puede comprender fácilmente que CARECE DE SIGNIFICACIÓN ALGUNA.

Entiéndase ahora el gravísimo error y la injusticia que se comete con niños a quienes se valora con esta prueba y con estos índices para concederles ayudas económicas o técnicas, clasificarlos o diseñar con ellos planes de intervención psicoeducativa.

Consideramos que los comentarios anteriores resultarían suficientes para descartar el empleo de este instrumento, no obstante, conviene profundizar en los aspectos siguientes:

a) la validez de contenido de cada escala resulta aceptable, pudiendo utilizarse como una prueba parcial que constituye una medida del nivel de desarrollo en esa habilidad. Es la combinación aditiva de diversos tests lo que niega la validez a la puntuación total.

b) la fiabilidad de las medidas es deficiente por varias razones; la principal de todas ellas es que al otorgar a cada elemento de cada test la posibilidad de valorarlo con 0, 1 y 2, una puntuación media puede explicarse por varias combinaciones de ceros, unos y doses, cuyo significado es, evidentemente, diferente. Este error, frecuente en los primeros años de desarrollo de la psicometría de las habilidades, ha ido sustituyéndose en los tests más actuales (por ejemplo: BAMADI) mediante un sistema de puntuación dicotómico: objetivo alcanzado (1) o no alcanzado (0). La suma de objetivos alcanzados en cada escala indica el grado de dominio de cada habilidad.

c) otro aspecto que afecta a la fiabilidad de las medidas es la limitación de tiempo, quizás explicable porque en la época de elaboración de estos tests se desconocía las diferencias individuales, de naturaleza biológica, que explica la diferente velocidad de ejecución (que no de procesamiento...) tanto cognitiva, como motriz. Hay niños rápidos y niños lentos, pero esta velocidad de ejecución es totalmente independiente del grado de eficacia en las ejecuciones. De acuerdo a este criterio de valoración los niños lentos han resultado evaluados como "con retraso en el desarrollo", según el MSCA.

Sobre la fiabilidad de las muestras de baremación

Dada la difusión de uso de este instrumento en servicios de Orientación Educativa de nuestras escuelas y Gabinetes Privados de Psicología y Pedagogía, consideramos que los profesionales no han caído en la cuenta de varios aspectos sobre la baremación de este instrumento. A saber:

1. El Manual reeditado en el año 1983, reedición del primero en 1977, es una versión en castellano de la edición en el año 1970, de la Psychological Corporation. Esto quiere decir que el test y sus baremos se elaboraron hace "la friolera" de 39 años.

2. La muestra de baremación es norteamericana, no habiendo sido baremada en España. En su uso comparamos ejecuciones de niños españoles con la de norteamericanos (de los años 70)

3. La muestra de edad está constituida por un centenar de niños de cada grupo de tipificación (eso sí: repartidos en medio urbano-rural y raza blanca-"no-blanca"), lo cual constituye un grupo muy reducido.

4. A pesar de que el manual indica los errores de medida, ni las tablas de baremos, ni las instrucciones de uso, incluyen los intervalos de confianza, lo que permitiría ajustar más a la realidad los resultados.

Nota: en el año 2005 el editor español puso en el mercado una nueva edición del test, hecho que nos resulta muy sorprendente ya que los defectos de construcción de esta prueba desaconsejan de todo punto su empleo en el siglo XXI, dado que existen otras alternativas psicotécnicas mucho más válidas y fiables.

En conclusión: apreciando en lo que vale la gran aportación que supuso este instrumento en el momento de su edición original, en nuestra opinión, el MSCA, puede muy bien formar parte de la historia de la evaluación psicológica.

Finalidad	**Valorar el nivel de desarrollo general de escolares entre 2.5 y 9 años de edad**
Validez de contenido	**Aceptable en cada escala**
Validez de constructo	**Inaceptable**
Fiabilidad	**Escasa**
Baremos	**Irrelevantes en función de su escasa validez de constructo**
Facilidad de Aplicación	**Media**

P.F.S.E.

Test de Actitudes

(Yuste Hernanz, C., 1991; Madrid: CEPE)

Pocos instrumentos diseñados para la evaluación psicopedagógica se pueden encontrar en España integrando tan gran número de despropósitos como éste. Comenzando por su denominación, tan poco afortunada como "test de Actitudes", que lleva al autor, tras una revisión escasa y sesgada de la evolución del constructo "actitud", a no dejar constancia alguna de la definición que asume para la construcción del test.

Con muy poca complejidad pudo muy bien definir actitud como "la predisposición cognitiva a actuar de determinada manera, dada una situación concreta". Esto hubiera facilitado mucho las cosas, porque, al margen de lo que opinara Thurstone a principios del siglo XX, resulta mucho más clara la declaración de Cook y Sellitz, citada por el autor en el mismo párrafo: "disposición fundamental que interviene junto con otras influencias en la determinación de una diversidad de conductas hacia un objeto o clase de objetos, las cuales incluyen declaraciones de creencias y sentimientos acerca del objeto y acciones de aproximación-evitación con respecto de él". Es decir que, de un modo más confuso, estos autores consideran la actitud tal y como indicamos al principio del párrafo:

"*una predisposición cognitiva a actuar de determinada manera, dada una situación concreta*"

Naturalmente, toda actitud tiene un componente afectivo, sobre todo porque no puede darse un instante en la vida consciente de una persona en la que no haya un componente afectivo... Una actitud favorable hacia el deporte conlleva un sentimiento a favor del mismo y al contrario.

Más adelante, el autor indica que (en base a sus propias investigaciones, de las que no da cita alguna) *las actitudes hacia determinados objetos se conforman en torno a factores de grupo*.... afirmación ésta cuyo significado concreto no alcanzamos a entender; para continuar comentando que en las analizadas en el cuestionario presente (suponemos que se refiere a la suma total de elementos del PFSE) se puede hablar de una actitud general.

Entendemos que es en este párrafo (segundo de la página 13 del manual) donde el autor intenta exponer su modelo conceptual sobre las actitudes que va a evaluar con el instrumento que ha diseñado. Por lo tanto, hemos resumido este modelo en los aspectos siguientes:

1. Es posible identificar *una actitud general, dependiente de una única estructura mental de aceptación o rechazo de determinadas conductas relacionadas con los diversos ambientes en los que el sujeto se desenvuelve: colegio, familia, amigos, la cual parece denominar "autoconfianza*"

2. Esta *actitud general estaría configurada por conocimientos, sentimientos y conductas consideradas "positivas" en todos estos ambientes*

3. Tales conocimientos, sentimientos y conductas *"la mente" los dirige y los valora con la finalidad de comprender el mundo en que se vive, proteger el propio concepto de sí mismo, ayudar a adaptarse y posibilitar la expresión de las propias tendencias*,....., añadiendo una frase de significado confuso: ... "la propia visión de su relación con el medio ambiente"

Ante esta descripción del fundamento conceptual del PFSE cabe pronunciarse a favor del mismo (comparto esos planteamientos) o bien discrepar al respecto (no comparto tales planteamientos)

Si el profesional no compartiese este modelo conceptual de las actitudes, no tendría sentido seguir adelante con el análisis del cuestionario: quedaría descartado como instrumento para usar en la práctica profesional o en la investigación,

Ahora bien, en el caso de compartir, aunque fuera provisionalmente tales planteamientos, deberíamos pasar a analizar la Validez de Contenido del PFSE.

Sobre la Validez de Contenido y Constructo

El autor considera el constructo "Actitud" como un factor único, denominado GLOBAL que, a su vez, se subdivide en varias áreas.

Esta estructura resulta de la aplicación de los análisis factoriales que aparecen en la sección correspondiente del manual. En este sentido, el autor no ha partido inicialmente de un modelo concreto, hallando justificación experimental al mismo, sino que ha construido el modelo "a posteriori", en función de los resultados obtenidos con los análisis matemáticos.

Las áreas en que divide el constructo Global son:

Personal, variable cuyo valor corresponde a la suma de las puntuaciones obtenidas en dos sub-variables que denomina **Autoconfianza** y **Fisiológica**.

Autoconfianza, sub-variable que expresa el grado de confianza en sí mismo y en las propias posibilidades.

Fisiológica, sub-variable que expresa el grado de satisfacción con el propio cuerpo, ausencia de temores a enfermedades o carencias físicas.

Familiar, variable cuyo valor corresponde a la suma de las puntuaciones obtenidas en dos sub-variables que denomina **Ante el Padre** y **Ante la Madre**.

Ante el Padre, sub-variable que expresa el grado de aceptación del padre, el grado de confianza y admiración que le suscita.

Ante la madre, sub-variable que expresa el grado de aceptación de la madre, el grado de confianza y admiración que le suscita.

Social: Extraversión, variable que expresa el deseo de contactos sociales.

Escolar: variable cuyo valor corresponde a la suma de las puntuaciones obtenidas en dos sub-variables que denomina **Ante el profesor** y **Ante el estudio**.

Espontaneidad: escala de sinceridad en las respuestas

En primer lugar, estimar que puede admitirse como una variable con significado relevante, la suma de las anteriores, sería como admitir que se pueden sumar las longitudes de pies, manos, dedos, tronco, piernas, cabeza y dar significado al número resultante. No sería igual si sumásemos longitudes de cabeza, tronco y piernas que nos informaría de la estatura; en cambio sumar las anteriores longitudes carecería de significado alguno. Pues bien, sumar Autoconfianza con Fisiológica no puede constituir nunca una variable con significado alguno, sencillamente porque es un principio de matemáticas elementales el no poder sumar "elefantes con margaritas".

Véanse algunos ejemplos de los elementos de cada subescala:

Autoconfianza:

2. ¿Crees que en general tienes buena suerte?

18. ¿hay alguna circunstancia de tu vida que te impide ser suficientemente feliz?

33. La mayor parte de los adultos con los que me relaciono confían en mi

34. ¿Te suelen salir bien las cosas que emprendes?

49. ¿crees que tienes defectos?

Fisiológica:

3. ¿Eres ágil para los deportes o la gimnasia?

4. A menudo pienso que soy demasiado nervioso.

51. Los alimentos que tomo me sientan bien

52. Me molesta saber que soy demasiado alto-bajo

Con relación a las actitudes Familiares se puede considerar algo parecido, ya que, según esta estructura del PFSE, al sumar las puntuaciones de la escala referente a cada uno de los progenitores, la puntuación total

media no permite discriminar si su valor se debe a una muy relación con uno de ellos y muy mala con el otro (dos posibilidades) o bien a una relación media con cada uno de ellos.

Carece pues de valor la puntuación **Familiar** como suma de ambas.

Los mismo sucedería con la Variable **Escolar**, que al ser suma de otras dos, su valor medio no discrimina si es buena con Profesores y Mala con le estudio, al contrario, o media con ambas.

A este respecto, resumiendo, **la validez de constructo del PFSE resulta sumamente deficiente,** lo cual constituye un primer criterio para descartar su uso profesional.

Sin embargo, al llevar a cabo el análisis del contenido de los diversos elementos, la situación se agrava notablemente, ya que su formulación, en muchos casos, no se ajusta al concepto de actitud propuesta por el autor. En efecto, si, según el autor: *actitud = conocimientos, sentimientos y conductas consideradas "positivas"* en ese caso, ¿cómo es posible que formule como considera una posible actitud de un niño o adolescente los pensamientos, sentimientos o conductas de otras personas diferentes a él, como padres, maestros o compañeros, tal y como representan elementos tales como los siguientes:

¿cómo te considera tu padre/madre? (22, 24)

¿hay alguna circunstancia de tu vida que te impide ser feliz?

¿te suelen insistir en que estudias poco? (45)

¿es exigente y severo tu padre/madre contigo? (37, 39)

¿se discute entre los miembros de tu familia? (31)

¿tus profesores están contentos con tu conducta en el colegio? (59)

dentro del hogar mi padre/madre contribuye a la armonía de la familia (70, 72)

a menudo me duelen los dientes o los oídos (84)

en este colegio están muy mal organizados los horarios (75)

No cabe duda que estas algunas de estas informaciones pueden resultar de interés en un caso concreto, pero tal utilidad de los datos no constituye validez de contenido del instrumento, pudiendo muy bien formar parte de un Cuestionario-Entrevista Conductual.

En resumen, la validez de contenido del PFSE es muy deficiente, resultando incongruente la formulación de los elementos con la definición aportada por el autor sobre la variable a evaluar, lo cual constituye un segundo criterio para descartar su uso profesional.

Sobre la Fiabilidad

La fiabilidad es el parámetro de un instrumento psicométrico que permite confiar en él. Muchos factores pueden influir en la fiabilidad, pero especialmente la forma de presentación del instrumento al usuario. Por ello, se suele cuidar la inteligibilidad de las frases, las palabras empleadas, los tiempos verbales,..., que no exijan un nivel de comprensión alto, ya que eso implicaría que no todos los sujetos, con la misma capacidad de comprensión del texto o de la forma de responder al instrumento, se encontraran ante una misma situación estimular.

El PFSE se ha construido siguiendo el modelo de un test clásico, el Cuestionario de Adaptación de Bell (Bell, H., 1937: California: Stanford University Press), con una estructura muy similar en cuanto a sus contenidos (lo que harían más que un test de actitudes un cuestionario de adaptación personal, familiar y social) y los mismos defectos de fiabilidad de él.

La forma de contestar al cuestionario, teniendo que elegir entre tres opciones, es confusa y no permite identificar claramente la actitud del sujeto ante una situación concreta.

Valga como ejemplo lo siguiente:

A menudo pienso que:

Soy un fracasado

Me faltan audacia y "decisión" para triunfar en la vida

Ninguna de las dos anteriores es verdad

Primera cuestión: la palabra audacia no es de uso común, lo que dificulta la comprensión del elemento afectando a la fiabilidad.

Segunda cuestión: el sujeto debe elegir como falta de audacia Y decisión, conjunción copulativa que no tiene por qué darse; puede faltar audacia, pero no faltar decisión. Lo cual también afecta a la fiabilidad.

Tercera cuestión: el autor otorga un punto a la opción en la que el sujeto solamente niega las anteriores, pero no informa de nada en concreto. Se valora con un punto, la "ausencia de algo".

Si, además, recordamos la crítica previa a la construcción del instrumento mediante el cual se suman magnitudes de naturaleza completamente diferentes, concluimos que la fiabilidad del PFSE es muy deficiente, lo cual constituye un tercer criterio para descartar su uso profesional.

Consideramos innecesario comentar nada sobre baremos poblacionales ya que, al tratarse de un cuestionario de variables de personalidad (actitudes, adaptación, autoconcepto, autoestima,...) carece de sentido comparar la puntuación de un sujeto con la media y la desviación típica del grupo población al que pertenece.

En conclusión: en nuestra opinión, el Cuestionario PFSE carece de validez de constructo, su validez de contenido es muy deficiente, así como la fiabilidad por lo que se considera un instrumento no eficaz en el análisis de casos de bajo rendimiento o fracaso escolar en alumnos de Educación Primaria o Secundaria.

Finalidad	**Valorar las actitudes del escolar**
Validez de contenido	**muy deficiente**
Validez de constructo	**nula**
Fiabilidad	**deficiente**
Baremos	**irrelevantes para las variables a medir**
Facilidad de Aplicación	**irrelevante**

Una alternativa al empleo de este test, para evaluar los niveles de adaptación en el ámbito familiar, escolar y personal (indicador de autoestima) lo constituyen las **Escalas Magallanes de Adaptación, EMA**.

PROLEC

Batería de Evaluación de los procesos lectores de los niños de Educación Primaria

(Cuetos, F., Rodríguez, B. y Tuano, E., 1996; Madrid: TEA)

Sobre la Validez de Constructo

La Batería PROLEC es, tal y como lo describen sus autores en el manual, un producto derivado conceptualmente de la Psicología Cognitiva y del Modelo de Procesamiento de la Información (Lindsey y Norman, 1977). Por ambas razones, su validez de constructo sería discutible "en origen". A saber:

1º. La Psicología Cognitiva lleva empeñada años en tratar de observar lo "inobservable". Esto es: los procesos que tienen lugar en el cerebro durante el funcionamiento para realizar algún tipo de tareas. No cabe duda a nadie que en el interior del cerebro, en sus estructuras neuronales, tienen lugar determinadas actividades -diferentes unas de otras- por medio de las cuales cuando el sujeto "ve" un caballo, es capaz de decir en voz alta o de escribir "caballo", en castellano, "cheval" en francés o "horse" en inglés. Hasta aquí no hay discusión. El parecido con el funcionamiento de un PC el cual, al pulsar el botón del icono "Word" inicia una serie de procesos electromecánicos que terminan con la apertura del conocido procesador de textos, diferentes a los que realiza si se pulsa el botón "Excel" o "PowerPoint", es tan sólo aparente. Pero, en ambos casos, lo único que somos capaces de "ver" y "medir" es **el producto de tales procesos**, pero **nunca vemos ni medimos los procesos** que han tenido lugar para obtener uno u otro producto.

2º. Pese a ello, en su momento se consideró como una posibilidad y por ello dio lugar a una "teoría" conocida como "Teoría de Procesamiento de Información Humana", falsada, tal como se explicó anteriormente.

Pues bien, con relación al modelo de procesamiento de la información, los trabajos realizados por diferentes equipos de investigación llegaron a la conclusión de que no podía mantenerse ya que:

- no se han puesto en evidencia que las neuronas se agrupen para realicen específicamente un tipo de funciones
- no tenemos suficientes neuronas en el cerebro para especializarse en tan gran número de funciones como lleva a cabo el cerebro humano (el animal tampoco)
- la velocidad de procesamiento de la información es muchísimo más alta que la que tendría lugar si verdaderamente el procesamiento fuera serial: primero una función, luego otra, luego otra,...
- no se ha identificado ninguna estructura neuronal que pudiera desempeñar la función de control y regulación de todas las funciones (una unidad central de procesamiento)
- solamente presuponer la existencia de una unidad central de procesamiento supondría admitir un elevadísimo riesgo adaptativo, ya que todas las funciones dependerían de ella (recuérdese lo que ocurre cuando se bloquean nuestros PCs)

Por lo tanto, hace ya bastantes años se desechó el modelo de procesamiento de información como una teoría plausible que explicase el funcionamiento del cerebro y se sustituyó por el Modelo Conexionista.

Pues bien, tanto desde la perspectiva cognitivista, que pretende evaluar lo inobservable, como desde los supuestos básicos de la teoría de procesamiento de la información (gráfico 1 en página 10 de manual) que ha sido falsada y, por lo tanto, descartada como una posibilidad real de funcionamiento del cerebro, se puede afirmar que **el PROLEC carece de validez de constructo.**

En cuanto a la Validez de contenido, dado que el objetivo o la finalidad del test consiste en evaluar los procesos lectores, estimamos una grave contradicción entre sus contenidos y dicha finalidad. La apariencia de la Batería es la de un producto psicotécnico que pretende evaluar "productos" y no "procesos". ¿Cuál si no es la utilidad de la valoración cuantitativa de los resultados y la aparición de baremos poblacionales? Los procesos no "se bareman".

Respecto a las pruebas que contiene el PROLEC:

IDENTIFICACIÓN DE LETRAS

1. Nombre o sonido de letras; el niño debe decir el nombre de unas letras

2. Lectura de palabras y pseudopalabras; el niño debe leer pares de palabras e indicar si son iguales o diferentes (es una prueba que requiere eficacia atencional y puede ser deficitaria en niños que leen bien con baja eficacia atencional y, por lo tanto no analizan todos los elementos de la palabra)

PROCESOS LÉXICOS

3. Decisión léxica; ante un listado de 30 palabras se pide al niño que indique si cada una de ellas se trata de una palabra real o inventada. Con esta prueba se trata de saber si conoce la palabra como un símbolo gráfico total (no sabe leerla) o bien si la lee. ¿cómo puede estar seguro el evaluador de que la respuesta se produce por medio de uno u otro proceso? Debe tenerse en cuenta que la probabilidad de responder al azar es del 50% (¿real o inventada?). Requiere además una cantidad de vocabulario disponible.

4. Lectura de palabras; se le solicita que lea en voz alta palabras y pseudopalabras. ¿cómo sabe el evaluador que los errores en lectura de pseudopalabras no implica sencillamente una falta de dominio en los mecanismo lectores y lo atribuye a diferentes procesos de lectura?

5. Lectura de pseudopalabras

6. Lectura de palabras y pseudopalabras

En ambos casos, las diferencias en eficacia de ejecución se atribuye a un proceso cognitivo diferente.

PROCESOS SINTÁCTICOS

7. Estructuras gramaticales o comprensión de oraciones; el niño debe leer tres frases situadas debajo de un dibujo cada una de las cuales describe la situación de un modo diferente, pero sólo uno es adecuado. Se supone que el niño tiene un nivel de lectura alto, lo cual requiere que haya superado satisfactoriamente las pruebas anteriores, las cuales quedarían invalidadas. En caso contrario, que no haya superado las anteriores, es técnicamente imposible que el niño supere esta prueba. Esta prueba requiere como variable de control conocer la eficacia atencional y el nivel intelectual del sujeto ya que ambas variables modulan el éxito en esta tarea, por lo que resulta irrelevante para identificar procesos cognitivos.

8. Comprensión de textos; se proporciona 4 textos al niño y se le formulan preguntas sobre el mismo. Cada texto tiene 4 oraciones y se le formulan 4 preguntas cuyas respuestas son 2 explícitas en el texto y 2 inferenciales.

Obviamente, el diseño de estas pruebas resulta acorde con el modelo conceptual del test y con sus objetivos. La cuestión es que el modelo propuesto es un modelo teórico, no un modelo real y que, lejos de estar validado, se encuentra falsado. De modo que cualquier conclusión que pudiera extraerse de este test resultaría siempre una hipótesis, nunca una evidencia empírica. La diferencia con este tipo de pruebas (cognitivas) y otras aparentemente similares pero de base conductual (TALE, TALE-2000) es que las segundas sí pueden llevar a cabo afirmaciones reales, de base empírica, ya que solamente aspiran a medir el grado de competencia, de eficacia, de capacidad, para llevar a cabo una tarea. Es decir, en el caso que nos ocupa, de informar pericialmente del grado de dominio de la habilidad lectora de un escolar.

En cuanto a la **validez criterial de la Batería,** los autores nos ofrecen un estudio sorprendente por su originalidad y por sus resultados: los profesores puntuaron a los escolares que participaron en la prueba en habilidad lectora según sus criterios subjetivos: lo que ellos estimaban como capacidad lectora, de 1 a 10, y se correlacionó esta puntuación con la obtenida en cada una de las pruebas. El "sorprendente" resultado es que ninguna correlación llegó ni siquiera al 50%, resultando escandalosas la correlaciones de 0,26 y de 0.38 en las pruebas más sencillas: nombre de letras e igual-diferente. Más aún, siguiendo al costumbre (muy mala por cierto) de sumar todo, una puntuación total de la batería (carente del más mínimo significado) correlacionó 0,56 con el criterio de los profesores. Estos resultados lo que acreditan empíricamente es que la Batería PROLEC no tiene mayor valor predictivo sobre el dominio de la lectura que "echar una moneda a cara-cruz".

Sobre la Fiabilidad

En cuanto a la fiabilidad de la Batería PROLEC, el único dato que aportan los autores es una consistencia interna de 0,92. Pero, ¿consistencia interna de qué escala?. Hemos de suponer que de la Batería considerada en su globalidad. En tal caso, ¿qué significación puede tener tal coeficiente?. ¿cómo es posible incluir tan diferentes clases de pruebas, que evalúan -supuestamente- diferentes "procesos cognitivos". Lo razonable sería estudiar la consistencia interna de cada una de las escalas.

En cuanto a la fiabilidad test-retest, ¿es consistente el resultado de esta Batería con el tiempo (dos meses quizás)? No se aportan datos al respecto.

Y, por otra parte, ¿en qué medida se puede fiar el evaluador de una prueba de comprensión lectora que tiene la mitad de los elementos textuales y la otra mitad inferenciales?. Cuando el sujeto saca la máxima puntuación en cada texto, se puede asegurar que domina ambas habilidades, pero cuando saca la puntuación media, ¿qué habilidad posee la de comprender lo que lee textualmente o lo que lee inferencialmente?

Si nos atenemos a la puntuación total, de 0 a 16, ¿qué valor otorgaremos a una puntuación de 6 o de 8 (media): ¿el sujeto comprende haciendo inferencias o comprende sólo lo que lee textualmente?

Finalmente, en cuanto a lo que se refiere a los Baremos obtenidos con la muestra de tipificación, los sujetos que integran estos baremos son grupos de, aproximadamente, 100 sujetos (102, 100, 98 y 103 respectivamente) por nivel escolar de 1º a 4º de Primaria. De todo punto inadmisible que el evaluador proceda a comparar los resultados de un escolar en evaluación con la media de un grupo tan reducido como un centenar.

Y una curiosidad: ¿qué fundamento dan los autores a dividir la muestra de tipificación en colegio público y colegio privado; colegio urbano, colegio rural? ¿acaso los cerebros de los niños tienen "procesos cognitivos" diferentes dependiendo del colegio al que van?. No se aportan resultados estadísticos que informen de alguna de estas diferencias.

Resumiendo: la Batería PROLEC no tiene acreditada su validez de constructo al encontrarse basada en un modelo, el de procesamiento de la información, que ya ha sido falsado. Tampoco su validez de contenido corresponde con el constructo en el que se fundamenta y su validez criterial es menor que el azar. La fiabilidad interna no se ha establecido de manera adecuada, es inexistente la información sobre la consistencia temporal (test-retest). Por otra parte, sus baremos son sumamente escasos, poco diversificados y no suficientemente representativos de la población a la que se dirige.

En conclusión: dadas las características de este instrumento, en nuestra opinión, carece de la validez de constructo, predictiva y fiabilidad suficientes como para su empleo en evaluaciones clínicas con valor pericial o en psicopedagógicas en casos de análisis de bajo rendimiento o fracaso escolar, pudiendo sustituirse por otros productos más actuales y relevantes para cada caso.

Finalidad	**Evaluar procesos cognitivos en la Lectura**
Validez de contenido	**Insuficiente**
Validez de constructo	**Nula**
Fiabilidad	**Muy deficiente**
Baremos	**Insuficientes**
Facilidad de Aplicación	**Moderada**

Para la evaluación de las habilidades de lectura y escritura puede considerar la posibilidad de emplear un test de fundamento conductual, adecuadamente baremado con muestras amplias y diversificadas como es el TALE-2000. En este caso encontramos un ejemplo de acumulación de conocimiento: se realizan cambios significativos en un test elaborado en los años 70, aprovechando el modelo conceptual y los elementos que han resultado útiles en el pasado y sustituyendo los inadecuados.

STAI-C

Inventario de Autoevaluación Ansiedad Estado/Rasgo para Niños

(Spielberger C.D. y cols., 1973. Edición española en 1989 por TEA Ediciones)

Inadecuadamente traducido del original "Inventario", como "Cuestionario", este instrumento adolece de serias deficiencias conceptuales y metodológicas que lo hacen poco o nada útil para la evaluación de casos de ansiedad infantil. En primer lugar, se debe destacar que este Inventario (listado de indicadores) se diseñó con fines de investigación en escolares de enseñanza primaria (6 a 12 años) y no como un instrumento de evaluación clínica.

En segundo lugar, el Manual Técnico del Instrumento no recoge ni los fundamentos teóricos ni los procedimientos de construcción, remitiendo al profesional a los manuales originales en inglés (publicados en 1966, 1970, 1971 y 1972). Esto hace especialmente difícil para los profesionales en general el análisis de estos aspectos del instrumento.

Pese a las carencias conceptuales mencionadas, del análisis del propio instrumento, podemos destacar lo siguiente:

a) En la parte A del instrumento se pretende valorar el nivel de ansiedad en el momento actual (no la ansiedad con relación a ninguna situación en concreto, sino la ansiedad en el momento actual. Esta valoración se refiere EXCLUSIVAMENTE al momento de la ejecución del cuestionario, ya que las instrucciones son muy concretas:

"señala la respuesta que diga mejor cómo TE SIENTES AHORA MISMO"

Evidentemente, esta información resulta de muy poca o ninguna utilidad en un caso clínico, ya que no nos informa para nada de la ansiedad del sujeto fuera del momento actual. El niño puede estar muy tranquilo durante la realización de cuestionario, en la sala de evaluación y en presencia del evaluador, pero este estado puede cambiar al salir de la sala y terminar la evaluación, o al contrario.

En cuanto a su VALIDEZ DE CONTENIDO, si la variable a medir es ANSIEDAD, variable fisiológica, carecen de validez alguna las respuestas a los elementos:

Me encuentro confuso

Me encuentro contrariado

Tengo miedo

Me encuentro atemorizado

Estoy preocupado

Me siento molesto

Me siento triste

En modo alguno estos elementos constituyen una medida directa o indirecta de respuestas fisiológicas de ansiedad. Sin embargo se acumulan a las respuestas a los elementos:

Me encuentro inquieto

Me siento nervioso

Me siento angustiado

Los cuales son tres formas diferentes de informar de lo mismo, pero que realmente son indicadores fisiológicos de ansiedad.

b) En la Parte B del cuestionario se pretende evaluar la presencia de ansiedad de manera generalizada en la vida del niño:

"señala la respuesta que diga mejor cómo TE SIENTES EN GENERAL"

y su validez de contenido es igualmente muy deficiente ya que incluye pensamientos (variable cognitiva) como una medida de ansiedad (variable fisiológica):

Me preocupa cometer errores

Me cuesta tomar una decisión

Me preocupo demasiado

Me preocupan las cosas del colegio

Me cuesta decidirme en lo que tengo que hacer

Me preocupa lo que los otros piensen de mi

Encuentro muchas dificultades en mi vida

En modo alguno estos elementos constituyen una medida directa o indirecta de respuestas fisiológicas de ansiedad. Sin embargo se acumulan a las respuestas a los elementos:

Noto que mi corazón late más rápido

Tengo sensaciones extrañas en el estómago

Y vuelve a incluir en la medida de la ansiedad, sentimientos de tristeza:

Siento ganas de llorar

Me siento desgraciado

Me siento menos feliz que los demás chicos

Por todo lo anterior, desacreditada totalmente la VALIDEZ DE CONTENIDO y CONSTRUCTO, no merecería la pena entrar a considerar la fiabilidad del instrumento, descartando de manera absoluta el uso de este test en evaluación clínica.

No obstante, merece la pena destacar un error común a otros muchos instrumentos de construcción similar: el hecho de puntuar de manera cuantitativa y no dicotómica cada elemento de cada escala hace imposible

interpretar una puntuación determinada. Por una parte constituye una auténtica barbaridad y un desatino psicométrico otorgar valor de 1 punto a "Nada".

Así, si un sujeto responde "Nada" a: Me encuentro inquieto + Me siento nervioso + Tengo miedo + Estoy preocupado + Me siento molesto + Me encuentro atemorizado + Me encuentro confuso + Me siento angustiado + Me siento contrariado + Me siento triste, según este sistema de puntuación, cuenta con 10 puntos.

Esta puntuación resultaría equivalente a responder "Algo" a: Me encuentro inquieto + Me siento nervioso + Estoy preocupado + Me encuentro atemorizado + Me siento angustiado , lo que conlleva puntuación 10.

o bien a: Tengo miedo + Estoy preocupado + Me siento molesto + Me encuentro atemorizado + Me encuentro confuso, lo que también conlleva puntuación 10.

Se comprende así la falta de confiabilidad absoluta en las puntuaciones de este Inventario.

Finalmente, ¿qué significado pueden tener unas tablas de baremos? ¿Comparar al individuo en evaluación con la media poblacional? Y ¿cómo interpretar una puntuación media? ¿Que no tiene ansiedad estado/rasgo relevante, o que tiene la misma ansiedad que la media de la población?. En este último caso, si toda la clase del Colegio/Instituto está "nerviosa" los escolares saldrán evaluados con ansiedad "normal"...

En conclusión: en nuestra opinión el instrumento adolece de tales deficiencias de construcción que lo hace descartable como medio de evaluación clínica en casos de niños con Problemas de Ansiedad o Estrés.

Finalidad	**Valorar el nivel de ansiedad situacional y la ansiedad general en Niños de 9 a 15 años**
Validez de contenido	**Nula**
Validez de constructo	**Nula**
Fiabilidad	**Pésima**
Baremos	**Irrelevantes**
Facilidad de Aplicación	**Alta**

TAMAI

Test Evaluativo Multifactorial de Adaptación Infantil

(Hernández y Hernández, Pedro,1988; Madrid: TEA Ediciones)

Sobre la validez de constructo:

Afirma el autor en la introducción del Manual de este Test lo siguiente:

"*Un adecuado análisis del rendimiento académico exige el conocimiento de factores de personalidad del alumno y, más concretamente, de su adaptación personal, escolar, social y familiar.....*

....Es más, el conocimiento de la adaptación de los educandos debe ser por sí mismo, un objetivo operativo de la educación, además del rendimiento académico.."

Bien, partimos de dos supuestos que, a falta de un modelo conceptual al que se adhiera el autor, constituyen dos afirmaciones personales y gratuitas del mismo, que, por supuesto no compartimos.

1. Con relación a la necesidad de conocer factores de personalidad para llevar a cabo un "adecuado análisis del rendimiento académico", hubiera sido muy deseable que el autor explicara de qué manera, qué factores de personalidad, influyen en el rendimiento escolar...

¿Es acaso la extraversión un factor de pronóstico favorable o desfavorable? ¿cuáles son los factores que lo favorecen y cuáles lo perjudican? Pero, sobre todo, ¿cuál es el modelo conceptual en el que se fundamenta el autor para incluir este test y sus variables de modo que contribuyan a la explicación del alto o bajo rendimiento escolar?.

La ausencia total de fundamento conceptual de esta prueba nos obliga a proseguir su análisis de una manera cautelosa, a la espera de encontrar información relevante más adelante.

2. Por otra parte, la educación siempre aparece en todos los manuales de Pedagogía como una acción dirigida a los educandos, por lo que conocer la adaptación social, familiar y personal de éstos, no tiene ningún sentido que sea un objetivo operativo de la misma. Para más detalles puede consultar a Bloom, un autor del máximo prestigio en este campo.

Otra cuestión muy diferente sería considerar que los Profesores tutores, en cumplimiento de su función tutorial, deben conocer diversas características de los escolares; entre ellas, sus niveles de adaptación, de modo que puedan considerar -en su caso- la existencia de factores de estrés en la vida del escolar, que pudieran ser detectados por aspectos de inadaptación familiar, social o personal.

Aún más: si el test trata de evaluar variables de los educandos: ¿qué papel juegan las variables siguientes?:

Educación Adecuada (Pa) (M) (estilo educativo de los padres evaluados por el hijo), subdividida, a su vez en subvariables:

Educación Asistencial-personalizada (Pa1) (M1)

Permisivismo (Pa2) (M2)

Restricción (Pa3) (M3)

Estilo Punitivo (Pa31) (M31)

Estilo Despreocupado (Pa32) (M32)

Estilo Perfeccionista (Pa33) (M33)

Discrepancia Educativa (Dis) (diferencias entre el estilo de educación del padre y de la madre)

Ambas (Pa) (M) y (Dis) no corresponden a características del sujeto en evaluación, ¿cómo es posible que más adelante el autor las considere de la misma naturaleza y efectúe una correlación (fiabilidad) entre estos elementos y los de escalas de características del sujeto.

Y, de otra parte, ¿qué sentido tiene la subdivisión en variables de segundo y tercer nivel de la variable Educación Adecuada?. Insistimos en la incongruencia de definir estas variables por el resultado de un ejercicio estadístico como es el Análisis Factorial de Correspondencia, tal y como indica el autor en la página 2 del manual. Aún más, ¿de donde procede la muestra poblacional de la que se ha generado este test? Si, como se indica en dicha página, el número de sujetos es de 1.200 (varones, por cierto; no hay mujeres en la muestra con la que **se construye este test**) de 3º de EGB hasta COU, esto proporciona un número de 1.200/10 grupos = 120 sujetos por grupo escolar. Probablemente sea todos de un mismo Centro Escolar o a lo sumo de dos Centros. ¿Cómo es posible dotar de estructura factorial a un test a partir de una muestra poblacional tan sesgada (sexo único, ciudad única, posiblemente centro único?)

Convendría recordar aquí que la estadística, los procedimientos estadísticos deben estar al servicio de la conceptualización científica y no al contrario.

El autor debía haber comenzado por diseñar un test de acuerdo a un modelo conceptual (fuera este cual fuese), lo que constituiría su hipótesis de trabajo científico. La administración a una población muestral amplia y diversificada, permitiría poner a prueba la hipótesis de la estructura conceptual del test. Pero no al contrario, tal y como lo ha hecho el Dr. Hernández.

Son diversas las pruebas que, en tono de humor, ponen de manifiesto los gravísimos errores de considerar los resultados estadísticos como una prueba de la realidad cotidiana. Por ejemplo:

- **las estadísticas muestran que casi todos los accidentes de circulación se producen entre vehículos que ruedan a velocidad moderada. Muy pocos ocurren a más de 150 Km. por hora. ¿Significa**

esto que resulta más seguro conducir a gran velocidad?. No, de ninguna manera. Con frecuencia, las correlaciones estadísticas no reflejan causas y efectos. Casi todo el mundo circula a velocidad moderada, y como es natural, la mayoría de los accidentes se producen a estas velocidades.

- **Si las estadísticas mostrasen que la mortalidad por tuberculosis es mayor en Segovia que en las demás provincias, ¿significaría esto que el clima segoviano favorece el contagio tuberculoso? Todo lo contrario. El clima segoviano es tan beneficioso para los tuberculosos que muchos acuden allí para restablecerse. Naturalmente, ésta es la causa de que aumenten allí los fallecimientos provocados por el mal.**

- **Un reciente estudio psicopedagógico ha mostrado que los niños de pie grande saben leer mejor que los de pie pequeño. ¿Permitirá el tamaño del pie medir la capacidad de lectura de los niños? No, desde luego. El estudio se hizo sobre escolares que están en crecimiento. Todo cuanto se demostró en él es que los niños mayorcitos, cuyos pies son más grandes, leen mejor que los pequeñines.**

Tomado de la web: http://platea.pntic.mec.es/jescuder/estadist.htm

****

La lectura del manual del test nos lleva de sorpresa en sorpresa ya que el autor, como comentamos anteriormente, no parte de una definición del constructo ADAPTACIÓN, sino que el constructo que pretende medir se define por el análisis factorial al que somete los resultados del test en una muestra de población. Esta ha sido una práctica frecuente en el siglo pasado en los instrumentos diseñados bajo el principio "*vamos a ver que encontramos*". Así, el manual omite el método seguido para elaborar los ítems de cada escala, lo cual nos permitiría analizar su validez de contenido.

De modo que, no habiendo definido por ninguna parte las variables que se pretenden medir, se incluyen en un test de adaptación elementos tales como:

Creo que soy bastante vago, Soy muy vergonzoso, Me gustaría tener menos edad, Soy muy miedoso, Soy muy chistoso y hablador, Soy muy cuidadoso con las cosas,...

los cuales podrían aceptarse o rechazarse a priori, si se hubieran definido las variables de manera concreta y operativa, lo cual no ha sucedido.

Sobre la consistencia interna del TAMAI:

El Manual del test **omite los datos sobre los índices de homogeneidad de cada ítem**, constitutivo de cada escala.

Las Escalas Principales se subdividen en otras, no por un principio conceptual, sino que, a la vista de los resultados factoriales, se describen como componentes de la variable principal a posteriori; por supuesto, sin justificación conceptual ni estadística expresa.

En conclusión: los posibles usuarios de la prueba deben utilizarla como un acto de fe, ya que no se ha fundamentado más allá que en un juego de números, tal y como se suele denominar a las operaciones matemáticas y especialmente a la estadística.

En cuanto al coeficiente de consistencia interna, el autor vuelve a sorprendernos, ya que nos da un valor único: índice de fiabilidad = 0.87

y ¿cuál es el significado de este índice? ¿el de una supuesta escala total en la que se mezclan los factores que se han evidenciado diferentes?

Y, en cuanto a la fiabilidad test-retest, ¿en qué medida es consistente la respuesta de los escolares a este instrumento en un intervalo de tiempo prudente?

No hay referencia alguna a este índice.

Sobre los baremos

El manual ofrece unos baremos poblacionales, obviamente obtenidos con la misma muestra con la que se ha factorizado; esto es: no constituyen en modo alguno unos baremos de población general.

Siendo así, que esta característica de la prueba la invalida como método de comparación de un sujeto con la población de referencia, todavía podemos incidir aún más en el hecho de que, para grupo de baremación las muestras son muy escasas: inferiores a 150 por grupo de edad y no aleatorizadas ni generalizadas.

Sobre la utilidad

El usuario de este instrumento, a la vista del manual, debe plantearse: ¿para qué me pueden servir a mí estos datos?

¿Voy a poder establecer una relación funcional entre cada uno de los datos obtenidos y el rendimiento escolar de un alumno en estudio?

¿Donde están los estudios de validez predictiva de cada una de estas variables y subvariables?

Sobre la facilidad de uso:

La utilización de la Hoja de Respuestas del TAMAI es cómoda y sencilla para los escolares, pero no así para el evaluador, quien debe emplear unas plantillas complejas.

Por todas las consideraciones anteriores, consideramos el TAMAI como un instrumento que carece de suficiente validez de constructo y de fiabilidad como para tener en cuenta sus resultados en procesos de análisis del bajo rendimiento o fracaso escolar.

Finalidad	**Valorar diversos niveles de adaptación de los escolares desde los 8 años de edad**
Validez de contenido	**Escasa**
Validez de constructo	**Muy discutible**
Fiabilidad	**Escasa**
Baremos	**Deficientes**
Facilidad de Aplicación	**Irrelevante**

Una alternativa al empleo de este test, para evaluar los niveles de adaptación en el ámbito familiar, escolar y personal (indicador de autoestima) lo constituyen las **Escalas Magallanes de Adaptación, EMA**, de las que pueden obtener amplia información en www.gac.com.es

TEST TOULOUSE-PIÈRON

Prueba perceptiva y de Atención

(E. Toulouse y H. Pièron, 1904)

(Versión española preparada por el Dr. Mariano Yela; Madrid: TEA)

Probablemente este test sea el primero, concebido a principios de siglo (1904), para medir la capacidad de atención sostenida (concentración, resistencia a la monotonía y la rapidez perceptiva)

En cuanto a la validez para medir la atención sostenida de los sujetos, dado que el tiempo normativo son 10 minutos, puede considerarse suficiente; no obstante, la puntuación directa que, a diferencia de otros tests (caras o formas idénticas) ya tiene en cuenta los errores y las omisiones, sigue sin valorar el número de figuras supervisadas, lo que hace difícil comparar la eficacia de sujetos de ejecución rápida y los de ejecución lenta.

Aún más, la velocidad perceptiva que pretende medir [ausente totalmente en los baremos del test] , es un objetivo técnicamente imposible, ya que el procedimiento a seguir es un método que incluye como tiempo total el de percepción más el de ejecución. El sujeto de percepción lenta (hablamos de milisegundos...) pero de ejecución rápida se confundirá siempre con el de percepción rápida (algunos milisegundos menos) pero de ejecución lenta. Ahora bien, como nadie destaca esta incongruencia conceptual, ahora tenemos el mismo error en tests mucho más actuales como el WISC-IV, el cual, cometiendo el mismo error metodológico, aspira a informar de la Velocidad de Procesamiento Visual, a través del tiempo de ejecución del subtest CLAVES.

Dada la ausencia de datos sobre la justificación estadística del test, poco podemos analizar al respecto. No obstante, en cuanto a su fiabilidad (test retest), probablemente, en atención a sus características sea Aceptable e indiscutiblemente Buena su consistencia interna.

En cuanto a las muestras de baremación, muy escasas y sin justificar su diversidad y aleatorización.

En cualquier caso, como prueba diseñada hace más de un siglo (1904), debemos considerar su valor como instrumento pionero en los estudios sobre atención sostenida, pero muy difícil de justificar su empleo en el siglo XXI, cuando se dispone de pruebas de calidad psicotécnica muy superior a ésta. Véase por ejemplo las Escalas Magallanes de Atención Visual, que, con muestras muy amplias, diversificadas y aleatorizadas, elaboradas hace muy pocos años, informa de nuevos parámetros atencionales como son la Atención Sostenida, la Eficacia Atencional y la Estabilidad de la Atención.

En conclusión: Apreciando en lo que vale la gran aportación que supuso este instrumento con el nacer del siglo XX, en nuestra opinión, el test de Toulouse-Pièron, puede muy bien formar parte de la historia de la evaluación psicológica y archivarlo en el lugar correspondiente.

Finalidad	**Valorar la capacidad de atención sostenida y la rapidez perceptiva a partir de los 9 años**
Validez de contenido	**Aceptable en función del constructo**
Validez de constructo	**Escasa**
Fiabilidad	**Aceptable**
Baremos	**Escasos, Anticuados y Poco Relevantes**
Facilidad de Aplicación	**Alta**

WISC-R

Escalas de Inteligencia para Niños de Wechsler

(Wechsler, D. 1974; versión española en 1993; Madrid: TEA)

A diferencia de la mayoría de los modelos conceptuales propuestos por diversos expertos en el estudio de la inteligencia, Binet, Terman, Spearman, Thurstone, Hebb, Raven, Das, Sternberg,… David Wechsler adoptó un punto de vista muy diferente para elaborar y evaluar el constructo inteligencia.

Obviamente, en tanto en cuanto la inteligencia es un "constructo", no cabe discutir sobre la "opinión que una persona, profesional o no, se forma sobre él". Solamente es posible llevar a cabo una discusión formal, ideológica, argumentada sobre diversos elementos, de modo que se pueda establecer algún tipo de conclusión, basada en la valoración de su aplicación y utilidades.

Desde el principio, cuando publicó las primeras escalas de inteligencia, el autor ha propuesto un modelo conceptual de la inteligencia amplio. Para Wechsler, la inteligencia es *el conjunto total de recursos de un individuo para adaptarse al medio*. O sea, más o menos, todas las habilidades de que dispone una persona para vivir, respondiendo a situaciones del contexto físico y social con más o menos éxito y bienestar personal, ya que eso es la adaptación.

Obsérvese cómo este modelo constituye una mezcla de inteligencia lógica, atención y memoria de conocimientos; no ajustándose, desde su propia definición, al acuerdo generalizado sobre el concepto de inteligencia.

En función de este concepto de inteligencia, David Wechsler elaboró una primera versión **en el año 1939** de una batería de pruebas conocidas como Escalas Wechsler de Inteligencia, que según los niveles de aplicación se denominarían: WPPSI, WISC y WAIS.

Estructuradas en forma de seis pruebas que requieren el empleo de habilidades lingüísticas y otras seis que no lo requieren, ofrece una medida global de cada una de las seis pruebas, denominadas Cociente Intelectual Verbal (CIV) y Cociente Intelectual Manipulativo (CIM). Una puntuación combinada de las puntuaciones anteriores representaría el Cociente Intelectual Global de una persona. Los efectos prácticos de este modo de establecer el CI son que dos personas con habilidades completamente diferentes pueden tener el mismo CI Global. Así, una persona con un CI Verbal muy bajo y un CI Manipulativo muy alto, tendrá un mismo CI global medio que otra con un CI Verbal muy alto y un CI Manipulativo muy bajo, lo que dejaría en evidencia la ausencia de validez discriminante de la prueba.

Evidentemente, si aceptásemos el concepto de inteligencia de Wechsler, podríamos asumir la estructura de sus tests. Ahora bien, ¿cómo es posible aceptar como inteligencia la capacidad-habilidad para …?

- recordar las estaciones del año, conocer de qué animal obtenemos la leche, el mes posterior a marzo,... (Información)
- retener en memoria una secuencia de números y repetirla en orden directo o inverso, inmediatamente (Dígitos)
- conocer el significado de palabras poco frecuentes (Vocabulario)
- construir una figura sin modelo a partir de partes de la misma (Rompecabezas)
- encontrar partes significativas de una figura incompleta (atención) (Figuras incompletas)
- retener en memoria una asociación de pares de figuras geométricas o figuras-números (claves)

¿Resulta admisible como indicador de inteligencia conocer el resultado del producto 5-1, 4+2,...?.

Conceptualmente, la definición de inteligencia de la que parte el autor y en la que se basa este test, es muy poco admisible, a la luz de las investigaciones y conocimientos del siglo XXI.

En el mejor de los casos, las escalas Wechsler tendrían que establecer una diferencia importante y significativa entre la capacidad-habilidad de desenvolvimiento práctico: experiencias de aprendizaje previas y destrezas de memoria, y la capacidad-habilidad de razonamiento; algo que ni su autor propone, ni resulta factible dada la estructura del test.

Algunos autores han considerado el WISC como una prueba factorial, sin caer en la cuenta de que los diversos factores, en caso de existir como tales, deberían tener un peso específico cada uno de ellos. Si el valor de cada factor fuera equivalente al de los demás, uno cualquiera de ellos podría ser suficiente indicador de la inteligencia del sujeto, estando de más el resto. Sin embargo, si cada factor tuviese un peso específico diferente, ¿cómo se explica que la puntuación total sea una "media de las puntuaciones en cada factor"? En realidad, el sistema de puntuación trata a las diversas subpruebas como factores con el mismo peso, sin que jamás se haya puesto en evidencia esa realidad. Así pues, desde el punto de vista conceptual, resulta inaceptable admitir este instrumento como un test para evaluar la inteligencia de los sujetos, a menos, claro está, que no importe tener una medida resultado de la suma de magnitudes diferentes; esto es, diversos tipos de destrezas y capacidades. Para más detalles puede consultarse el análisis realizado sobre la versión IV de estas escalas, en esta misma obra.

Desde un punto de vista metodológico, cada prueba proporciona una medida aceptable -en el caso de disponer de baremos obtenidos a partir de muestras amplias y representativas de la población de referencia- de alguna capacidad o destreza concreta: vocabulario (semántica), cubos y rompecabezas (organización viso-espacial y coordinación motriz), comprensión y semejanzas (razonamiento lógico), dígitos (memoria verbal inmediata),...

Si el método de valoración de las distintas capacidades-habilidades fuera el de un Perfil de Habilidades, mediante el cual, las distintas puntuaciones no se suman para obtener CI alguno, sino que se analizan y valoran por separado; en tal caso, podría resultar útil para comprender cada situación personal y diseñar Programas de Intervención. Sin embargo, la propuesta de D. Wechsler: la suma algebraica de todas las puntuaciones obtenidas en cada subprueba para obtener una puntuación global, lleva a situaciones frecuentemente absurdas y poco o nada operativas. En efecto, supóngase una elevada destreza en conocimientos prácticos, memoria y cálculo aritmético y un déficit significativo en razonamiento. La suma de todas las puntuaciones indica que el sujeto tiene un "nivel medio", ignorando el hecho de un importante déficit de razonamiento, que podría explicar algunas de sus dificultades escolares o personales. Necesariamente debemos concluir que el empleo de estas escalas en la forma que propone su autor es, metodológicamente, inadecuado, no resultando útiles nada más que para detectar sujetos con déficits importantes o bien con destacadas destrezas en áreas diversas.

Por otra parte, los estudios de consistencia interna de las distintas subpruebas, mostrados a través de las matrices de correlaciones entre ellas resultan muy variables según los distintos grupos de edad, lo que impide aceptar las propiedades psicométricas del instrumento. En cuanto a la validez de pronóstico, ésta siempre se establece con respecto al criterio de éxito escolar, resultando bastante aceptable lo cual es comprensible dado los contenidos, cuasi-curriculares de las subpruebas verbales. Sin embargo, con relación al 25% de la población con dificultades de rendimiento escolar, las correlaciones entre los CIs obtenidos y el rendimiento suele ser inferior a 0.40 lo cual no ha sido nunca explicado de manera consistente por los defensores de este instrumento (véase manual del test)

Evidentemente, los riesgos de error, al realizar evaluaciones de capacidades o habilidades intelectuales de escolares con problemas de conducta o de rendimiento escolar empleando este instrumento, son muy elevados, dada la escasa validez y fiabilidad de los resultados que proporciona con relación a estos subgrupos de población.

Por otra parte, si nos atenemos a los baremos con que cuentan los profesionales españoles, los distintos grupos de baremación están formados por un número ínfimo de sujetos. El grupo de varones de 6 años de edad es de 26 sujetos, así como el de niñas (en total 52 sujetos). El caso más llamativo es el de varones de 9 años de edad, constituido por 21 sujetos.

En ningún grupo de edad, la totalidad de niños y niñas alcanza el centenar de sujetos. Lo cual significa que el evaluador está comparando los resultados de cualquier escolar con la media obtenida por un grupo de menos de 100 sujetos de la población general. Si tenemos en consideración la posibilidad de comparar por separado niños y niñas, los grupos de comparación no llegan a cincuenta nada más que a los 12 y 13 años (solo varones). Tampoco se conoce la procedencia geográfica y de centros de reclutamiento de estos sujetos. Esto resulta inaceptable desde el punto de vista de la fiabilidad de la medida.

Para agravar esta situación, en el momento actual, finales de 2009, nos encontramos con ***baremos de hace más de 16 años***, obtenidos con unas escalas ***diseñadas hace más de 35 años***.

En conclusión: A pesar del amplio uso de las Escalas Wechsler para la Evaluación de la Inteligencia de Niños, en su versión Revisada (1974) las consideramos de tan escasa validez de constructo y tan deficiente fiabilidad que estimamos debería abandonarse su empleo tanto con fines periciales, como para elaborar hipótesis explicativas de los casos de bajo rendimiento escolar o déficits de adaptación social.

Finalidad	**Valorar la inteligencia de 6 a 16 años**
Validez de contenido	**Moderada**
Validez de constructo	**Muy escasa**
Fiabilidad	**Muy escasa**
Muestra de Baremación	**Insuficiente, no identificada**
Facilidad de Aplicación	**Media**

Nota: como observación generalizada en los centros de evaluación psicopedagógica, el WISC-R no identifica adecuadamente a los escolares con nivel intelectual de tipo "border-line" (nivel intelectual significativamente por debajo de la media)

WISC-IV

Escalas de Inteligencia para Niños de Wechsler, versión IV

(Psychological Corporation, 2002; Madrid: TEA, 2003)

Sobre la Validez de Constructo y Contenido

Desde hace bastantes años, al publicarse las *Escalas Wechsler-Bellevue de Inteligencia para Adultos* (1939) dio comienzo una auténtica saga de escalas de inteligencia, de amplia difusión y empleo en casi todo el mundo: WISC (1949), WIPPSI (1967), WISC-R (1974), WISC-III (1991) y WISC-IV. Las Escalas Wechsler se han ido actualizando cada 10 años aproximadamente, con la supuesta finalidad de compensar el efecto Flynn; esto es, la mejora progresiva de la inteligencia en la población, lo que deja obsoletos los baremos cada diez años. Pese a esta difusión, las escalas de Wechsler adolecen de graves errores de constructo y de fiabilidad, cuyo análisis detallado es el objetivo de este documento, siendo, no obstante, necesario tener en cuenta el contexto sociocultural en el que se desarrollaron. En los años 30, no se disponía de la gran cantidad de investigaciones sobre la inteligencia que hoy tenemos; ni había modelos conceptuales que se hubieran puesto a prueba empírica de manera adecuada, válida y fiable.

En este contexto, se debatía sobre la inteligencia como una aptitud o capacidad general (el factor "g" cuya existencia defendió en un principio Spearman, 1927, y posteriormente Thurstone, 1938 y Carroll, 1993)[7] y la inteligencia como un constructo constituido por diversos factores del mismo orden; es decir, de valor equivalente entre sí, cuya suma era la inteligencia total del individuo.

En el momento actual y desde los años 90, resulta relativamente sencillo criticar estas escalas. Las diversas investigaciones del siglo XX pusieron claramente de manifiesto que los modelos factoriales de la inteligencia estaban equivocados y que "*Spearman tenía razón*": la inteligencia es una capacidad unidimensional con manifestaciones o aplicaciones en todos los ámbitos adaptativos del ser humano: social, personal, escolar, laboral,… El mismo Thurstone, que comenzó defendiendo una estructura factorial de la inteligencia (prueba de ello es su test de Aptitudes Mentales Primarias, P.M.A.) terminó admitiendo un único factor "g" de inteligencia (Paíno Quesada, Susana, 2009; Tema 4. Evaluación de la inteligencia, pág. 3, en "Aspectos Evolutivos y Educativos de la Deficiencia Mental". Universidad de Huelva). Los tres autores aportan evidencias empíricas que sugieren que "*la inteligencia es una capacidad mental general que engloba habilidades diferentes*" (Alonso Tapia, 2002). Por supuesto, que otros autores, Catell (1963), Horn y Catell, (1966), Gardner (1983, 1993, 1998), Sternberg (1988), Das, Naglieri y Kirby (1994) y Greenspan y Love (1997), entre otros, han propuesto un modelo del constructo multifactorial (citados en AAMR, 2002). *Sin embargo, la mayoría de estos modelos de inteligencias múltiples carecen por el momento de base empírica y de validación psicométrica, por lo que actualmente persiste el consenso de que la capacidad intelectual está*

[7] Citados por la AAMR (American Association on Mental Retardation, 2002) y Rodríguez y De Pablo, 2004)

conceptualizada y representada de la mejor manera por un factor común de inteligencia general (Aymamí, Mª. N., 6ª edición 2006; en Vallejo Ruiloba, J., "Introducción a la Psicopatología y la Psiquiatría", páginas 334 y 335. Barcelona: Masson)

Medir la inteligencia es un procedimiento mucho más seguro ahora que conocemos que el tiempo de ejecución de una tarea intelectual no es un indicador de más o menos inteligencia por una diferencia en segundos.

Los diseñadores de tests han aprendido a mejorar la fiabilidad de las medidas y a seleccionar los elementos que integran las diferentes escalas de manera que el sujeto no llegue a los más difíciles (donde se inicia la discriminación entre varios) con efectos de fatiga por la longitud de las mismas. Por ello, aún reconociendo el esfuerzo intelectual que supuso la elaboración de estas escalas por parte de David Wechsler, consideramos que seguir empleando estos instrumentos constituye un error conceptual y metodológico de los profesionales, tanto clínicos como educativos, que perjudica la evaluación, con el consiguiente riesgo de error de diagnóstico. Así pues, si equivocamos el juicio diagnóstico, aumenta la probabilidad de negar las ayudas técnicas a que tiene derecho legal y moral el niño o adolescente (en algunas Comunidades autónomas, como la de Madrid, los dictámenes de los Equipos de Orientación requieren un mínimo puntaje de C.I. para asignar a un escolar la categoría de Alumno con Necesidades Educativas Especiales). Además, si la educación requiere del profesorado *la adecuación de los métodos y ritmos a las características aptitudinales y de personalidad del educando*… (Ley de Educación), un error en la identificación del nivel de capacidad intelectual de un alumno, requisito actitudinal fundamental en todos los aprendizajes, conllevaría, para el Técnico que lleva a cabo la evaluación con un instrumento técnicamente inadecuado, la responsabilidad de impedir al alumnado que su profesorado llevase a cabo las adaptaciones educativas a que la ley le da derecho.

Habida cuenta de que el WISC-IV es la versión más actual de las Escalas de Wechsler, publicado inicialmente en el año 2003, nos centraremos en su análisis dando por descontado que algunos aspectos críticos del mismo son comunes a las versiones anteriores.

En primer lugar, destacamos que el autor estableció (hace ahora 70 años) como definición del constructo "inteligencia" la siguiente: *"el conjunto total de recursos de un individuo para adaptarse al medio"*

En función de esta definición del autor, la inteligencia sería el conjunto de todas las habilidades de que dispone una persona para vivir. Claro que, en tal caso, podría haber incluido como medida de la inteligencia otras muchas cualidades y destrezas: sensoriales, cognitivas y motrices. Obsérvese pues, cómo este modelo constituye una adición de inteligencia lógica, atención y memoria de conocimientos; encontrándose desde su propia definición, en oposición radical al acuerdo generalizado sobre el constructo inteligencia que la relaciona estrictamente con el empleo de procesos cognitivos.

En la versión actual, las pruebas que integran el WISC-IV son las siguientes:

A diferencia de las versiones anteriores en las que este instrumento agrupaba las diversas pruebas en tres escalas: una verbal, otra manipulativa y una total, en la versión actual se agrupan en cuatro áreas:

Comprensión Verbal:

Semejanzas, tradicional prueba de versiones anteriores

Se resuelve con conocimientos culturales y memoria

Vocabulario, tradicional prueba de versiones anteriores

Se resuelve con conocimientos culturales, memoria y fluidez verbal

Comprensión, tradicional prueba de versiones anteriores

Se resuelve con conocimientos culturales, memoria y fluidez verbal

Información, tradicional prueba de versiones anteriores

Se resuelve con conocimientos culturales y memoria

Adivinanzas, prueba novedosa

Se resuelve con conocimientos culturales y memoria

En realidad es una forma inversa de Información.

Razonamiento perceptivo:

Cubos, tradicional prueba de versiones anteriores

Se resuelve con habilidades de organización visual en el plano

No requiere más razonamiento lógico que el comprender que las figuras se construyen con trozos.

Conceptos, prueba novedosa

Se resuelve con razonamiento

Matrices, prueba novedosa

Se resuelve con razonamiento

Figuras incompletas, tradicional prueba de versiones anteriores

Se resuelve con conocimientos culturales, memoria y eficacia atencional

Memoria:

Dígitos, tradicional prueba de versiones anteriores

Se resuelve con escasa atención sostenida y memoria inmediata

Letras y números, prueba novedosa

Se resuelve con escasa atención sostenida y memoria inmediata

Aritmética, tradicional prueba de versiones anteriores

Se resuelve con habilidades de cálculo (memoria) y de razonamiento con elementos numéricos

Procesamiento de la información:

Claves, tradicional prueba de versiones anteriores

Se resuelve con escasa atención sostenida, memoria inmediata y eficacia atencional

Símbolos, prueba novedosa

Se resuelve con escasa atención sostenida y eficacia atencional

Animales, prueba novedosa

Se resuelve con escasa atención sostenida y eficacia atencional

Una vez descritos los componentes del test, podemos proceder al análisis de su validez de contenido y constructo.

Para efectuarlo, nada mejor que tomar la afirmación que se ofrece en la página 59 del manual del test:

*… **aunque los constructores de tests son responsables de aportar pruebas empíricas iniciales sobre la validez,… es el usuario quien debe evaluar si estos datos apoyan el uso que se desea realizar del test para determinado fin…"***

Pues bien, a pesar de que el manual del test nos ofrece un conjunto de análisis factoriales con los cuales se pretende establecer la validez de constructo del mismo (página 64 y siguientes), y suponiendo que es a estos datos a los que se refiere en dicho párrafo como "pruebas empíricas sobre la validez…" no podemos por menos que considerar que las Escalas de Inteligencia de Wechsler versión IV, carecen de validez de constructo para medir la inteligencia tal y como la entienden la mayoría de autores e investigadores en este campo.

Con excepción hecha de Kauffman (K-BIT, K-ABC) y otros escasos autores que siguen el modelo propuesto por Wechsler (por ejemplo Carlos Yuste y Galve Manzano con el BADYG-R y otros), ningún autor de relevancia en este campo, considera que pruebas que se puedan resolver sin el recurso de procesos cognitivos que impliquen razonamiento (análisis-comparación y síntesis) podrían sumarse a otras pruebas que requieran de estos procesos para su resolución satisfactoria, de tal modo que el resultado final sea una medida del constructo "inteligencia". En tal sentido, el WISC-IV sólo puede aspirar a medir el *nivel-de-inteligencia-según-Wechsler de un niño o adolescente*.

Véase cómo, conscientes los autores y adaptadores del WISC-IV de esta situación, en el apartado correspondiente a estudiar su Validez Concurrente, es decir, a estudiar el grado de correlación de esta prueba con otras pruebas de contenidos diferentes para evaluar el mismo constructo, no aparece ningún estudio que no sea con las mismas escalas: WISC-III, WAIS-III, WASI (Wechsler Abbreviated Scale if Intelligence)

Ciertamente, se ha correlacionado el WISC-IV con otras pruebas, como el WIAT-II (Test de Rendimiento), pero esta correlación no corresponde a validez concurrente con otra prueba de Inteligencia. Lo mismo se

puede decir de la correlación con la CMS (escala de memoria) o de la correlación con el BarOn EQ (de inteligencia emocional). Sería muy deseable que los autores hubieran estudiado la correlación entre los índices del WISC-IV y el test de Raven (claro exponente de evaluación de la inteligencia), del test de Sternberg, o de otros tantos tests de evaluación de la inteligencia que carecen de pruebas que se pueden resolver exclusivamente con conocimientos culturales, de memoria a corto o largo plazo, de habilidades motrices o de organización perceptiva.

Resulta de todo punto lógico que no se mencionen tales estudios –si se han realizado- ya que resultaría muy difícil explicar la escasa correlación entre ambos instrumentos, cuyo fundamento no sería otro que el de su origen: las escalas de inteligencia de Wechsler sólo sirven para evaluar lo que Wechsler llamó inteligencia:

"*Lo que medimos con los tests de inteligencia no es lo que aparentemente pretende medir el test, la información del sujeto, su percepción espacial o su capacidad de razonar. Lo que miden los tests de inteligencia -lo que esperamos y deseamos que midan- es algo mucho más importante: la capacidad del sujeto de comprender el mundo que le rodea y los recursos que posee para enfrentarse con sus exigencias y desafíos*".

Algo que conviene tener presente en todo momento:

Si lo que deseaba evaluar D. Wechsler era…*"la capacidad del sujeto de comprender el mundo que le rodea y los recursos que posee para enfrentarse con sus exigencias y desafíos*", entonces le faltan muchas pruebas a las Escalas Wechsler… Pero, en todo caso, no es con estas escalas con las que se puede medir de manera válida y fiable la capacidad de razonamiento lógico de los niños y adolescentes.

Compruébese cómo la estructura de estas Escalas indica ya por sí misma las aptitudes pretendidamente a medir:

1. Comprensión Verbal
2. Razonamiento Perceptivo
3. Memoria
4. Velocidad de Procesamiento

A la vista de lo cual, podemos admitir

a. la validez de contenido de la variable denominada "Comprensión verbal"

b. la validez de contenido de la variable denominada "Razonamiento Perceptivo", ya que, de hecho, es la que más razonamiento requiere para la resolución de pruebas tales como "conceptos" y "matrices" (innovación necesaria en estas escalas)

c. la validez de contenido de la variable denominada "Memoria", puesto que Dígitos y Letras y Números no requieren de otros recursos para resolverlas

Sin embargo, en cuanto a la variable denominada "Velocidad de Procesamiento", resulta inaceptable admitir su validez de contenido. Incomprensible, además, que no se haya levantado ninguna voz crítica con esta variable cuando se presentó por primera vez en la versión anterior de estas Escalas (WIS-III, no publicada en España).

La Velocidad de Procesamiento es un concepto tomado de la neuropsicología, en base a los planteamientos de la Teoría del Procesamiento de la Información (Lindsey y Norman, 1977). A pesar de que esta teoría ya ha sido suficientemente falsada como modelo explicativo del funcionamiento cerebral (sustituida por el modelo conexionista), lo cierto es que nuestras estructuras neuronales, bien en forma de procesamiento en serie (más lento y propuesto por la teoría de procesamiento de la información) o bien en forma de procesamiento en paralelo (más rápido y ajustado a la realidad, propuesto por el modelo conexionista), "procesan información".

Es evidente que los receptores sensoriales "traducen la información externa" a impulsos nerviosos, que transmiten a las áreas corticales (pasando por el tálamo); que de allí pasan a otras áreas y que, finalmente, se produce el fenómeno "perceptivo". Todo esto se realiza en un tiempo, ya que no es "instantáneo". Pero tras la percepción se inicia la fase de "ejecución de una tarea": bien sea ésta marcar con un aspa, subrayar, colorear o pulsar un botón,…

Las tareas propuestas en las Escalas de Wechsler exigen que el sujeto perciba visualmente los estímulos, lo que le lleva un tiempo de búsqueda motriz (Claves, Símbolos y Animales), un procesamiento visual que precederá a un procesamiento cognitivo (¿qué tengo que hacer?, al que le sigue una ejecución motriz (infinitamente más lenta que el procesamiento neuronal) ¿Qué cerebro con formación universitaria puede admitir que el tiempo que tarda un escolar en realizar cada una de estas pruebas puede constituir una medida de la "velocidad de procesamiento"?

El tiempo que registra el evaluador es la suma de los tiempos de percepción visual + procesamiento cognitivo + procesamiento de ejecución (toma de decisión) + ejecución motriz.

¿Qué significado se puede atribuir a los escolares de temperamento impulsivo en esta prueba?¿Algo diferente de los escolares reflexivos? ¿Acaso los investigadores ignoran los trabajos publicados por Tallal y Piercy, 1975, Tallal y cols., 1975, 1976, 1985; o los de Morrison y cols. en Dartmounth, 1977, sobre evaluación de la velocidad de procesamiento en niños con dislexia?

En resumen, a pesar de los juegos matemáticos que constituyen los estudios factoriales aportados por los autores del original y de la adaptación española de las Escalas Wechsler en esta versión IV, no está acreditada en modo alguno la validez de constructo de esta prueba. No se deben olvidar nunca las enseñanzas de los profesores de métodos matemáticos: el matemático proporciona con los datos que le facilita el científico unos resultados determinados, pero es el científico quien debe dar significado a los datos, y no al contrario.

Todo test debe construirse de acuerdo a un modelo conceptual bien explicitado, sus escalas deben poseer validez de contenido (expresar inequívoca y unívocamente diversas manifestaciones de la magnitud que se desea medir) y, después, validar matemáticamente dichas escalas. En el caso de las Escalas de Wechsler nunca se la cumplido este criterio y la versión IV no es diferente a las demás. Así pues, desde un punto de vista meramente conceptual, resulta inaceptable admitir este instrumento como un test para evaluar la inteligencia de los sujetos, a menos, claro está, que no importe tener una medida de algo que no clarifica nada, ya que mezcla aditivamente diversos tipos de destrezas y capacidades.

Si, por el contrario, el método de valoración de las distintas capacidades-habilidades fuera el de un Perfil de Habilidades, mediante el cual, las distintas puntuaciones no se suman para obtener C.I. alguno, sino que se analizan y valoran por separado; en tal caso, podría resultar útil para comprender cada situación personal y diseñar Programas de Intervención (Curso Máster de Asesoramiento, Orientación e Intervención Educativa). Sin embargo, la propuesta de D. Wechsler: la suma algebraica de todas las puntuaciones obtenidas en cada subprueba para obtener una puntuación global, lleva a situaciones frecuentemente absurdas y poco o nada operativas. Supóngase, por ejemplo, en la escala de Razonamiento Perceptivo una elevada puntuación en Cubos y en Figuras Incompletas, pero un déficit significativo en Conceptos y Matrices. La suma de todas las puntuaciones indicaría que el sujeto tiene un "nivel medio", ignorando el importante déficit de razonamiento, que podría explicar algunas de sus dificultades escolares o personales, mucho mejor que un Déficit en Organización Perceptiva. Por cierto, ¿se ha preguntado usted alguna vez qué relación tiene la construcción de cubos y la percepción de detalles en figuras, con los aprendizajes curriculares de matemáticas, lengua, sociales, música, idiomas,…?

Necesariamente debemos concluir que el empleo de estas escalas en la forma que proponen sus autores resulta, metodológicamente, inadecuado, no resultando útiles nada más que para detectar sujetos con déficits importantes o bien con destacadas destrezas en múltiples áreas.

Sobre la Fiabilidad

Ausente la acreditación de la Validez de Constructo de los diferentes Índices de estas Escalas, resultaría innecesario proceder a valorar su fiabilidad. No obstante, por su interés consideramos conveniente comentar los siguientes aspectos que, a nuestro juicio, constituyen importantes errores metodológicos de la prueba que afectan gravemente a su fiabilidad:

1. Por una parte, las pruebas que permiten valorar las respuestas a los diversos elementos que las componen en forma de puntajes 0, 1 y 2, afectan seriamente al significado que debe darse a las puntuaciones, ya que un mismo número total de la prueba, por ejemplo 18 puede obtenerse con 18 elementos valorados como "1" o como 9 elementos valorados como "2", con todo tipo de posibilidades intermedias. La puntuación 18 no representa un mismo nivel de habilidad para unos que para otros, luego esa forma de medir "no es muy fiable". Esto afecta a Semejanzas, Dígitos, Comprensión y Vocabulario.

2. Por otra parte, las pruebas que mejoran la puntuación por la rapidez en su ejecución, proporcionan una idea equivocada al evaluador, ya que, por una parte, los sujetos de ejecución rápida las terminan antes y los de ejecución lenta (más reflexivos) tardan algo más. Obviamente, no se pueden controlar los estados de motivación, ansiedad de ejecución, temperamento, hábitos de respuesta, de los sujetos y asignarles mayor capacidad intelectual porque son unos 4-5 segundos más rápidos que el grupo normativo. Aún más, la experiencia en los pasados años con las versiones anteriores de este instrumento ha permitido comprobar que cada vez que un escolar pasaba estas pruebas obtenía mayores puntuaciones, pareciendo que estaba mejorando su inteligencia. La realidad es que el escolar realizaba más deprisa las pruebas con las que ya estaba familiarizado, pero no necesariamente respondía a más elementos. Esto afecta lógicamente a Cubos, Claves y Animales (anteriormente afectaba a Historias)

La existencia de pruebas para cuya ejecución satisfactoria hace falta emplear más de una habilidad, constituye un riesgo para la fiabilidad de sus resultados. Por ejemplo, en la prueba de Cubos se requiere destreza motriz para manejar de manera adecuada los cubos, además de habilidad perceptiva. Un buen resultado permite asegurar que el sujeto domina ambas habilidades, pero un resultado deficitario no se puede asignar unívocamente a falta de habilidad de organización perceptiva, ya que podría explicarse por su torpeza motriz o su empeño en mantener los cubos en una posición determinada.

3. Finalmente, la longitud de algunas pruebas puede afectar a la fiabilidad de las medidas ya que cuando los sujetos llegan a los elementos de mayor dificultad, que le requieren mayores esfuerzos de atención y de razonamiento, se pueden sentir fatigados y, no siendo conscientes de la importancia del procedimiento evaluador, dar respuestas aleatorias o renunciar al esfuerzo que requiere buscar la respuesta adecuada. Este efecto se multiplica cuando se acumulan las pruebas, no siendo suficientes breves períodos de descanso.

Con respecto a la fiabilidad test-retest, la Tabla 5.5 de la página 52 muestra dos resultados: el primero es que se cumple la predicción que realizamos en el apartado 2 anterior siendo mayores las puntuaciones obtenidas en el retest que en el test. El segundo es que los coeficientes de correlación son aceptables en todos los casos. Finalmente, con relación a los baremos españoles, hemos de destacar que la cifra global de sujetos evaluados, 1.485 de entre 6 años y 16 años y 11 meses, resulta muy escasa al dividir la muestra en grupos de tipificación de 4 en 4 meses. Si bien en ninguna parte del manual hemos encontrado la razón por la que estas Escalas proporcionan baremos diferentes cada cuatro meses, el hecho real convierte los grupos de baremación en 135 sujetos por cada año de edad y 135/4 = 34 sujetos por grupo de baremación. Bien, este dato constituye el argumento final para considerar el WISC-IV como un instrumento de escasa fiabilidad. El evaluador, cuando establece la comparación de los resultados de un niño o adolescente en estas Escalas con la Media del grupo de edad, lo está haciendo con relación a un grupo de referencia de 34 sujetos. Esto significa que el grupo normativo no llega a tener sujetos para todo el rango de puntuaciones de cada escala, lo cual hubiera podido comprobarse si el equipo de adaptación del WISC-IV hubiera incluido

en el Manual las curvas de distribución de las puntuaciones en cada escala o de los índices generales, en los grupos de baremación.

En resumen, la nueva versión de las Escalas de Wechsler presenta similares errores conceptuales y metodológicos que las versiones anteriores. En nuestra opinión, no se ha acreditado la Validez de Constructo de la prueba, entendida como una medida de la inteligencia, en el sentido universal y popularmente aceptado de que ésta es una cualidad específicamente humana, relacionada con la capacidad de pensar para encontrar soluciones a situaciones novedosas (ya que para situaciones habituales se cuenta con los aprendizajes/hábitos). Especialmente discutible es la validez de constructo del índice Velocidad de Procesamiento, que de ninguna manera puede considerarse como tal dado el procedimiento que se emplea para obtenerlo (sumando tiempo de procesos intelectuales y tiempo de ejecución de la tarea). Por otra parte, al margen del contenido de las escalas, discutido en el párrafo anterior, su fiabilidad es muy escasa, tanto por la metodología que se emplea para medir las diversas habilidades, como por la escasez de los baremos proporcionados.

En conclusión: aunque reconocemos que las escalas de Wechsler han sido durante los pasados años el instrumento más frecuentemente usado por los psicólogos educativos, estimamos que sus fundamentos conceptuales han quedado obsoletos en función de los avances en el conocimiento sobre las funciones intelectuales y que sus graves déficits de contenido y de fiabilidad, incluyendo sus escasos baremos, la hacen inadecuada para su empleo en procesos periciales, en el análisis de las dificultades de aprendizaje, así como en el estudio de casos de bajo rendimiento o fracaso escolar.

Curiosidad: el autor de estas escalas, falleció en 1981 a los 85 años de edad, por lo cual no cabe atribuirle ninguna de las deficiencias que pudiéramos encontrar en las versiones WISC-III y WISC-IV. Aunque podamos discrepar sobre su concepto de inteligencia y su modo de medirla, la realidad es que su trabajo fue de gran valor y relevancia, especialmente en el campo de la detección de lesiones cerebrales en soldados. Muy probablemente el mayor valor de las Escalas Wechsler no estuvo nunca en el campo de la educación, sino en el de la neuropsicología. Sus apreciaciones sobre las zonas dañadas a través de las diferentes subpruebas deficitarias pudo ser el origen del valor que se atribuyeron a sus escalas en el campo de la educación. Sin embargo, el desarrollo de instrumentos tecnológicos mucho más fiables permitió ir descartando el empleo del WISC y el WAIS en el campo de la neurofisiología.

Lamentablemente, debido al importante negocio que constituyó la comercialización de estas Escalas para la Psychological Corporation, ésta no quiso renunciar a continuar ofreciendo al mercado de la educación y la psiquiatría unas escalas tan bien conocidas. Por ello, tras el fallecimiento de su autor, se ocupó de encargar el diseño de unas nuevas escalas, estructuralmente muy diferentes de las diseñadas originalmente por D. Wechsler, pero conservando su denominación original y cambiando solamente el número de su versión.

Esto ha ocasionado que en el siglo XXI se siga comercializando un instrumento, a nuestro juicio, conceptualmente obsoleto, mal diseñado estructuralmente y muy deficientemente validado y baremado. Los perjudicados con su empleo son, necesariamente los escolares a quienes se evalúa y valora de manera sistemática y, en ocasiones, de manera exclusiva con estas Escalas. Subsiguientemente, quienes se ven perjudicados en su prestigio profesional son aquellos que confían en los grupos editores y en quienes, con algún prestigio académico o profesional, lejos de destacar sus errores e insuficiencias, solamente mencionan algunas de sus posibles ventajas.

Finalidad	**Valorar la inteligencia de 6 a 16 años**
Validez de contenido	**Moderada**
Validez de constructo	**Muy escasa**
Fiabilidad	**Muy escasa**
Baremos	**Insuficientes**
Facilidad de Aplicación	**Media**

NOTA IMPORTANTE

Los comentarios, análisis y valoraciones sobre los tests anteriormente citados se han llevado a cabo mediante el empleo de los manuales técnicos en la versión de la que han dispuesto los distintos consultores. Lamentamos la posible existencia de errores o inexactitudes si en versiones posteriores de estos manuales se han llevado a cabo modificaciones o anexos no citados.

Por otra parte, cualquier error o insuficiencia detectada en este libro con posterioridad a su edición será indicada en el sitio web www.preocupados.es como rectificación o aclaración, en la página correspondiente al test en cuestión.

En el mismo sitio web se irán añadiendo progresivamente valoraciones de otros tests, de tal modo que los profesionales interesados puedan descargar e imprimir en formato pdf añadiéndolas a este libro.

SOBRE LOS AUTORES...

E. Manuel García Pérez y Ángela Magaz Lago son ambos Psicólogos Especialistas en Psicología Clínica, mediante titulación otorgada por el Ministerio de Educación en atención a su curriculum académico y profesional; así como, expertos en Psicología Educativa. Dirigen un equipo de profesionales que ha aumentado progresivamente de cuatro a más de veinte miembros, en el momento actual.

Ambos son miembros, en algún caso de honor, de diversas Asociaciones Profesionales nacionales e internacionales: Asociación Española de Terapia del Comportamiento (AETCO), Asociación Peruana de Análisis y Modificación del Comportamiento, Centro Peruano de Investigaciones Psicológicas y Modificación del Comportamiento, Colegio Oficial de Psicólogos del Perú, Academia Peruana de Psicología, Asociación Europea de Evaluación Psicológica (EAPA), de la International Society of Quality of Life, Asociación Española de Psicología Conductual, European Educational Research Association, Sociedad Española para el estudio de la ansiedad y el estrés, Asociación de Psiquiatría Infanto-Juvenil, Asociación Iberoamericana de Diagnóstico y Evaluación Psicológica e International Association of Applied Psychology.

Directores y Profesores del Máster en Psicología, Especialidades Psicología de la Educación, Psicología Clínica y de la Salud, Atención Temprana y Psicopedagogía Clínica, de la Universidad Internacional "Menéndez Pelayo", con promociones en Madrid, Bilbao, Perú y Chile. Miembros de la plantilla docente de otras Universidades Internacionales, como profesores visitantes, tutores de prácticum u honoríficos: Universidad de Buenos Aires, Universidad de La Plata y Universidad del Aconcagua (Argentina), Benemérita Universidad de Puebla (México), Universidad Federico Villarreal, Univ. Nacional de Educación, Universidad Ricardo Palma (Perú), Universidad Autónoma de Madrid y Universidad de Alcalá de Henares.

Profesores invitados en diversos cursos de formación para posgraduados y de actualización científico-profesional, en Perú (Universidad César Vallejo, Inca Garcilaso de la Vega, UNIFé, Universidad San Martín de Porres), en Argentina (Universidad de Buenos Aires, Universidad Nacional de Rosario), en España (Universidad de Valladolid, Universidad de Badajoz, Universidad de Alcalá de Henares, Universidad de Navarra, Universidad del País Vasco, Universidad Internacional Menéndez Pelayo), Colegios Públicos, Privados y Centros de Formación del Profesorado de diversas Comunidades Autónomas (Galicia, Asturias, País Vasco, La Rioja, Castilla-León, Castilla La Mancha, Andalucía, Extremadura, Comunidad Valenciana y Madrid)

Consultores de diversas Administraciones en el País Vasco, Cantabria y Comunidad Valenciana.

Han desarrollado una amplia labor investigadora en el campo de la psicología de la salud y de la educación, con la publicación de numerosas obras de naturaleza diferente: libros, tests psicométricos, instrumentos de evaluación conductual, programas de intervención psicoeducativa y protocolos de evaluación e intervención.

Directores del Grupo **ALBOR-COHS**, los lectores interesados pueden acceder a más información sobre sus obras y actividades de consultoría, investigación y formación en la página www.gac.com.es

BIBLIOGRAFÍA

Anastasi, A. (1968, 78). *Tests Psicológicos*. Madrid: Aguilar.

Anastasi, A. (1987). What test users should know about the interpretation os test scores. *Keynote address at Joint Committee on Testing Practices Second Test Publishers Conference*, Rockville, Maryland. (Citado de Fremer, 1996).

Bartram, D., Lindley, P.A.y Marshall, L. (1992 b). *Update to the Review of Psychometric Tests for Assessment in Vocational Training*. Leicester: BPS Books.

Evers, A. (1996). Regulations concerning test qualifications and test use in The Netherlands. *European Journal of Psychological Assessment, 12,* 153-159.

Fernández Ballesteros, R. (2005). *Introducción a la Evaluación Psicológica*. Madrid: Pirámide.

García Pérez, E.M. y Magaz, A. (2005). *Escalas Magallanes de Detección de Déficit de Atención: EMA-DDA*. Bilbao: COHS. Consultores en CC.HH. s.l.

García Pérez, E.M. y Magaz, A. (2005). *Escalas Magallanes de Identificación de Déficit de Atención: ESMIDAs*. Bilbao: COHS. Consultores en CC.HH. s.l.

García Pérez, E.M. y otros (2005). *Escalas Magallanes de Inteligencia para Niños: EMIN-6*. Bilbao: COHS. Consultores en CC.HH. s.l.

García Pérez, E.M. y Magaz, A. (2000). *Escalas Magallanes de Hábitos Asertivos: EMHAS*. Bilbao: COHS. Consultores en CC.HH. s.l.

García Pérez, E.M. y Magaz, A. (1998). *Escalas Magallanes de Adaptación: EMA*. Bilbao: COHS. Consultores en CC.HH. s.l.

Hambleton, R. K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Coor.), *Psicometría*. Madrid: Universitas.

Huff, Darrell (1954, 1982). *How to lie with statistics*. New York: W.W. Norton & Company, Inc. (reedición 1993; ISBN: 0-393-31072-8

Martínez Arias, R. (1995). *Psicometría: teoría de los tests psicológicos y educativos*. Madrid: Síntesis.

Muñiz, J. y Hambleton, R. K. (1996). Directrices para la traducción y adaptación de los tests. Papeles del Psicólogo, 66, 63-70.

Tuleja, Tad (1992). *Verdades de Mentira*. Colombia: Editorial Voluntad

Toro T., J., Cervera, M. y Urío, C. (2002). *Escalas Magallanes de Lectura y Escritura: TALE-2000*. Bilbao: COHS. Consultores en CC.HH. s.l.