Formación Continuada a Distancia
Consejo General de la Psicología de España

1

Vigésimo segunda Edición octubre-diciembre 2013

LOS TESTS PSICOLÓGICOS EN LA PRÁCTICA PROFESIONAL

Paula Elosua
UNIVERSIDAD DEL PAÍS VASCO

FORMACIÓN continuada a distancia

ISSN 1989-3906

Formación Continuada a Distancia
Consejo General de la Psicología de España

Contenido

FORMACIÓN continuada a distancia

Documento base.

Los tests psicológicos en la práctica profesional

1. INTRODUCCIÓN

Desde la aparición del test de Binet-Simon (1905) para la medición de aptitudes intelectuales superiores, la disponibilidad y el uso de los tests se han extendido a todos los ámbitos de actuación de la psicología (educativo, social, jurídico, organizacional, deportivo, clínico, investigador…). Se utilizan como herramientas de apoyo en el quehacer diario y asistencial profesional en la toma de decisiones; decisiones que afectan en mayor o menor grado a personas o a instituciones. Un diagnóstico clínico, una evaluación educativa, el diseño y evaluación de un programa de intervención, la selección de personal o un peritaje judicial entre otros, demandan la recogida de información objetiva que se lleva a cabo por medio de instrumentos métricos diseñados con esa finalidad, los tests. En apenas el siglo de historia que tienen los tests en occidente, los avances en el campo de la investigación psicométrica y psicológica han perfilado un campo marcado por la mejora en los modelos psicométricos, por la innovación en nuevos formatos de ítems/tests, y por una concienciación sobre la importancia de un uso correcto de los tests. Esta última surge como consecuencia de los usos abusivos e incorrectos detectados, y derivados de un conocimiento insuficiente (Anastasi, 1954) de las fortalezas y debilidades de los instrumentos de medida.

Básicamente un test es un *instrumento de medición estandarizado* que sirve para recoger información sobre una muestra de conducta (American Educational Research Association, American Psychological Association, y National Council on Measurement in Education, 1999). Aún siendo conscientes de que los desarrollos técnicos y psicométricos de las últimas décadas en el área de los tests hacen difícil una clasificación exhaustiva de los tipos de tests disponibles (Olea, Abad y Barrada, 2010), sería posible catalogarlos en función de los siguientes criterios: a) *campo de estudio* (personalidad, inteligencia, neuropsicología, intereses vocacionales, rendimiento, competencias, aptitudes, actitudes…), b) *forma de administración* (individual/ colectivo, secuencial/adaptativa, online/offline), c) *soporte* (papel/ ordenador), d) *información requerida* (autoinformes, ejecución, observación), e) *tipo de interpretación* (normativa, criterial, ipsativa), y/o f) *tipo de respuesta* (ejecución típica, ejecución máxima).

El objetivo de los tests es recoger información métrica que podrá ser utilizada junto con el apoyo de otros recursos en un proceso evaluativo. Constituyen el aspecto más conocido y de mayor impacto social relacionado con la investigación psicométrica. Sean conocidos como cuestionarios, exámenes, escalas, instrumentos, autoinformes o medidas, los tests comparten exigencias formales cuyo cumplimiento los facultará para su uso en la práctica profesional. Alcanzar el propósito con el que fueron diseñados depende de que se cumplan dos requisitos: a) que su construcción siga

principios psicométricos que garanticen su calidad técnica (Wilson, 2005), b) que sean utilizados de acuerdo a criterios que permitan salvaguardar éstos. Sólo un uso apropiado garantizará la validez de las interpretaciones.

En este marco definido por los requisitos de calidad psicométrica y el uso correcto de los tests , el objetivo de este curso formativo es ofrecer al profesional un panorama de actualidad que recoja los avances llevados a cabo para la mejora del uso de los tests. Para ello dividimos el curso en 5 apartados en los que a) definimos los componentes de un tests, b) presentamos los proyectos internacionales relacionados con las prácticas para la mejora del uso, c) ofrecemos una introducción de los conceptos psicométricos que influyen directamente en la correcta aplicación/interpretación/elección de los tests (fiabilidad, validación, estandarización), y d) presentemos el cuestionario para la evaluación de tests amparado por la Federación Europea de Asociaciones de Psicólogos (EFPA) que utiliza el Consejo General de la Psicología para la evaluación de los tests publicados en España.

Se trata de un documento reducido en extensión que viene acompañado de dos fichas de aconsejada lectura. La primera se refiere a las directrices relacionadas con el uso de los tests elaborada por la Comisión Internacional de Tests (Ficha 1), y la segunda es el modelo evaluativo de tests de la EFPA (Ficha 2).

2. COMPONENTES DE UN TEST

El resultado obtenido tras la administración de un test es un una representación observada de una variable latente. Este resultado, habitualmente en forma numérica y conocido como puntuación observada o puntuación empírica tiene ciertas propiedades métricas que la caracterizan: a) no es una medida directa del rasgo evaluado, b) no es un valor numérico interpretable en una escala de razón, c) está afectada por errores de medida y, d) su interpretación y valoración están sujetas a los cuatro principios básicos por los que se rige la medición psicométrica: fiabilidad, validez, estandarización y uso.

La inferencia psicológica – el uso de la puntuación obtenida con fines métricos y/o evaluativos- exige la consideración y conjunción de estos *cuatro componentes* en la interpretación de las puntuaciones.

- El componente de *fiabilidad* se relaciona directamente con el error aleatorio de medida. Ninguna medida está libre de error, y la estimación de su efecto sobre la puntuación es imprescindible para su interpretación. Los modelos psicométricos se encargan de la estimación del efecto de los errores aleatorios sobre las puntuaciones. El modelo psicométrico más utilizado en España para la construcción de tests es el basado en la Teoría Clásica de Tests (TCT; ver Abad, Olea, Ponsoda y García, 2011; Elosua, 2011; Muñiz, 2000) según el cual la puntuación observada X es una combinación lineal de la puntuación verdadera (V) y el error aletaorio de medida (E; X=V+E). Coexiste con la teoría clásica de tests, la Teoría de Respuesta al Ítem (TRI; ver Elosua, Hambleton y Muñiz, en prensa; Hambleton y Swaminathan 1985; Lord, 1980). La TRI postula que la probabilidad de respuesta correcta a un ítem es una función de las características del ítem y del nivel de la persona en la variable medida y, estima además el error de medida como una función de ésta.
- El segundo componente a considerar en la interpretación de una puntuación es su *validez*. La validación, se relaciona con el modelo sustantivo, teoría psicológica o base racional utilizada en la construcción del test. Un test es algo más que un procedimiento estandarizado para la obtención de valores escalares. Es un instrumento de medida, y como tal tiene que justificarse. Es necesario demostrar que la puntuación es un indicador del constructo o variable no observada que se desea medir. Independientemente del error de medida que afecta a la puntuación, ésta tiene que quedar justificada a través de los consiguientes estudios de validación.
- El tercer componente de un test hace referencia a su carácter de medida estandarizada. La estandarización asegura que tanto las instrucciones, como la administración, corrección e interpretación se realizan siguiendo pautas de actuación normalizadas.
- Sin embargo, estos tres componentes, fiabilidad, validez y estandarización, carecerían de sentido si el uso del test no respetara los objetivos, la finalidad, el contexto de aplicación o las características definitorias de la población a la que éste va dirigido. Si el uso de los tests no es correcto no puede asegurarse el cumplimiento de sus propiedades psicométricas. Hoy somos más conscientes que nunca de las consecuencias derivadas de un uso incorrecto de los tests (Messick, 1995).

Todo test se construye sobre estas cuatro bases. Los modelos psicométricos se encargarán del análisis formal de las puntuaciones. Los estudios de validación tienen como cometido el análisis sustantivo de las puntuaciones obtenidas. El proceso de estandarización garantizará en la medida de lo posible, que las diferencias encontradas no son debidas a una aplicación/corrección incorrecta del test, y finalmente un uso correcto preservará las propiedades psicométricas del test.

3. HACÍA UNA MEJORA EN EL USO DE LOS TESTS

El uso correcto de los tests es uno de los cuatro pilares sobre los que descansa la interpretación de las puntuaciones. El desarrollo de la psicometría en los últimos años viene marcado por la construcción y estudio de nuevos modelos formales, pero también, por una toma de conciencia respecto a la relevancia social de los tests que no ha existido hasta ahora. La atención prestada a la importancia del correcto uso de los tests en la práctica profesional queda reflejada en los proyectos de mejora en el uso iniciados desde organizaciones nacionales e internacionales.

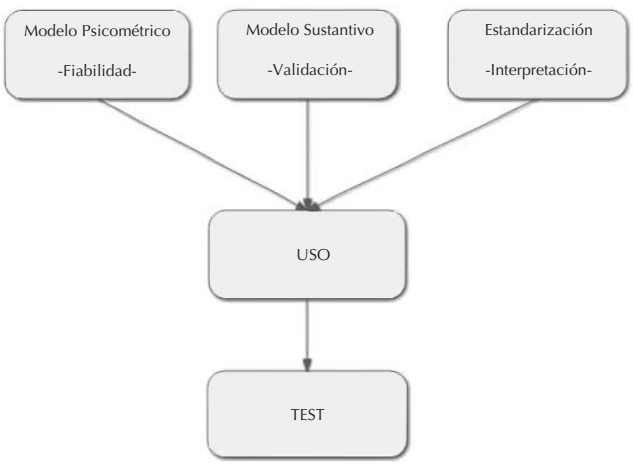
A partir de las iniciativas surgidas en las Comisiones de Tests de la Federación Europea de Asociaciones de Psicólogos (EFPA, *European Federation of Psychologists' Associations*), del Consejo General de Colegios Oficiales de Psicólogos (CGCOP) y de la Comisión Internacional de Tests (ITC, *International Test Commission*) asistimos a un proceso de descripción, revisión y formación en el uso de los tests. Los proyectos amparados por estas organizaciones incluyen, estudios sobre las actitudes de los psicólogos hacia los tests (Evers y col., 2012; Muñiz y Fernández-Hermida, 2010), análisis sobre las condiciones que favorecen una correcta evaluación (Muñiz y Bartram, 2007) o la revisión/actualización de los cuestionarios para la valoración de los tests publicados en Europa (Bartram, 2011; Evers y col., 2013; Muñiz, y col., 2011).

Los proyectos se enmarcan dentro de una estrategia formativa destinada a mejorar el uso de los tests (Muñiz, 2010). Los objetivos fijados dentro de esta estrategia formativa se centran en la formación del profesional, en la diseminación de información sobre la calidad y características de los tests disponibles, y en el desarrollo de directrices.

Las directrices cumplen una importante misión en la difusión de los avances formales y sociales. Son documentos que conjugan rigor y sencillez en textos de fácil lectura y comprensión. Ofrecen normas generales que son importantes en el proceso y en la evaluación del resultado de la construcción/adaptación y uso de los tests. Las directrices sobre construcción/uso de los tests de la AERA, APA y NCME (1999), las directrices para la evaluación en contextos laborales y de organizaciones (ISO, 2011), las directrices sobre evaluación psicológica (Fernández-Ballesteros y col., 2001) o las directrices de la Comisión Internacional de Tests relacionadas con el uso de los tests (**Ficha 1**), la adaptación de tests (Muñiz, Elosua y Hambleton, 2013), o la evaluación por internet (http://www.intestcom.org) son excelentes ejemplos de esta labor. Definen marcos de referencia de inexcusable seguimiento, en los que se recogen los avances metodológicos de mayor impacto, y describen las prácticas más aconsejadas que garantizan un uso correcto de los tests.

Los resultados de las iniciativas llevadas a cabo concluyen que: a) los psicólogos participantes en el último estudio sobre actitudes hacia los tests (Muñiz y Fernández-Hermida, 2010) reconocen que la formación recibida en el grado de Psicología puede no ser suficiente para la correcta utilización de la mayoría de los tests, b) la actitudes de los psicólogos respecto al uso de los tests es positiva (Muñiz y Fernandez-Hermida, 2010), c) el conocimiento psicométrico avanza de tal forma que la distancia entre psicometría teórica y práctica profesional es hoy mayor que nunca (Elosua, 2012; Elosua & Iliescu, 2012) , y d) las directrices que intentan aunar y exponer los avances metodológicos y sociales en teoría de tests son documentos de adhesión que se suscriben pero en muchos casos no se conocen, a pesar de su papel motor sobre la mejora en los usos.

FIGURA 1
COMPONENTES DE UN TEST



4. CONCEPTOS PSICOMÉTRICOS BÁSICOS

Los tests son instrumentos que se rigen por principios psicométricos. El desarrollo de la psicometría como área de conocimiento encargada de la medición en psicología ha estado marcado desde sus orígenes por la distinción entre planos de actuación psicométrica diferentes, que quedan perfectamente reflejados en los contenidos de revistas como *Psychometrika*, o la más cercana al psicólogo aplicado *Papeles del Psicólogo*. Es innegable, la distancia entre los contenidos de ambas, pero el conocimiento psicométrico está presente y nutre a las dos.

El profesional que administra o corrige un test, no

es un psicómetra, no tiene que serlo. No obstante en la medida que utiliza un instrumento que se rige por requisitos técnicos, un uso correcto exige la comprensión/aplicación de los principios básicos constituyentes de los tests: fiabilidad, validez y estandarización.

4.1. *Fiabilidad y Error de Medida*

Fiabilidad. La fiabilidad en términos psicométricos tiene varias acepciones entre las que destacan la que la entiende como estabilidad temporal y aquella que la interpreta como un indicador de consistencia interna. La primera significa que las puntuaciones son constantes en el tiempo (test-retest), mientras que la segunda es un indicador de las relaciones entre los ítems que componen el test. Desde la aparición del artículo de Cronbach (1951), el uso del coeficiente alfa como indicador de consistencia interna de las puntuaciones es una constante en la documentación que acompaña a los tests y a los artículos sobre construcción/adaptación. Su éxito puede justificarse sobre tres puntos; es fácil de estimar, es sencillo de interpretar (Nunnally, 1978) y no es complicado mejorar su valor.

La fiabilidad de una puntuación es un valor situado en el intervalo 0, y 1. A medida que se acerca a 1 aumenta la precisión de la medida. Si bien es deseable que los coeficientes de fiabilidad se aproximen a 1, su valor medio es diferente en función del tipo de test y de su propósito. Para los tests de uso individual relacionados con la medida de aptitudes el valor se aproxima a 0,92. Si el objetivo del test es la estimación grupal, este valor suele ser algo inferior, situándose cercano a 0,85. Las escalas de personalidad presentan coeficientes de fiabilidad en torno a 0,75, los exámenes abiertos tiene fiabilidades de 0,65, y los tests utilizados en la evaluación educativa alcanzan el valor de 0,95. La exigencia sobre la fiabilidad de los tests ha de ser mayor en aquellos contextos en los que las consecuencias derivadas de las puntuaciones son importantes para el evaluando. En este sentido, la exigencia sobre la fiabilidad de un test utilizado en investigación que no tiene repercusiones sobre las personas que lo han respondido será menor que la exigible a aquellos tests sobre los que se basan decisiones individuales.

Error típico de medida. La consistencia no es la única acepción relacionada con la precisión de las medidas; el concepto de error típico de medida ocupa un lugar preeminente en la evaluación individual. El error típico de medida (ETM) cuantifica el error aleatorio en torno a la puntuación verdadera, y en los contextos evaluativos en los que el objetivo final es la interpretación de una puntuación, su relevancia es mayor que el de la consistencia interna; aporta una vía para expresar la incertidumbre con relación a las puntuaciones que no es ofrecida por el coeficiente alfa de Cronbach. Solo a partir del valor del ETM podrían construirse enunciados como: "Con un 95% de probabilidad la puntuación de la persona X se sitúa entre los valores 34 y 48".

La estimación del error típico de medida desde la teoría clásica es tan sencilla como el cálculo del coeficiente alfa de Cronbach; a pesar de ello, la mayoría de los programas informáticos y los manuales publicados no ofrecen información sobre este índice.

$$ETM = DT_X\sqrt{1 - r_{XX'}}$$

Donde, DT_X es la desviación típica de la muestra
$r_{XX'}$ es la fiabilidad del test.

Ejemplo: La puntuación obtenida por María en el WISC-V fue 105 y la de Pedro fue 115. ¿Se puede afirmar que el potencial académico de María es menor que el de Pedro? La fiabilidad del WISC-IV es 0,95 y la desviación típica en la muestra es 14.

a.- El error típico de medida asociado al WISC-IV es: $ETM = 15\sqrt{1 - 0{,}95} = 3{,}35$
b.- El Error típico de medida tiene una distribución normal.
c.- El intervalo de confianza del 95% es ±ETM×1,96: 3,35×1,96=6,57 (aprox. 7)
d.- El intervalo de confianza para cada puntuación es:
María: 105±7 = 98 a 112
Pedro: 115±7= 108 a 121

Los resultados aconsejan cautela en la interpretación absoluta de los resultados.

Consistencia desde un modelo. A pesar del extendido uso del coeficiente alfa como estimador de la fiabilidad, un número cada vez mayor de psicómetras desaconseja su uso, y propone alternativas de estimación y de definición de

la fiabilidad construídas sobre modelos alternativos de medida a la teoría clásica de tests (McDonald, 1999). Las nuevas aproximaciones al estudio de la consistencia de las puntuaciones se basan bien en los modelos de respuesta al ítem bien en modelos factoriales que ahondan en el problema de la homogeneidad con relación al factor medido. Estimar la consistencia desde esta perspectiva factorial ayudaría a eliminar a) el uso incorrecto del coeficiente alfa como indicador de unidimensionalidad, b) los problemas derivados del uso de estadísticos que no cumplen las asunciones del modelo referidas en este caso al carácter continuo de las variables o a la tau-equivalencia de las medidas (Elosua y Zumbo, 2008; Zumbo, Gadermann y Zeisser, 2007), y c) permitirían profundizar en el estudio de la estructura interna del test.

En el estudio de la fiabilidad merecen un apartado destacado los modelos construidos desde la teoría de respuesta al ítem (TRI) en tanto en cuanto suponen un avance importante respecto a la teoría clásica de tests en los procesos de construcción de tests y análisis de ítems (Embretson y Reise, 2000; Lord, 1980; van der Linden y Hambleton, 1997). Los modelos de respuesta al ítem favorecen el estudio de: a) la invarianza tanto de las personas como de los ítems b) la equivalencia entre grupos, y c) la estimación condicional de los errores de medida (función de información). Las ventajas y propiedades formales de la TRI hacen de ella un marco teórico atractivo y efectivo en la resolución de problemas asociados a la medida en psicología; entre los que cabría mencionar la equiparación y comparabilidad de puntuaciones, el análisis del funcionamiento diferencial de los ítems, la construcción de tests adaptativos o la elaboración de informes evaluativos.

4.2. *Validación*

La validación se refiere al proceso de recogida de información que tiene por finalidad justificar las inferencias derivadas de las puntuaciones. Si bien tradicionalmente se han distinguido tres tipos de validez (contenido, criterio y constructo) y son las formas recogidas de manera mayoritaria en los manuales de tests, la teoría actual sobre validez, prefiere hablar de validación y bajo este término común recoge evidencias (pruebas) que justifiquen las inferencias o conclusiones construidas a partir de las puntuaciones obtenidas por un test. En la búsqueda de información por medio de pruebas o evidencias es importante recordar que siempre hay que evaluarlas con referencia al propósito del test.

Desde una perspectiva histórica la definición y evolución del concepto de validez ha quedado reflejada en las sucesivas ediciones de las directrices conjuntas de la APA (ver Elosua, 2003). La concepción tripartita de la validez que diferenciaba entre validez de constructo, validez predictiva y validez de contenido se superó en 1985 aceptándose a partir de esta fecha una concepción unitaria. En la edición de 1999 se postulan cinco fuentes de evidencia en el proceso de validación de una puntuación (contenido, proceso de respuesta, estructura interna, relaciones con otras variables y consecuencias), y se incide en el aspecto práctico de la validez. El giro adoptado implica ligar la validez de las puntuaciones a su uso (perspectiva que se mantiene en la próxima edición).

Las definiciones de validez ofrecidas en los años 30 que diferencian tres tipos de validez, son las más comúnmente adoptadas en los manuales clásicos de psicometría y en los manuales de los tests. Reflejan la idea de que un test es válido para aquello con lo que correlaciona (Guilford, 1946; Kelley, 1927), o equiparan la validez del test al grado en que el test mide lo que pretende medir. La operacionalización de la primera acepción se lleva a cabo por medio de la correlación entre un test y un criterio. La segunda acepción más acorde o cercana a los modelos factoriales, se materializa habitualmente por medio del análisis factorial exploratorio. Ambas técnicas fueron diseñadas en los albores de la psicometría (Spearman, 1904; Thurstone, 1932), han influenciado las primeras definiciones de validez, están integradas en los planes de estudio de las facultades de psicología españolas, y forman parte de los módulos de software para el análisis de datos en ciencias sociales.

La nueva teoría sobre la validez (validación) sitúa el foco de atención sobre la validación de la interpretación propuesta; no se habla ya de validez del test. El objetivo es justificar una interpretación de las puntuaciones basada en razones, argumentos, que se recogen durante el proceso de validación (Kane, 1992, 2006; Sireci, 2007; Zumbo, 2007).

Al igual que ha evolucionado el concepto de validez (validación), los modelos factoriales y de regresión originales han dado paso a metodologías más potentes y explicativas diseñadas para el estudio de las relaciones entre variables observadas y latentes; los modelos de ecuaciones estructurales (SEM, *structual equation modeling)*. Desde la década de los años 70 se ha producido un rápido desarrollo de los modelos teóricos SEM (Bollen, 1989; Millsap y Maydeu-Olivares, 2009) y del software para su estimación (Elosua, 2009; R Core Team, 2012). SEM representa una familia de técnicas estadísticas multivariantes potentes y flexibles entre las que se incluyen los modelos factoriales confirmatorios, que permiten modelar las relaciones entre variables latentes e indicadores, asumiendo en todo caso la presencia

de errores de medida. Las ventajas asociadas al uso de los modelos de ecuaciones estructurales en el proceso de validación incluyen, la estimación de la fiabilidad y el error de medida, la construcción de modelos teóricos explicativos y su contraste simultáneo para diferentes grupos.

4.2.1. Sesgo

Uno de los conceptos que con más fuerza ha emergido en la medición psicológica como consecuencia, entre otros factores, del uso indebido de los tests, es el concepto de sesgo. Es un término con connotaciones políticas, sociales, estadísticas y psicométricas, que comienza a cobrar relevancia en la década de los 20 debido a las controversias y procesos judiciales surgidos en Estados Unidos referidos a la parcialidad de los tests respecto a determinadas minorías étnicas (Cole y Moss, 1989; Jensen, 1980).

Desde un punto de vista estrictamente psicométrico el sesgo es un error sistemático originado por deficiencias en el test o en el modo en que éste es usado, que produce una distorsión en el significado de las puntuaciones y contamina su interpretación. Validez y sesgo son dos caras de una misma moneda. El sesgo siempre supondrá falta de validez, y la falta de validez puede ser el origen del sesgo. Para maximizar una, y consecuentemente minimizar otra, el test ha de incorporar una descripción detallada del uso propuesto y sólo con referencia a él podrán interpretarse las puntuaciones.

Es necesario evaluar la posible *infrarrepresentación* del constructo o la existencia de *varianza irrelevante*. La infrarrepresentación del constructo se refiere al grado en que el test no incluye aspectos relevantes del mismo. Esta carencia, puede corromper el significado de las puntuaciones porque los ítems no son una muestra representativa, no abordan algún proceso psicológico importante, o elicitan un tipo de conducta o componente que no está entre los propósitos del test. La varianza irrelevante del constructo hace referencia al grado en que las puntuaciones del test se ven afectadas por procesos que son extraños a su propósito. Las puntuaciones pueden estar sistemáticamente influenciadas por componentes que no forman parte del constructo.

La inclusión de estos aspectos dentro del proceso de validación supone adoptar un punto de vista multidimensional sobre el origen del sesgo. Se acepta el hecho de que un instrumento de medida puede estar sesgado si mide dimensiones diferentes en dos (o más) grupos y además las distribuciones multidimensionales de estos grupos respecto a ellas difieren (Ackerman, 1992; Camilli y Shepard, 1994; Mellenbergh, 1989). Por ejemplo, en un país en el que coexisten 5 idiomas oficiales (español, catalán, gallego, valenciano y euskera) el idioma familiar o dominante del evaluando puede ser un factor que origine sesgo. Si bien es posible asumir que todos los ciudadanos catalanes entienden el español, cuando el catalán es la lengua materna y además es la lengua de instrucción, el administrar un test en español a un niño catalán podría generar un componente intencionalmente no medido (dominio de la lengua española) que vulneraría las condiciones de aplicabilidad del test y podría ser el origen de un sesgo por razón de idioma. En estas condiciones quedarían comprometidas tanto la interpretación de las puntuaciones como su comparabilidad con respecto a la muestra en la que se construyó el test.

Las implicaciones más directas de estos principios suponen que el proceso de validación ha de incluir una atención cuidadosa a posibles distorsiones en el significado de las puntuaciones causadas por la influencia de variables ajenas a los objetivos propuestos. Si el origen puede situarse en una inadecuada representación del constructo o aspectos tales como el formato de la prueba, condiciones de administración, o nivel del lenguaje utilizado, el proceso continuo de recogida de evidencias para garantizar la validez debería de considerarlas objeto de análisis.

4.2.2. Fuentes de evidencia en los estudios de validación

La recogida de evidencias para prestar una base científica a la interpretación de las puntuaciones en un uso concreto puede provenir de diversas fuentes. La importancia otorgada a cada una de ellas dependerá siempre de los objetivos del test, que determinarán en cada caso el tipo de evidencia más significativa. Los últimos estándares (AERA, APA y NCME, 1999) diferencian entre fuentes relacionadas con el contenido, la estructura interna, las relaciones con otras variables, el proceso de respuesta, y las consecuencias del test, que en ningún modo suponen distintos tipos de validez sino aspectos complementarios. Citaremos las tres primeras (para un análisis más detallado véase Elosua 2003; Zumbo, 2007).

4.2.2.a Evidencias centradas en el contenido

Examinan el grado en que los ítems de un cuestionario son relevantes y representativos del constructo medido. En un

proceso de validación de contenido se analiza la representatividad del ítem, su relevancia, pero también el formato de respuesta, el modo de puntuación/corrección, su adecuación lingüística con respecto a las características de la muestra evaluada, o las instrucciones ofrecidas en el manual (Haynes, Richard & Kubany, 1995).

4.2.2.b Evidencias centrada en la estructura

El estudio de la evidencia centrada en el análisis de la estructura interna evalúa el grado en que las relaciones entre los ítems y los componentes del test conforman el constructo que se quiere medir y sobre el que se basarán las interpretaciones. Generalmente, su objetivo es especificar y demostrar la existencia de una "estructura simple" (Thurstone, 1947) que resulta de la agrupación de ítems en núcleos dimensionalmente homogéneos entre ellos, y dimensionalmente distintos entre sí.

Cuando se analiza la estructura de un test se intenta "dividir" la puntuación total en partes (factores) que con un significado sustantivo intenten explicarla. El procedimiento estadístico más utilizado para ello es el análisis factorial. El análisis factorial engloba un conjunto de técnicas de análisis multivariado cuyo objetivo es resumir la información contenida en un grupo de variables observadas por medio de un número reducido de variables hipotéticas, conocidas normalmente como factores. Existen básicamente dos aproximaciones al estudio factorial, el exploratorio y el confirmatorio. En el primer caso el procedimiento se utiliza desde una aproximación descriptiva y en el segundo la aproximación permite cotejar y/o refutar estadísticamente modelos alternativos.

El estudio y conocimiento de la estructura factorial de un test tiene importantes repercusiones teóricas y aplicadas porque a) la solución factorial está ligada al modelo teórico utilizado en la construcción del test, y b) determina el modo en que ha de puntuase el test y c) fija la manera en que debe estimarse la fiabilidad. Por ejemplo, si los análisis muestran que bajo el test subyace un único factor o existe un factor dominante, quedaría justificada la suma de los valores individuales obtenidos en cada uno de los ítems para obtener un único indicador del constructo medido, es decir, una puntuación total. Sin embargo, si la estructura factorial describe múltiples factores, es a cada uno de ellos a los que habría de asociar una puntuación parcial. En este caso, se deberían de estimar coeficientes de fiabilidad diferenciados para cada una de las escalas parciales (Elosua, 2008) y no estaría justificado el utilizar un único indicador.

4.2.2.c Evidencia centradas en las relaciones con otras variables

Se centra en la búsqueda de pruebas que relacionen la puntuación con algún criterio que se espera prediga el test, o con otros tests que hipotéticamente midan el mismo constructo, constructos relacionados o constructos diferentes (AERA, APA y NMCE, 1999). Los resultados obtenidos de este modo evalúan el grado en que las relaciones hipotetizadas son consistentes con la interpretación propuesta.

Relaciones test-criterio. El objetivo es constatar la relación entre las puntuaciones obtenidas por el test y un criterio. Un criterio es un indicador de un constructo, de un diagnóstico, o de una conducta que el test intenta predecir. La bondad de los estudios centrados en las relaciones test/criterio dependerá de la calidad del criterio seleccionado y de la calidad de la medida que de él se haya tomado. El usuario habrá de valorar la calidad del criterio con referencia al propósito del test. La información reportada en el manual a modo de coeficiente de validez, es la correlación obtenida en una muestra representativa entre las puntuaciones en el test y la medida del criterio. A mayor coeficiente de validez mayor será la relación entre ambas medidas.

Validez predictiva. En determinados ámbitos del uso de test, como en el organizacional, en el que un test es usado para selección de personal, la información sobre la relación entre el test y un criterio ha de ampliarse con información referida a la precisión de la predicción. Para ello se tiene en cuenta el número de errores por falsos positivos, falsos negativos, la razón de selección (porcentaje de candidatos que serán seleccionados) o la tasa base. La cantidad de información aportada por una prueba más allá de la tasa base puede determinarse utilizando las tablas de Taylor-Russell (1939). Las tablas muestran el porcentaje de solicitantes seleccionados que se espera tengan "éxito" en la situación de selección en función del coeficiente de validez, de la tasa base y de la razón de selección. El uso de las tablas supone una definición dicotómica del "éxito" en una situación de selección; se han elaborado enfoques similares para criterios continuos basados en la teoría de la decisión y en la teoría de la utilidad (Raju, Burke y Normand, 1990).

Validez discriminante. Clasificación. En la práctica clínica la información test/criterio tampoco es suficiente, y es necesario aportar información sobre la capacidad discriminatoria del instrumento, es decir, sobre su sensibilidad y especificidad. La sensibilidad se refiere a la proporción de casos diagnosticados como afirmativos, a partir del criterio o regla de decisión. La especificidad es la proporción de casos diagnosticados como negativos, a partir de la regla de decisión (ver Tabla 1).

Habitualmente se valoran la especificad y sensibilidad para las distintas puntuaciones obtenidas por un cuestionario a través de curvas ROC (*Receiver Operating Characteristic*, o Característica Operativa del Receptor (López-Jáuregui y Elosua, 2013). La curva ROC resume, en un único gráfico, sensibilidad y especificidad para todos los puntos de corte en una prueba, es decir, la capacidad predictiva para la detección de casos verdaderamente positivos. El área bajo la curva ROC es una medida de la eficacia predictiva del modelo, independientemente del punto de corte que se establezca con la misma (Swets, 1986; Swets y Pickett, 1982). Mediante la curva ROC empírica se puede comparar visualmente el rendimiento de varias pruebas sobre los mismos casos, o sobre casos diferentes. Cuanto más alejada se encuentre la curva de la diagonal, mayor será la capacidad predictiva de la prueba. Asimismo, se podrá concluir si una prueba detecta más que por puro azar, simplemente comprobando que la curva se separa de la diagonal (Figura 2).

Evidencia convergente y discriminante. Una de las formas más utilizadas en la recogida de información durante el proceso de validación de un instrumento consiste en evaluar su relación (correlación estadística) con medidas similares obtenidas por medio de otros tests (validación convergente), o con medidas diferentes con las que teóricamente el test no presenta relación (validación discriminante). Las correlaciones convergentes habrían de ser altas, y las correlaciones discriminantes bajas. El estudio de un conjunto de correlaciones convergentes y discriminantes se analiza en el marco de los modelos de ecuaciones estructurales a través del análisis de lo que se conoce como matriz multirasgo-multimétodo.

4.3. *Estandarización*

La puntuación empírica obtenida en un test no puede interpretarse en términos absolutos en tanto en cuanto las escalas a partir de las cuales se obtienen las puntuaciones no son escalas de razón. La interpretación de la mayor parte de los tests psicológicos se lleva a cabo con referencia a la norma (test normativos) o con referencia al criterio (test criteriales). Si bien existen también tests ipsativos – todavía con poca tradición - en los que la interpretación es intrasujeto.

TABLA 1
SENSIBILIDD Y ESPECIFIDAD

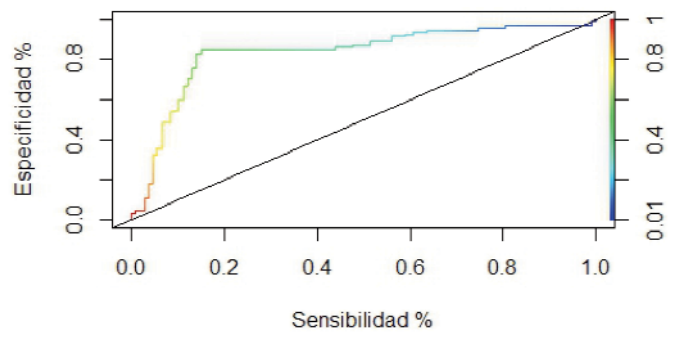
ESCALA	CRITERIO (estado real)	
	Positivo **SI** es un caso clínico	**Negativo** **NO** es un caso clínico
Positivo **SI** puede ser caso clínico	Verdadero positivo (VP)	Falso positivo (FP) -Falsas alarmas-
Negativo Puede **NO** ser caso clínico	Falso negativo (FP) -omisiones	Verdadero negativo (VN)

$$Sensibilidad = \frac{VP}{VP + FN} \times 100$$

$$Especificidad = \frac{VN}{VN + FP} \times 100$$

VP= Verdaderos Positivos
FP = Falsos Positivos
VN= Verdaderos Negativos
FN = Falsos Negativos

FIGURA 2
CURVA ROC



4.3.1. Interpretación normativa

La muestra de estandarización o normalización es aquella muestra representativa de la población a la que el test va dirigido y que ha sido utilizada en el proceso de construcción del test. El manual del test ha de incluir información relevante sobre ella y sobre su distribución en función de las variables relacionadas con el constructo o muestra de conducta evaluada. El contenido de esa información es crucial para la valoración de la aplicabilidad del test. Si la muestra de normalización no es equiparable a las personas/grupo a la que se va a administrar el test, las propiedades psicométricas del test podrán verse comprometidas. Un test puede incluir más de una muestra de normalización. Éste sería el caso de aquellas variables cuya distribución varía en función de la edad, grado, sexo, nivel socioeconómico… o cualquier otra que tenga relación con los resultados del test (ver Figura 3).

La puntuación empírica obtenida en un test no puede ser interpretada en términos absolutos en tanto en cuanto las escalas a partir de las cuales se obtienen las puntuaciones no tienen un 0 absoluto; ¿Qué se puede inferir de la información contenida en el siguiente enunciado? "Javier obtuvo 16 puntos

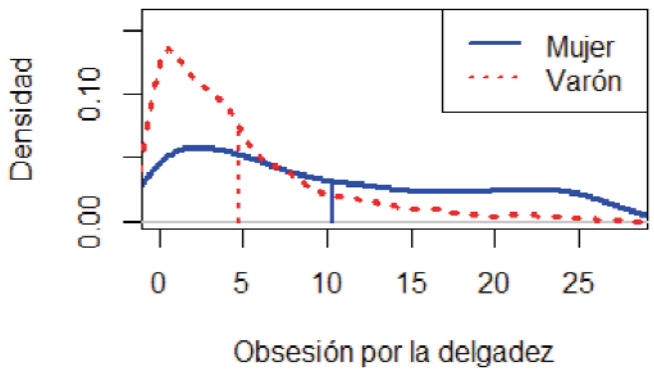
en una prueba de razonamiento lógico". Nada, si no se dispone de más información sobre la escala utilizada. La puntuación 16 sólo tiene sentido, referida al test a partir del cual se obtuvo, y con referencia a la muestra (grupo normativo) en la que se estandarizó el test. Sólo con ambas referencias (test y grupo) se podrá concluir si la puntuación 16 se sitúa por encima o por debajo de la media aritmética del grupo, o se podrá conocer cuantas desviaciones estándar por encima o por debajo de la media se sitúa el valor 16; de hecho, la media aritmética y la desviación estándar se consideran el punto 0 y la unidad de medida de la escala.

La transformación de puntuaciones engloba el conjunto de procedimientos o medidas encaminadas a situar la puntuación individual con referencia al grupo, y así definir un contexto para su interpretación sin referencia a los parámetros de escala.

Percentiles. Los percentiles son las escalas transformadas más comunes; son especialmente útiles cuando las distribuciones son asimétricas y no se ajustan a la distribución normal (p.e, cuando se analizan variables clínicas en poblaciones no clínicas; ver Figura 3).

La interpretación de los percentiles es sencilla tanto para el profesional como para el evaluado. Un percentil indica el porcentaje en la muestra de baremación que se sitúa por debajo de un determinado valor. En la lectura de los percentiles es importante tener presente que los percentiles, a) no son una transformación lineal de la puntuación, b) maximizan las diferencias en la zona centro de la distribución, y c) minimizan las diferencias entre puntuaciones brutas en las colas de la distribución. Es decir, diferencias menores en el centro de la distribución (valores cercanos a la media aritmética) se corresponden con diferencias mayores en los valores percentiles; sin embargo, como en las colas de la distribución hay un número reducido de personas, diferencias pronunciadas en los valores brutos pueden corresponderse con diferencias mínimas en los valores percentiles asociados. Por ejemplo, la tabla siguiente muestra una transformación en percentiles de una escala que tiene un valor mínimo de 25 y un máximo de 81. La media de la distribución es 50 y su desviación típica es 10. Puede comprobarse que en los valores próximos a la media un incremento de un punto (p.e 51-52) supone el cambio del P50 al P56; pero en la cola derecha de la distribución el incremento de 5 puntos en la puntuación empírica (p.e. 70- 75) se asocia con un cambio del P98 al P99.

FIGURA 3
EJEMPLO DE DISTRIBUCIÓN ASIMÉTRICA Y DIFERENTE PARA DOS GRUPOS NORMATIVOS (MUJERES/VARONES)



Puntuaciones típicas. Las puntuaciones típicas indican la localización de una puntuación con referencia a la media de la distribución. Ofrecen una medida de distancia, positiva o negativa, con referencia a la media. Son transformaciones lineales de las puntuaciones brutas, y como tales no modifican la forma de la distribución. Las puntuaciones típicas permiten comparar la localización de una persona en varios tests diferentes. Las puntuaciones típicas más comunes son las puntuaciones típicas estandarizadas (puntuaciones z) que se distribuyen con una media aritmética de 0 y una distribución típica de 1, y las puntuaciones T (Media aritmética de 50 y desviación típica de 10). Las puntuaciones de Cociente Intelectual (CI) son también puntuaciones típicas distribuidas con una media de 100 y una desviación típica de 15 (Escalas Wechsler) o 16 (Standford-Binet).

La correcta interpretación de las puntuaciones típicas ha de tener en cuenta si la distribución muestral se ajusta o no a la distribución normal. En el primer caso, podrían hacerse afirmaciones del tipo,

TABLA 2
EJEMPLO DE TRANSFORMACIÓN PERCENTIL

Per.	0	1	2	3	4	5	6	7	8	10	11	13	15	17	20	23	26	28	31	34	38
X	25	26	28	30	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48

41	45	50	56	59	63	65	68	71	75	77	81	83	86	87	89	92	94	95	96	98	99	100	Per.
49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	69	70	75	81	X

"el 95,44% de la población se encuentra en el intervalo definido por la media ± dos desviaciones típicas" (ver Figura 4). Sin embargo, si la distribución original no es una distribución normal, sentencias como la anterior serían incorrectas.

Puntuaciones normalizadas. Una distribución siempre y cuando no se aleje de la normalidad puede "normalizarse" por medio de transformaciones matemáticas que permitirán estimar las puntuaciones típicas correspondientes. Este tipo de puntuaciones se denominan puntuaciones típicas normalizadas. Se encuentran en esta categoría los *Estaninos* o *Eneatipos* (Media=5; DT=2) y los *Decatipos* (media=55; DT=2). Los primeros dividen la distribución en 9 segmentos mientras que los segundos lo hacen en 10 (Ver Figura 5).

En la lectura y valoración de las tablas de baremación incluidas en los tests es importante recordar que las poblaciones cambian a lo largo del tiempo y las tablas pueden quedar obsoletas.

4.3.2.Intepretación criterial

Los tests normativos, o test de norma de grupo, se interpretan con relación a la ejecución de un grupo y se basan como tal en un modelo de diferencias individuales. Sin embargo, si hubiera un criterio objetivo con el que comparar la puntuación, ésta podría interpretarse con referencia a su nivel alcanzado con respecto al dominio evaluado. En la interpretación criterial, existe un dominio bien definido en términos de conducta que permite definir puntos de corte en función de niveles prefijados. En este marco, interesa especialmente la composición analítica del contenido de la medida o la especificación de los procesos implicados en toda ejecución.

TABLA 3
TRANSFORMACIONES LINEALES MÁS COMUNES

Escala	Media aritmética	Desviación Típica
Z	0	1
T	50	10
WAIS	100	15
Standford-Binet	100	16
MMPI	50	10
Delta	4	13
Estanino o Eneatipo	5	2
Decatipos	5,5	2

FIGURA 4
DISTRIBUCIÓN NORMAL:
PUNTUACIONES TÍPICAS Y PORCENTAJES ACUMULADOS

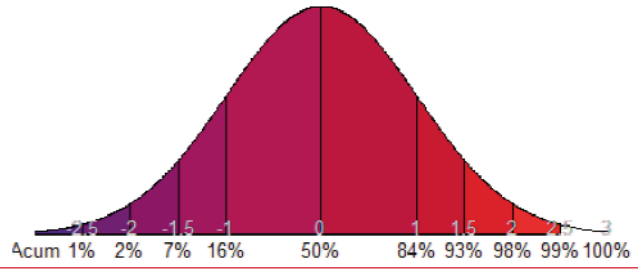
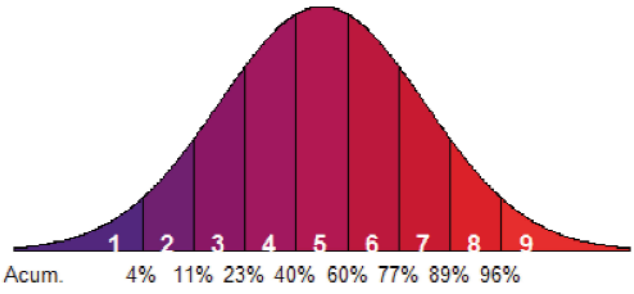


FIGURA 5
DISTRIBUCIÓN NORMAL:
ESTANINOS Y PORCENTAJES ACUMULADOS



La interpretación criterial de una puntuación (Berk, 1984; Hambleton, 1983; Popham, 1984) se apoya en la definición de un referente externo respecto al cual se comparan los niveles de ejecución. En este sentido, se informa sobre el porcentaje de aciertos en un dominio determinado, sobre la probabilidad de responder correctamente a un ítem, o sobre la probabilidad de presentar determinada patología o rasgo. La adopción de una interpretación criterial de las puntuaciones exige que se especifique claramente el método y el procedimiento utilizados para su determinación (Cizek y Bunch, 2007; Hambleton y Pitoniak, 2006). Por ejemplo en la evaluación de la capacitación lingüística, el Marco Europeo Común de Referencia para las lenguas (MECR) establece una serie de niveles para todas las lenguas (A1, A2, B1, B2, C1, C2). La determinación de estos niveles favorece la comparación u homologación de los distintos títulos emitidos por las entidades certificadoras (DELE- diploma de Español como segunda lengua, exámenes Cambridge…). En la tabla 4 se muestran las especificaciones que definen el nivel A1.

Con una especificación como la anterior, el objetivo del test es obtener información para valorar el grado de competencia conseguido; es un grado externo, fijado de antemano que no depende del grupo de personas que realice la prueba. Piénsese de

lo absurdo de la situación en que una persona obtiene un nivel de capacitación en función del resto de las personas que realizan la prueba.

La lectura criterial de los resultados adquiere una relevancia mayor cuando el objetivo de la evaluación es un diagnostico diferencial. En estos contextos los argumentos que se aportan durante el proceso de validación tienen que incluir información sobre la plausibilidad de las inferencias con relación a un diagnóstico. En este contexto de uso no es un argumento válido la descripción de las medias aritméticas obtenidas por diferentes grupos.

La distinción teórica (norma-criterio), no es mutuamente excluyente, pero su utilización conjunta ha de estar basada en una documentación que justifique tanto la una como la otra.

4.3.3.Administración, corrección e informes

La administración, corrección y elaboración del consiguiente informe de evaluación son tareas ineludibles en el establecimiento de procedimientos estandarizados de medida. Solo cuando las oportunidades y las condiciones de examen son equitativas puede hablarse de medidas estandarizadas. El estudio de las condiciones óptimas de examen no excluye, al contrario, exige, considerar medidas de acomodación para aquellos evaluandos que las necesiten, bien por no alcanzar el nivel de dominancia lingüístico necesario para una correcta evaluación, bien por la presencia de discapacidades motoras o de otro tipo.

La importancia de una adecuada elaboración de los informes de evaluación está siendo reconocida de forma unánime por la comunidad psicométrica (Hambleton y Zenisky, 2013). La información sumativa, diagnóstica y normativa incluida en un informe ha de estar redactada de forma inteligible y clara para el destinatario final. Tiene que ofrecer información sobre la calidad de la medida y su interpretación conforme al propósito del test de una manera efectiva.

5. EVALUACIÓN DE LOS TESTS

En la selección de un test para su uso es importante que el profesional evalúe o pueda acceder a evaluaciones externas sobre los tests disponibles. Con esta finalidad la Comisión de tests del Consejo General de Colegios Oficiales de Psicólogos inició el año 2010 un proceso de revisión de los tests utilizados en España. Hasta la fecha (septiembre 2013) se han llevado a cabo dos ediciones del proceso de revisión –está en marcha la tercera - que han dado lugar a informes detallados sobre los tests evaluados (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Alvarez y Peña-Suarez, 2011; Ponsoda y Hontangas, 2012). La evaluación ha utilizado como referente el modelo de evaluación de tests de la EFPA (Evers y col., 2013; Prieto y Muñiz, 2000) (Ficha 2).

El modelo de evaluación tiene tres partes destinadas a la descripción del instrumento (descripción general, clasificación, Medición y corrección, Informes computerizados, Condiciones de compra y coste) a su evaluación (calidad del modelo sustantivo, calidad de los materiales, baremos, fiabilidad, validez, calidad de los informes generados, evaluación final) y a la referencias. El proceso de revisión consiste en un análisis independiente de cada test por dos revisores ciegos con experiencia en el área de los tests que es coordinado por un editor. Los informes completos del proceso de revisión de los 20 tests evaluados hasta la fecha pueden consultarse en http://www.cop.es/index.php?page=evaluacion-tests-editados-en-espana

TABLA 4
ESPECIFICACIONES PARA EL NIVEL DE COMPETENCIA LINGÜÍSTICA A1

Comprender		Hablar		Escribir
Comprensión auditiva	**Comprensión de lectura**	**Interacción oral**	**Expresión oral**	**Expresión escrita**
Reconozco palabras y expresiones muy básicas que se usan habitualmente, relativas a mí mismo, a mi familia y a mi entorno inmediato cuando se habla despacio y con claridad.	Comprendo palabras y nombres conocidos y frases muy sencillas, por ejemplo las que hay en letreros, carteles y catálogos.	Puedo participar en una conversación de forma sencilla siempre que la otra persona esté dispuesta a repetir lo que ha dicho o a decirlo con otras palabras y a una velocidad más lenta y me ayude a formular lo que intento decir. Planteo y contesto preguntas sencillas sobre temas de necesidad inmediata o asuntos muy habituales.	Utilizo expresiones y frases sencillas para describir el lugar donde vivo y las personas que conozco.	Soy capaz de escribir postales cortas y sencillas, por ejemplo para enviar felicitaciones. Sé rellenar formularios con datos personales, por ejemplo mi nombre, mi nacionalidad y mi dirección en el formulario del registro de un hotel.

TABLA 5
TESTS EVALUADOS EN LAS EDICIONES DE 2011 Y 2012

TESTS EVALUADOS EN 2011		TESTS EVALUADOS EN 2012	
WISC-IV	Escala de inteligencia de Wechsler para niños-IV	CEAM	Cuestionario de estrategias de Aprendizaje y Motivación
MMPI-2-RF	Inventario Multifásico de la Personalidad de Minnesota-2 Reestructurado	ESCOLA	Escala de Conciencia Lectora
16PF-5	Dieciséis Factores de Personalidad, quinta edición	ESPERI	Cuestionario para la detección de los trastornos del comportamiento en niños y adolescentes
PROLEC-R	Batería de Evaluación de procesos Lectores, revisada	EPV-R	Escala de predicción del riesgo de violencia grave contra la pareja. Revisada
EFAI	Evaluación Factorial de la Aptitudes Intelectuales	WNV	Escala no verbal de aptitud intelectual de Wechsler
NEO PI-R	Inventario de Personalidad NEO Revisado	BDI-II	Inventario de Depresión de Beck-II
EVALUA	Batería Psicopedagógica	BAI	Inventario de Ansiedad de Beck
IGF	Batería de Inteligencia General y Factorial	RIAS	Escalas de inteligencia de Reynolds: RIAS y RIST
		PAI	Inventario de Evaluación de la Personalidad
		MPR	Escalas de desarrollo Merrill-Palmer revisadas
		CompeTEA	CompeTEA
		BAS-II	Escalas de aptitudes intelectuales

El objetivo del proyecto es evaluar la calidad de los tests con el propósito de ofrecer a los usuarios la información necesaria para tomar decisiones sobre los instrumentos de medida.

6. EPÍLOGO

Los tests son instrumentos de medida que asisten al profesional en su quehacer diario. Se rigen por principios psicométricos de incuestionable cumplimiento, que únicamente un uso correcto puede salvaguardar. Rigor técnico y uso apropiado se conjugan para que los tests alcancen el propósito para el que fueron construidos.

Las organizaciones nacionales e internacionales relacionadas con el uso de los tests tratan de mejorar la práctica profesional utilizando como argumento principal la formación. Aunque la principal base formativa del profesional es el grado ofrecido en nuestras universidades, éste puede no ser suficiente (Muñiz y Fernández-Hermida, 2010). Los contenidos de los planes de estudio se asientan sobre la inercia marcada por años de tradición y usos, y se retroalimenta con respecto a las prácticas establecidas. La psicometría ha evolucionado, y la distancia entre teoría psicométrica y práctica profesional es hoy mayor que nunca. Teniendo en cuenta esa distancia, este breve documento ha intentado incidir desde una perspectiva aplicada en aquellos puntos que pueden favorecer una mejor comprensión de los que es un test con el fin de mejorar su uso.

AGRADECIMIENTOS

Trabajo financiado parcialmente por el Ministerio de Economía y Competitividad (PSI2011-30256) y por la Universidad del País Vasco (GIU11-33)

REFERENCIAS

Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid: Síntesis

Ackerman, T. A. (1992). Didactic Explanation of Item Bias, Item Impact and Item Validity from a Multidimensional Perspective. *Journal of Educational Measurement, 29*, 67-91.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for European Psychologist educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi, A. (1954). *Psychological testing*. New York: Macmillan.

Bartram, D. (2011). Contributions of the EFPA Standing Committee on Tests and Testing (SCTT) to standards and good practice. *European Psychologist*, 16, 149-159.

Binet, A., y Simon, Th. A. (1905). Méthode nouvelle pour le diagnostic du niveau intellectuel des anormaux. *L'Année Psychologique*, 11, 191-244.
Berk, A. (1984). *A guide to criterion-referenced test construction*. The Johns Hopkins Univ. Press, Baltimore
Bollen, K.A. (1989). *Structural Equations with Latent Variables.* New York: Wiley
Camilli, G., y Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. London: Sage.
Cizek, G. J., y Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.
Cole, N. S., y Moss, P. A. (1989). Bias in Test Use. En R.L.Linn (Ed.), *Educational Measurement. Third Edition* (pp. 201-219). New York: American Council on Education and Macmillan Publishing Company.
Cronbach, L.J. (1951). Coefficient alpha and the internal consistency of tests. *Psychometrika,16*, 297–334.
Embretson, S. E. y Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
Elosua, P. (2003). Sobre la validez de los tests. *Psicothema* ,15(2), 315-321.
Elosua, P. (2008). Una aplicación de la estimación Bayes empírica para incrementar la fiabilidad de las puntuaciones parciales. *Psicothema,* 20, 497-503
Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema* 21,4, 652-655.
Elosua, P. (2011). Psicometría. Conceptos básicos y aplicaciones prácticas con Rcommander. Leioa, UPV/EHU. Accesible en http://www.ehu.es/argitalpenak/images/stories/libros_gratuitos_en_pdf/Ciencias_Sociales/ Psicometria.%20Conceptos%20basicos%20y%20aplicaciones%20practicas%20con%20Rcommander.pdf.
Elosua, P. (2012). Tests publicados en España: Usos, costumbres y asignaturas pendientes. *Papeles del Psicólogo*, 33, 12-21.
Elosua, P., Hambleton, R., y Muñiz, J. (en prensa). *Teoría de la Respuesta al item aplicada con R*. Madrid: La Muralla.
Elosua, P., e Iliescu, D. (2012). Tests in Europe. Where we are and where we should to go. *International Journal of Testing*, 12, 157-175.
Elosua, P. y Zumbo, B. (2008). Reliability coefficients for ordinal response scales. *Psicothema,* 20, 896-901.
Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., Vizcarro, C., et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment*, 17, 187-200
Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., FernándezHermida, J.R., y col. (2012). Testing practices in the 21st Century: Developments and European psychologists' opinions. *European Psychologist*, 17, 300-319.
Evers, A., Muñiz, J., Hagemeister, C., Høstmælingen, A., Lindley, P., Sjöberg, A., & Bartram, B. (2013). Assessing the quality of tests: Revision of the EFPA review model. *Psicothema*, 25, 283-291.
Fernández-Ballesteros, R., De Bruyn, E.E.E., Godoy, A., Hornke, L.F., Ter Laak, J., Vizcarro, C., Westhoff, W., Westmeyer, H., y Zaccahgnini, J.L. (2001). Guideliness for the Assessment Process (GAP): A proposal for discussion. *European Journal of Psychological Assessment*, 17, 187-200.
Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.
Hambleton, R. K. (1983): Application of item response models to criterion referenced assessment. *Applied Psychological Measurement*, 7, 1,33-44.
Hambleton, R. K. y Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston, Kluwer: Nijhoff Publishing.
Hambleton, R. K. y Pitoniak, M. (2006). Setting performance standards. . En R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.
Hambleton, R.K. y Zenisky, A. (2013). Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design. En Kurt F. Geisinger, Kurt F. (Ed); *APA handbook of testing and assessment in psychology,* Vol. 3 (pp. 479-494). Washington, D.C.: American Psychological Association,
Haynes, S. N., Richard, D. R., y Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, 7, 238-247.
ISO (2011). *Procedures and methods to assess people in work and organizational settings (part 1 and 2)*. Geneva: Author
Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.
Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
Kane, M. T. (2006). Validation. En R. Brennan (Ed.), *Educational measurement*, 4th ed (pp. 17-64). Westport, CT: Praeger

Kelley T.L. (1927). *Interpretation of educational measurements*. Yonkers, NY, World Book Company.
López-Jáuregui, A., y Elosua, P. (2013, Julio). A Comparison of Classification Methods for Identifying the Presence/Absence of Eating Disorders. Comunicación presentada en el *13th European Congress of Psychology* . Estocolmo.
Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
McDonald, R. P. (1999). *Test theory. A unified treatment.* Mahwah, NJ, Lawrence Erlbaum Associates.
Mellenbergh, G. J. (1989). Item Bias and Item Response Theory. *International Journal of Educational Research, 13*, 127-143.
Messick, S. (1995). Validity of psychological assessment. *American Psychologist* , 50, 741–749.
Millsap, R. y Maydeu-Olivares, A. (Eds.) (2009). *Handbook of Quantitative Methods in Psychology*. London: Sage.
Muñiz, J. (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.
Muñiz, J. (2010, Julio). Estrategias para mejorar el uso de los tests. Comunicación presentada en el *Congreso Iberoamericano de Psicología, Oviedo.*
Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., y Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment,* 17, 201-211.
Muñiz, J., Elosua, P., y Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda. *Psicothema*, 25, 151-157.
Muñiz, J. y Fernández-Hermida, J.R. (2010). La opinión de los psicólogos españoles sobre el uso de los tests. *Papeles del Psicólogo*, 31, 108-121.
Muñiz, J. y Fernández-Hermida, J.R., Fonseca-Pedrero, E., Campillo-Alvarez, A. y Peña-Suarez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo*, 32, 113-128.
Nunnally, J.C. (1978). *Psychometric theory* , New York: McGraw-Hill.
Olea, J., Abad, F., y Barrada, J. R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del Psicólogo*, 31, 94-107.
Ponsoda, V., y Hontangas, P. (2013). Tests editados en España (Segunda evaluación). *Papeles del Psicólogo*, 34, 82-90.
R Core Team (2012). R: *A language and environment for statistical computing*.Vienna: R Foundation for Statistical Computing.
Popham, W. J. (1983). *Evaluación basada en criterios*. Madrid: Magisterio Español
Prieto, G., y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo*, 77, 65-71.
Raju, N. S., Burke, M. J., y Norrmand, J.(1990). A new approach to utility analysis. *Journal of Applied Psychology*, 75: 3-12.
Sireci, S. G. (2007). On validty theory and test validation. Educational Researcher, 36,477-481.
Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology* 15, 201-293.
Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin* 99, 100–117.
Swets, J. A., y Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection theory*. New York: Academic.
Taylor, H.C., y Russell. J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection. *Journal of Applied Psychology*, 23, 565-578
Thurstone, L. L. (1924/1973).*The Nature of Intelligence*. London: Routledge.
Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: University of Chicago Press.
Van der Linden, W. y Hambleton, R.K. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.
Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum
Zumbo, B. D., Gadermann, A. M. y Zeisser, C. (2007). Ordinal Versions of Coefficients Alpha and Theta For Likert Rating Scales. *Journal of Modern Applied Statistical Methods*, 6, 21-29.
Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. En C. Rao y S. Sinharay (Eds.). Handbook of statistics. Vol.26. Psychometrics (pp. 45-79). Amsterdam: Elsevier Science

Ficha 1.

Directrices internacionales para el uso de tests

INTRODUCCIÓN

Los tests constituyen una de las tecnologías más utilizadas por los psicólogos en el ejercicio de la profesión. Como ocurre con la tecnología de otras áreas científicas, los tests pueden utilizarse correcta o incorrectamente. Las directrices que se proponen en este documento tienen como objetivo fundamental mejorar el uso que los profesionales hacen de los tests.

Las directrices han sido elaboradas originalmente por la *Comisión Internacional de Tests (ITC)*, bajo la dirección del profesor Dave Bartram. La versión en castellano que presentamos aquí ha sido realizada en el seno de la *Comisión de Tests* del Colegio Oficial de Psicólogos (COP). El Colegio, a través de sus representantes en la citada Comisión Internacional de Tests y en la Comisión Europea sobre Tests de la Federación Europea de Asociaciones Profesionales de Psicólogos (EFPPA), que también ha respaldado las directrices, ha participado activamente en la elaboración de las mismas, consciente de la necesidad de contribuir a mejorar la práctica de los tests en nuestro país.

Los documentos más utilizados en la elaboración de las directrices, junto con otros en castellano, aparecen citados al final; documentación complementaria en castellano puede consultarse en la sección de la *Comisión de Tests* incluida en la página web del COP (http://www.cop.es/tests/).

Las directrices reflejan principios generales implicados en el uso adecuado de los tests, pero no pretenden uniformar las diferencias legítimas existentes entre la práctica en los diferentes países o áreas profesionales. Existen bastantes ra-

zones por las que se necesitan unas directrices para el uso de los tests con carácter internacional. Veamos algunas de ellas.

a) Los países difieren ampliamente en el grado de control legal que pueden ejercer sobre el uso de los tests y sus consecuencias para las personas evaluadas. Por tanto, la existencia de un conjunto de directrices internacionalmente aceptadas proporcionará a las asociaciones nacionales de psicología y otras organizaciones profesionales una buena base documental para completar directrices ya existentes o proceder a crearlas.
b) El derecho a comprar y utilizar materiales psicométricos varía mucho de unos países a otros. En algunos países el uso de los tests está restringido a los psicólogos, en otros a quienes estén registrados por los distribuidores de las pruebas, e incluso en algunos los usuarios pueden obtener el material psicométrico libremente, sin restricciones.
c) Algunos tests bien conocidos han aparecido en Internet sin la autorización de autores y editores, violando los derechos de *copyright*, y sin ninguna precaución en relación con la seguridad del test.
d) En el ámbito de la psicología del trabajo, el aumento de la movilidad internacional de los trabajadores ha incrementado la demanda de tests para utilizarlos con aspirantes de diferentes países. A menudo los tests son aplicados en un país para empresas ubicadas en otros países.
e) En países como Estados Unidos e Inglaterra, entre otros, se están llevando a cabo trabajos sobre evaluación a distancia vía Internet en las áreas de Trabajo y Educativa. Esto plantea todo un conjunto de problemas en relación con las normas de aplicación, control del proceso de evaluación y seguridad de los resultados.

Por todas estas razones, y otras que se podrían enumerar, parece conveniente disponer de unas directrices internacionales que permita una referencia común a la hora de establecer normas nacionales para la mejora del uso de los tests.

Las presentes directrices recogen el trabajo de especialistas en evaluación psicológica y educativa (psicólogos, psicómetras, editores y constructores de tests) de diferentes países, entre ellos España. La idea no ha sido tanto el inventar unas nuevas directrices, como juntar las hebras comunes que se encuentran en las distintas directrices, códigos de práctica y estándares ya existentes, formando una estructura coherente que resulte comprensible y de uso práctico.

DIRECTRICES

Objetivo

El objetivo fundamental de las directrices es mejorar el uso de los tests, describiendo la forma adecuada de utilizarlos.

Un usuario competente utilizará los tests de forma adecuada, profesional y ética, prestando la debida atención a las necesidades y derechos de las personas implicadas en el proceso de evaluación, y teniendo muy en cuenta las razones para utilizar los tests, así como el contexto en el cual se lleva a cabo su aplicación. Este objetivo se alcanzará asegurándose de que el usuario de los tests tiene las competencias y conocimientos necesarios para llevar a cabo el proceso evaluativo.

Ámbito de aplicación

Resulta difícil dar una definición muy precisa de *test* o de su *práctica*, pues al hacerlo pueden dejarse fuera algunos procedimientos que debieran incluirse, o, por el contrario, incluir otros que deberían quedar fuera. En estas directrices los términos *test* y *práctica de los tests* se utilizan en sentido amplio. Las directrices son aplicables a muchos procedimientos de evaluación que no suelen denominarse *tests*, o que incluso tratan de evitar su designación como *tests*. Más que proponer una definición cerrada, se describe a continuación el ámbito que tratan de cubrir las directrices.

- ✔ Los tests incluyen un amplio abanico de procedimientos utilizados en la evaluación psicológica, educativa y ocupacional.
- ✔ Los tests incluyen procedimientos para la medición de conductas tanto normales como anormales o disfuncionales.
- ✔ Los tests son procedimientos diseñados para ser aplicados bajo condiciones controladas o estandarizadas, y conllevan la utilización de protocolos de puntuación rigurosos.
- ✔ Estos procedimientos proporcionan medidas de ejecuciones y conllevan la obtención de inferencias a partir de muestras de conducta. También pueden incluir procedimientos que proporcionan clasificaciones cualitativas u ordenamientos de las personas.

Cualquier procedimiento utilizado de la forma descrita arriba puede considerarse como un test, independientemente de su forma de aplicación, el tipo de profesional que lo haya construido, o si requiere contestar a ítems o ejecutar ciertas tareas u operaciones.

Los tests deben de estar apoyados por datos empíricos sobre su fiabilidad y validez para medir los objetivos que se proponen. Hay que aportar datos que justifiquen las inferencias que se hacen a partir de las puntuaciones de los tests. Estos datos tiene que estar disponibles para los usuarios de los tests, así como para los profesionales e investigadores que deseen llevar a cabo una evaluación o revisión independiente.

Las directrices que se presentan aquí son aplicables a todos los procedimientos descritos, se autodenominen o no tests. Deben de tenerse en cuenta para cualquier procedimiento de evaluación utilizado en situaciones en las que la evaluación tiene serias implicaciones para las personas, pudiendo causarles daños personales o psicológicos si no se realiza adecuadamente. Las directrices no se aplican al uso de materiales que pueden tener una semejanza superficial con los tests, pero que los propios participantes reconocen que se utilizan con fines de diversión o entretenimiento (por ejemplo, los cuestionarios de estilos de vida de los periódicos y revistas).

Destinatarios

Las directrices se aplican al uso de los tests en la práctica profesional, por tanto van dirigidas fundamentalmente a:

- ✔ Los compradores y vendedores de materiales psicométricos
- ✔ Los responsables de elegir los tests y el uso que se hará de ellos
- ✔ Quienes aplican, puntúan e interpretan los tests
- ✔ Quienes aconsejan a otros basándose en los resultados de los tests (consultores en ámbitos educativos y laborales, orientadores escolares y profesionales, etc.)
- ✔ Las personas encargadas de dar información sobre los resultados de los tests a las personas evaluadas
- ✔ Los constructores de tests
- ✔ Los editores de tests
- ✔ Las personas implicadas en el entrenamiento de los usuarios de los tests
- ✔ Los propios evaluados y las personas cercanas a ellos (padres, cónyuges, etc.)
- ✔ Organizaciones profesionales y otras asociaciones interesadas en el uso de los tests - Legisladores y responsables políticos
- ✔ Usuarios de los tests con fines de investigación

Si bien las directrices no pretenden cubrir todo tipo de técnica o situación evaluativa, muchas de las directrices pueden ser aplicables en situaciones evaluativas que no se identifican estrictamente con la práctica de los tests, tales como los centros de evaluación para la selección de personal, las entrevistas estructuradas y semiestructuradas, orientación escolar y profesional, etc.

Aspectos contextuales

Las directrices son aplicables internacionalmente, pero a la hora de utilizarlas a nivel local en un país determinado deben de tenerse en cuenta diversos factores o aspectos que pueden modular su aplicación. Entre otros, estos factores contextuales incluyen:

- ✔ Diferencias sociales, políticas, institucionales, lingüísticas y culturales
- ✔ La legislación del país en el que se utilizan los tests
- ✔ Directrices y estándares ya existentes establecidos por asociaciones profesionales
- ✔ Diferencias entre la evaluación individual y la colectiva
- ✔ Diferencias en función de la situación evaluativa (educativa, clínica, trabajo, otras)
- ✔ Quienes son los receptores de los resultados (los propios evaluados, los padres o responsables legales, los constructores de los tests, los empleadores, etc.)
- ✔ Diferencias según el uso de los resultados de los tests (por ejemplo, toma de decisiones frente a orientación)
- ✔ Grado en el que la situación permite la posibilidad de comprobar la precisión de las predicciones y su posible modificación a la luz de los resultados posteriores

Conocimientos

Unos conocimientos psicológicos psicométricos sólidos y una comprensión profunda de todos los aspectos implicados en el proceso evaluativo constituyen la base fundamental para el uso pertinente de los tests. Los expertos suelen estar de acuerdo en que la causa más importante del uso inapropiado de los tests es una formación deficiente de los usuarios. Los conocimientos concretos requeridos en cada caso variarán en función de la situación y área de aplica-

ción, por tanto no es fácil describir de forma general los conocimientos requeridos para la utilización adecuada de los tests en todas las situaciones, depende de cada caso. Por ejemplo, no se requieren los mismos conocimientos para aplicar un test colectivo a un grupo que para interpretar las puntuaciones, o construir el propio test, por citar un caso obvio. Las directrices no hacen descripciones detalladas de estos conocimientos, ni lo pretenden. No obstante, a la hora de aplicarlas a una situación concreta habrá que tener muy presente y describir con precisión los conocimientos requeridos para el caso. Esta descripción ha de cubrir las distintas áreas de conocimientos implicadas en la situación de que se trate, entre ellas deberían de incluirse las siguientes: a) Conocimientos sobre teoría de los tests y propiedades técnicas de los tests, tales como fiabilidad, validez, estandarización, sesgo, análisis de ítems, etc. b) Conocimiento de los tests y principios de la medición para entender adecuadamente los resultados. c). Conocimientos sobre la teoría, modelos y constructos medidos, que permita una elección pertinente de las pruebas e interpretación de los resultados. d). Conocer el abanico de tests disponibles, así como los editores correspondientes. e). Avances técnicos recientes, tales como los tests informatizados, bancos de ítems, etc. Otros muchos conocimientos específicos han de ponerse en funcionamiento para llevar a cabo una utilización adecuada de los tests, cada una de las directrices puede servir como indicador de la naturaleza de éstos.

1. USO ÉTICO DE LOS TESTS

Los usuarios competentes deberían:

1.1 Actuar de forma ética y profesional

1.1.1. Mantener y promover estándares éticos y profesionales

1.1.2. Estar al corriente de los debates profesionales y éticos sobre el uso de los tests en su campo de especialización

1.1.3. Llevar a cabo una política explícita sobre los tests y su uso (véase ejemplo en el Apéndice A)

1.1.4. Asegurarse de que las personas para las que trabajan, o con quienes trabajan, se acogen a estándares éticos y profesionales adecuados

1.1.5. Actuar con el debido respeto para la sensibilidad de las personas evaluadas y de otras personas o instituciones implicadas

1.1.6. Presentar la práctica de los tests de forma positiva y equilibrada cuando interactúan con los medios de comunicación

1.1.7. Evitar las situaciones en las que pueda parecer que existen determinados intereses en los resultados de la evaluación, o en las cuales la evaluación pueda dañar su relación con los clientes

1.2. Asegurarse de que son competentes para el uso de los tests

1.2.1. Trabajar de acuerdo con los principios científicos

1.2.2. Establecer y mantener elevados estándares personales de competencia

1.2.3. Conocer los límites de la propia competencia y no actuar fuera de ellos

1.2.4. Mantenerse al día de los cambios y avances en relación con el uso de los tests y de su construcción, incluyendo los cambios de normas y legislación, los cuales pueden influir en los tests y su uso

1.3. Responsabilizarse del uso que hacen de los tests

1.3.1. Ofrecer solo servicios y usar tests para los cuales están preparados

1.3.2. Aceptar la responsabilidad por los tests elegidos y por las recomendaciones proporcionadas

1.3.3. Dar una información clara y adecuada a los participantes en el proceso evaluativo sobre los principios éticos y las disposiciones legales que regulan el uso de los tests

1.3.4. Asegurarse de que el contrato entre los evaluados y los evaluadores es claro y se ha comprendido (véase un ejemplo de contrato en el Apéndice B)

1.3.5. Estar atentos a cualquier consecuencia imprevista del uso de los tests

1.3.6. Esforzarse para evitar cualquier tipo de daño o perjuicio a las personas evaluadas

1.4. Asegurarse de que los materiales del test están seguros

1.4.1. Asegurar un almacenaje seguro y controlar el acceso a los materiales del test

1.4.2. Respetar el *copyright* y los acuerdos que existan sobre el test, incluyendo cualquier prohibición sobre la copia y transmisión de los materiales, bien sea electrónicamente o de otra forma. Asimismo se respetarán rigurosamente los términos del acceso de otras personas, cualificadas o no, a los materiales.

1.4.3. Proteger la integridad del test, no entrenando a las personas con los propios materiales del test, o con otros materiales de prácticas que puedan influir de forma inapropiada en el rendimiento de las personas en el test

1.4.4. Asegurarse de que la tecnología del test no se expone públicamente de tal modo que su utilidad quede deteriorada

1.5. Asegurarse de que los resultados de los tests se tratan confidencialmente

1.5.1. Especificar quienes tendrán acceso a los resultados y definir los niveles de confidencialidad
1.5.2. Explicar los niveles de confidencialidad antes de dar los resultados
1.5.3. Limitar el acceso a los resultados únicamente a quienes tengan la necesidad y derecho a conocerlos
1.5.4. Obtener las autorizaciones pertinentes antes de proporcionar los resultados a otros
1.5.5. Proteger los datos archivados de tal forma que sólo puedan acceder a ellos quienes tengan derecho a hacerlo
1.5.6. Establecer una directrices claras en relación con el tiempo que se van a mantener archivados los datos
1.5.7. Suprimir el nombre y otros datos identificatorios de los resultados si así lo solicita la persona evaluada
1.5.8. Suprimir el nombre y otros datos identificativos de las bases de datos de los resultados, con fines de investigación, elaboración de baremos, u otros tratamientos estadísticos

2. UTILIZACIÓN ADECUADA DE LOS TESTS

2.1. Estimar la utilidad potencial de los tests en una situación evaluativa

Los usuarios competentes deberían:

2.1.1. Ofrecer una justificación razonada para el uso de los tests
2.1.2. Asegurarse de que se ha llevado a cabo un análisis riguroso de las necesidades del cliente, categoría diagnóstica, condiciones, o trabajo para el que se utilizará la evaluación
2.1.3. Comprobar que los conocimientos, destrezas, aptitudes, u otras características, que miden los tests correlacionan con las conductas pertinentes en el contexto en el que se van a llevar a cabo las inferencias
2.1.4. Buscar otras fuentes fuentes adicionales de información
2.1.5. Sopesar las ventajas e inconvenientes de utilizar tests frente a otras fuentes de información
2.1.6. Asegurarse de que se utilizan todas las fuentes de información colateral disponibles

2.2. Elegir tests técnicamente correctos y adecuados a cada situación

Los usuarios competentes deberían:

2.2.1. Examinar toda la información disponible sobre los tests potencialmente adecuados antes de elegir un test concreto
2.2.2. Comprobar que la documentación técnica sobre el test proporciona suficiente información para evaluar los siguientes aspectos:
a. Amplitud y representatividad del contenido del test, adecuación de los grupos normativos utilizados, nivel de dificultad de los contenidos, etc.
b. Precisión de la medición y fiabilidad para las poblaciones pertinentes
c. Validez para las poblaciones pertinentes y su aplicabilidad para el uso que se hace del test
d. Ausencia de sesgo para los grupos con los que se utilizará
e. Aceptación por parte de quienes están implicados en su uso, incluyendo la pertinencia y validez aparente percibidas
f. Aspectos prácticos, tales como tiempo requerido, coste, o recursos que se necesitan
2.2.3. Evitar el uso de tests que tengan una documentación técnica inadecuada o poco clara
2.2.4. Utilizar tests sólo para aquellos objetivos para los cuales se dispone de una validez empírica adecuada y pertinente
2.2.5. No aceptar un test basándose únicamente en su validez aparente, recomendaciones de otros usuarios, o consejos de quienes tienen intereses comerciales
2.2.6. Responder a las preguntas de las personas implicadas (personas evaluadas, padres, supervisores, representantes legales, etc.), dándoles suficiente información para que entiendan por qué se eligió el test

2.3. Prestar atención a los aspectos relacionados con el sesgo de los tests

Cuando los tests se van a utilizar con personas de diferentes grupos (por ejemplo: género, cultura, educación, etnia, origen, o edad, entre otros), los usuarios competentes de los tests harán todos los esfuerzos posibles para asegurarse de que:

2.3.1. Los tests son imparciales y adecuados para todos los grupos evaluados
2.3.2. Los constructos que se están midiendo son relevantes para cada uno de los grupos evaluados
2.3.3. Existen datos disponibles sobre las diferencias de rendimiento de los grupos en el test
2.3.4. Hay datos disponibles sobre el Funcionamiento Diferencial de los Ítems cuando ello es pertinente
2.3.5. Hay datos sobre la validez que apoyan el uso del test en diferentes grupos
2.3.6. Se minimizan los efectos de las diferencias grupales no relacionadas con el objetivo de la medición
2.3.7. Las directrices sobre la imparcialidad de los tests se interpretan dentro del marco de la legislación al respecto existente en cada país.

Cuando se utilizan los tests en más de un idioma (idiomas distintos, dialectos, lenguaje de signos, etc.) los usuarios competentes harán todos los esfuerzos posibles para asegurarse de que:

2.3.8. Las versiones de los distintos idiomas o dialectos hayan sido elaboradas utilizando una metodología rigurosa
2.3.9. Los constructores hayan sido sensibles a los aspectos de contenido, culturales e idiomáticos
2.3.10. Quienes aplican los tests sean capaces de comunicarse perfectamente en el idioma en el que se aplica el test
2.3.11. El dominio de la lengua (en la que se aplicará el test) de las personas evaluadas sea comprobado sistemáticamente, utilizándose la versión más adecuada, o una bilingüe si fuese necesario

Cuando se utilizan los tests con personas que tienen alguna discapacidad, los usuarios competentes harán todo lo que sea posible para asegurarse de que:

2.3.12. Se ha recabado consejo de los expertos acerca de los efectos de la discapacidad sobre el rendimiento en el test
2.3.13. Se han consultado las personas a evaluar y se ha dado un tratamiento adecuado a sus necesidades y deseos
2.3.14. Se han llevado a cabo los ajustes oportunos cuando se evalúa a personas con discapacidades auditivas, visuales, motoras, dislexia, u otras
2.3.15. Se ha contemplado la posiblidad de utilizar procedimientos de evaluación alternativos en vez de modificaciones o ajustes de los tests
2.3.16. Se ha solicitado consejo a expertos en el caso de que el grado de modificación requerido por el test esté más allá de la experiencia y conocimientos del usuario
2.3.17. Las modificaciones, cuando sean necesarias, se ajustan a la naturaleza de la discapacidad y se han diseñado para minimizar el impacto sobre la validez de las puntuaciones
2.3.18. La información relativa a cualquier ajuste o modificación hechos en el test o en su aplicación se comunica a quienes interpretan o utilizan las puntuaciones del test, para así facilitar una interpretación apropiada de las puntuaciones

2.4. Hacer los preparativos necesarios para la aplicación del test

El usuario competente hará todo lo posible para:

2.4.1. Proporcionar en el momento oportuno una información clara a las personas implicadas en la evaluación acerca de la finalidad del uso de los tests, la mejor forma de prepararse para la sesión de tests y los procedimientos a seguir
2.4.2. Aconsejar a los evaluados acerca de los idiomas y dialectos para los que es apropiado el test
2.4.3. Proporcionar información a quienes van a hacer el test sobre el tipo de práctica permitida, así como la documentación donde pueden encontrar ejemplares y materiales oportunos
2.4.4. Explicar claramente a los evaluados sus derechos y deberes (véase Apéndice B)
2.4.5. Obtener el consentimiento explícito de los evaluados o de sus responsables o representantes legales antes de aplicar los tests
2.4.6. Cuando los tests sean opcionales, explicar a las personas implicadas las consecuencias de hacerlos o no, para que puedan tomar una decisión con fundamento
2.4.7. Hacer los necesarios ajustes prácticos para asegurarse de que:
   a. Los preparativos coinciden con los establecidos en el manual del test
   b. Los locales y otras facilidades para aplicar los tests se han reservado con antelación, el entorno físico es accesible, seguro, tranquilo, libre de distracciones, y se ajusta a las necesidades
   c. Hay suficiente material disponible y se ha comprobado que no han quedado señales de usuarios previos en los cuadernillos o en las hojas de respuestas

d. Las personas implicadas en la aplicación de los tests son competentes
e. Se han hecho los ajustes oportunos para aplicar los tests a las personas con alguna discapacidad (véase Apéndice C)

2.4.8. Prever posibles problemas y solventarlos mediante la preparación de materiales e instrucciones

2.5. Aplicar los tests adecuadamente

Los usuarios competentes deberían:

2.5.1. Establecer una buena relación con las personas evaluadas, dándoles la bienvenida y dirigiéndose a ellas de forma positiva
2.5.2. Tratar de reducir la ansiedad de las personas a las que se va a evaluar, evitando crear o reforzar ansiedad innecesaria
2.5.3. Eliminar fuentes potenciales de distracción, tales como alarmas de relojes de pulsera, teléfonos móviles, buscas, etc.
2.5.4. Asegurarse de que todas las personas disponen de los materiales necesarios para responder al test antes de comenzar éste
2.5.5. Supervisar convenientemente la aplicación de los tests
2.5.6. Dar las instrucciones en la lengua dominante de las personas evaluadas siempre que sea posible, incluso cuando el test está diseñado para proporcionar datos sobre el conocimiento o dominio de una lengua distinta de la dominante
2.5.7. Ajustarse estrictamente a las instrucciones del manual del test, haciendo los ajustes pertinentes para las personas con alguna discapacidad
2.5.8. Leer las instrucciones pausada y claramente
2.5.9. Dar el tiempo adecuado para hacer los ejemplos
2.5.10. Observar y anotar las posibles desviaciones de los procedimientos estándar del test
2.5.11. Registrar los tiempos de respuesta con precisión cuando se requiera
2.5.12. Asegurarse de que están todos los materiales al final de cada sesión
2.5.13. Realizar la aplicación de modo que permita una supervisión adecuada y una comprobación de la identidad de cada una de las personas evaluadas
2.5.14. Permitir a los ayudantes hacerse cargo de la aplicación sólo si han sido entrenados adecuadamente
2.5.15. Asegurarse de que durante la sesión aplicación no se deja desatendidas o sujetas a distracción a las personas evaluadas
2.5.16. Proporcionar una asistencia adecuada a las personas evaluadas que muestran signos excesivos de ansiedad o desazón

2.6. Puntuar y analizar los resultados de los tests con precisión

Los usuarios competentes deberían:

2.6.1. Seguir al pie de la letra los procedimientos estandarizados de puntuación
2.6.2. Asegurarse de la precisión al asignar las puntuaciones, especialmente en aquellos casos en los que entra en juego el juicio de los evaluadores. Para cerciorarse de la precisión puede volver a puntuarse una muestra de las personas evaluadas, comprobando así la coincidencia entre las puntuaciones
2.6.3. Llevar a cabo las transformaciones de las puntuaciones directas a otros tipos de escalas pertinentes
2.6.4. Elegir los tipos de escala más convenientes de acuerdo con el uso que se vaya a hacer de las puntuaciones del test
2.6.5. Comprobar la precisión de las transformaciones de las escalas, así como la de cualquier tipo de análisis o tratamiento que se haga con los datos
2.6.6. Asegurarse de que no se sacan conclusiones erróneas debido a la utilización de baremos desfasados, o inadecuados para las personas evaluadas
2.6.7. Calcular las puntuaciones compuestas cuando proceda, utilizando las fórmulas y ecuaciones propuestas en el manual del test
2.6.8. Inspeccionar los resultados para detectar posibles errores o anomalías en las puntuaciones
2.6.9. Describir e identificar con precisión los resultados, normas, tipos de escalas, fórmulas, etc. utilizados

2.7. Interpretar los resultados adecuadamente

Los usuarios competentes deberían:

2.7.1. Tener una buena comprensión profesional de las bases teóricas y conceptuales del test, de la documentación técnica y de las directrices para el uso e interpretación de las puntuaciones

2.7.2. Tener una buena comprensión profesional de las escalas utilizadas, de las normas y baremos, así como de las limitaciones de las puntuaciones

2.7.3. Tratar de minimizar cualquier sesgo que pueda existir hacia las personas evaluadas en la interpretación de las puntuaciones del test

2.7.4. Utilizar normas o grupos de comparación apropiados cuando estén disponibles

2.7.5. Interpretar los resultados a la luz de la información disponible sobre la persona evaluada (edad, género, escolaridad, cultura, etc.), teniendo en cuenta las limitaciones técnicas del test, el contexto de la evaluación, y las necesidades de las personas o instituciones con intereses legítimos en el resultado del proceso evaluativo

2.7.6. Evitar la generalización de los resultados de un test a rasgos o características de la persona que no han sido medidos por el test

2.7.7. Tener en cuenta la fiabilidad y el error de medida de cada escala, así como otros factores que puedan alterar artificialmente los resultados a la hora de interpretar las puntuaciones

2.7.8. Tener muy en cuenta los datos disponibles sobre la validez del constructo medido en relación con las características de los grupos evaluados, tales como cultura, edad, clase social, género, etc.

2.7.9. Utilizar puntos de corte en la interpretación de las puntuaciones sólo cuando se disponga de datos empíricos sobre su validez

2.7.10. Ser conscientes de los estereotipos sociales que pueden existir sobre las personas evaluadas (en relación con su cultura, edad, clase social, género, etc.), evitando interpretar los tests de forma que se perpetúen dichos estereotipos

2.7.11. Tener en cuenta cualquier variación individual o colectiva que se haya hecho respecto al procedimiento estándar en la aplicación de las pruebas

2.7.12. Tomar en consideración cualquier experiencia previa que la persona evaluada haya tenido con el test, en el caso de que se disponga de datos sobre los efectos de dicha experiencia sobre el rendimiento en la prueba

2.8. Comunicar los resultados de forma clara y precisa

Los usuarios competentes deberían:

2.8.1. Identificar las personas o instituciones pertinentes que pueden recibir los resultados de los tests

2.8.2. Elaborar informes orales o escritos para los receptores de los resultados, siempre con el consentimiento explícito de las personas evaluadas o de sus representantes legales

2.8.3. Asegurarse de que el nivel técnico de los contenidos de los informes es adecuado para su comprensión por los receptores

2.8.4. Dejar muy claro en los informes que los resultados de los tests son confidenciales, y especificar el tiempo que se mantendrán archivados los resultados

2.8.5. Dejar claro que los datos de los tests representan una sola fuente de información que debe analizarse conjuntamente con otras fuentes

2.8.6. Explicar el peso que debe de darse a las puntuaciones de los tests en relación con otras fuentes de información acerca de las personas evaluadas

2.8.7. Proporcionar la información sobre los resultados en un lenguaje comprensible para el receptor, de modo que se minimice la posibilidad de interpretaciones incorrectas

2.8.8. Utilizar una forma y estructura para el informe que encaje en el contexto de la evaluación

2.8.9. Si procede, proporcionar información a quienes toman las decisiones acerca de como pueden usar los resultados de los tests para mejorar sus decisiones

2.8.10. Explicar y fundamentar la utilización que se hace de los resultados en los tests para la clasificación de las personas en categorías con fines diagnósticos, u otros.

2.8.11. Incluir un resumen claro en los informes escritos, y, cuando sea pertinente, recomendaciones concretas

2.8.12. Dar información a las personas evaluadas de forma constructiva y positiva

2.9. Revisión de la adecuación del test y de su uso

Los usuarios competentes deberían:

2.9.1. Seguir y revisar periódicamente los posibles cambios en la población de personas evaluadas, así como los criterios utilizados para la validez

2.9.2. Estar atentos a posibles impactos negativos de los tests

2.9.3. Ser conscientes de la necesidad de reanalizar la utilización de un test si se ha llevado a cabo algún cambio en su forma, contenidos, o forma de aplicación

2.9.4. Tener presente la necesidad de reconsiderar la validez del test si se ha cambiado la finalidad para la que se utilizaba

2.9.5. Siempre que sea posible, validar los tests para los usos para los que fueron elaborados y participar en los trabajos rigurosos de validación que se lleven a cabo

2.9.6. Ayudar en la medida de sus posibilidades a mantener al día la información sobre los baremos, fiabilidad y validez del test, proporcionando los datos pertinentes a los constructores, editores o investigadores

APÉNDICE A
DIRECTRICES PARA EL ESTABLECIMIENTO DE POLÍTICAS SOBRE EL USO DE LOS TESTS

Las directrices que siguen se refieren a la necesidad de las organizaciones de considerar su política sobre la práctica de los tests de forma sistemática, asegurándose de que todas las personas implicadas en la evaluación tienen clara esta política. La necesidad de una política explícita sobre la práctica de los tests no atañe sólo a las grandes organizaciones. Las empresas de tamaño medio o pequeño, al igual que las grandes, deben de prestar atención a la política sobre la práctica de los tests, de la misma forma que lo hacen en relación con la salud y seguridad laboral, igualdad de oportunidades, discapacidades, u otras áreas relacionadas con una práctica adecuada en el tratamiento y atención al personal.

Aunque las siguientes consideraciones pueden necesitar una adaptación para su utilización por parte de usuarios individuales de los tests cuando actúan como profesionales, sigue siendo importante que tengan muy clara su propia política y que puedan comunicarla a otros.

Los objetivos de una política sobre el uso de los tests serían:

✔ Asegurar que se alcanzan las metas personales y organizacionales
✔ Evitar el posible uso inadecuado de los tests
✔ Demostrar explícitamente el compromiso con una práctica adecuada
✔ Asegurar que el uso de los tests se ajusta a los fines establecidos
✔ Asegurar que los tests se utilizan de forma no discriminativa
✔ Asegurar que las evaluaciones se basan en información comprensiva y pertinente
✔ Asegurar que los tests son utilizados por personas cualificadas para ello

Una política sobre el uso de los tests deberá abordar los siguientes aspectos.

✔ Uso adecuado de los tests
✔ Seguridad de los materiales y puntuaciones
✔ Quién puede aplicar, puntuar e interpretar los resultados
✔ Cualificaciones de quienes van a usar los tests
✔ Entrenamiento de los usuarios
✔ Preparación de las personas a evaluar
✔ Acceso a los materiales y seguridad de los tests
✔ Acceso a los resultados y confidencialidad de las puntuaciones
✔ Información sobre los resultados a las personas evaluadas
✔ Responsabilidades hacia los evaluados, antes, durante y después de la sesión de tests
✔ Responsabilidades individuales de los usuarios de los tests

Cualquier política que se establezca sobre tests tiene que ser revisada periódicamente, actualizándola de acuerdo con los avances producidos en los tests o en su utilización. Hay que permitir el acceso e informar sobre la política establecida a las partes implicadas en la utilización de los tests. La responsabilidad de la política de una organización debe recaer en un usuario cualificado, que tenga autoridad para asegurar su implementación y cumplimiento.

APÉNDICE B
DIRECTRICES PARA DESARROLLAR CONTRATOS ENTRE LAS PARTES IMPLICADAS EN LA EVALUACIÓN

Los contratos entre los usuarios de los tests y las personas evaluadas deben de estar de acuerdo con los criterios de una práctica adecuada, la legislación correspondiente y la política establecida sobre el uso de los tests. Los puntos que siguen se ofrecen como un ejemplo del tipo de aspectos que un contrato de este tipo debe abarcar. Los detalles variarán en función del contexto de la evaluación (ocupacional, educativo, clínico, jurídico, etc.), y de las leyes y regulaciones locales y nacionales.

Los contratos entre los usuarios, las personas evaluadas y otras partes implicadas aparecen con frecuencia implícitos, no formalizados explícitamente, al menos en parte. Dejando claros los objetivos, el papel y las responsabilidades de cada una de las partes ayuda a evitar los malentendidos, los perjuicios y la necesidad de acudir a los tribunales.

El usuario se compromete a:

1. Informar a las personas evaluadas de sus derechos en cuanto a cómo se utilizarán sus puntuaciones de los tests, así como sus derechos de acceso a ellas
2. Advertir con anterioridad suficiente sobre cualquier carga económica que conlleve el proceso de evaluación, quién es el responsable de los pagos y fechas en las que deben de hacerse efectivos
3. Tratar a las personas evaluadas con cortesía, imparcialidad y respeto, independientemente de su raza, género, edad, discapacidad, etc.
4. Utilizar tests de calidad probada, adecuados a las personas evaluadas y a la situación de evaluación
5. Informar antes de la aplicación de los tests sobre la finalidad de la evaluación, el tipo de test, a quién se enviarán los resultados y el uso que hará de ellos
6. Avisar con antelación donde se aplicarán los tests, cuando se darán los resultados, y si pueden o no obtener una copia del test una vez hecho, de su hoja de respuestas, o de sus puntuaciones
7. Utilizar personas cualificadas para la aplicación de los tests y la interpretación de los resultados
8. Asegurarse de que las personas a evaluar saben si el test es opcional o no, y, cuándo lo sea, asegurarse de que conocen las consecuencias de hacerlo o no hacerlo
9. Asegurarse de que las personas que van a hacer el test conocen las condiciones (si las hubiera) en las que podrían volver a hacer el test, ser calificados de nuevo, o anulárseles la prueba
10. Asegurarse de que las personas evaluadas saben que se les explicarán los resultados después de hacer el test tan pronto como sea posible y en términos que les resulten comprensibles
11. Asegurarse de que las personas evaluadas saben que sus resultados son confidenciales hasta el punto que permitan la ley y una práctica profesional adecuada
12. Informar a las personas evaluadas sobre quienes tendrán acceso a los resultados y las condiciones bajo las cuales se proporcionarán
13. Asegurarse de que las personas evaluadas conocen los procedimientos para elevar quejas o comunicar algún problema

El usuario comunicará a las personas evaluadas que se espera de ellas:

14. Un trato cortés y respetuoso con el resto de las personas durante el proceso de evaluación
15. Preguntar antes del test si no están seguras acerca de por qué se aplica el test, cómo se aplicará, qué tienen que hacer exactamente, y que se hará con los resultados
16. Informar sobre cualquier circunstancia que consideren que puede invalidar los resultados del test, o que deseen que se tenga en cuenta
17. Seguir las instrucciones de la persona que aplica el test
18. Ser consciente de las consecuencias que se siguen de no hacer el test, si deciden no hacerlo, y estar preparado para aceptar dichas consecuencias
19. Si se requiere abonar alguna cantidad para hacer la prueba, los pagos se harán en la fecha acordada

APÉNDICE C
APLICACIÓN DE LOS TESTS A PERSONAS CON ALGUNA DISCAPACIDAD

Cuando se modifica la forma de aplicación de un test para aplicarlo personas con alguna discapacidad hay que proceder con sumo cuidado y experiencia. Como siempre, si existe alguna legislación local o nacional específica debe respetarse, así como el derecho a la privacidad de la persona evaluada. Cuando se recoge la información sobre los tipos y niveles de discapacidad hay que limitarse a la información relativa a la capacidad de cada persona para llevar a cabo las tareas requeridas por el test. Hay que poner especial cuidado cuando los tests se utilizan en el ámbito ocupacional, bien sea para selección, promoción, u otros fines.

No hay una regla sencilla para aplicar un test correctamente que sea generalizable a todas las personas con algún tipo de discapacidad. Queda a juicio del profesional si es mejor utilizar una forma alternativa de evaluación o modificar el test, o su forma de aplicación. En la práctica no suele ser posible baremar los tests modificados en muestras amplias de personas con discapacidades equivalentes, de modo que se pueda asegurar la comparabilidad del test con la versión estándar. No obstante, cuando haya datos al respecto, por ejemplo, en el caso de los efectos producidos por la modificación del tiempo de aplicación, el uso de Braille, o versiones auditivas de los tests en casete, deben de utilizarse como guía para las modificaciones que se lleven a cabo. Aunque una estricta estandarización de la versión modificada tal vez no sea posible, siempre que lo sea deben de llevarse a cabo estudios piloto con muestras pequeñas.

Dada la escasez de información sobre el rendimiento en los tests (modificados o no) de las personas con alguna discapacidad, suele ser más adecuado utilizar los resultados de forma más cualitativa. Puede tomarse como una indicación de la característica evaluada, complementándola con información recogida por otros métodos.

En el caso de aplicaciones individuales el evaluador suele poder ajustar los procedimientos evaluativos a las capacidades de la persona evaluada. Sin embargo, en el caso de aplicaciones colectivas aparecen problemas especiales, como ocurre por ejemplo en selección de personal. En estos casos puede haber dificultades prácticas para modificar la forma de aplicación para determinadas personas dentro del marco de una aplicación colectiva. Además, los tratamientos diferenciados podrían ser vistos por algunos como discriminatorios. Por ejemplo, si se concede más tiempo para hacer el test, las personas discapacitadas podrían tener la sensación de que son tratadas de forma *diferente*, mientras que quienes no tienen ninguna discapacidad pueden pensar que el tiempo extra concedido proporciona a los otros una ventaja.

Consejos sobre necesidades especiales pueden obtenerse habitualmente de las organizaciones de discapacitados, así como de las propias personas con discapacidades a las que se van aplicar los tests. Suele ser de gran ayuda dirigirse directamente a las personas y preguntarles si hay alguna circunstancia que consideran que debe de tenerse en cuenta. En muchos casos esta consulta va a permitir llevar a cabo las modificaciones oportunas en el contexto de aplicación de la prueba, sin tener que modificar el propio test.

Las siguientes reglas generales pueden utilizarse como guía a la hora de decidir si se modifica el proceso evaluativo y cómo hacerlo.

1. Si la discapacidad no afecta al rendimiento en el test, no es necesario hacer ajustes en la prueba
2. Cuando la discapacidad influye en el rendimiento en el test hay que distinguir si esta influencia es incidental o forma parte del constructo medido. Si es incidental hay que modificar el test, pero si forma parte del constructo medido no. Por ejemplo, una persona con artritis en las manos tendrá problemas con un test de velocidad que conlleve escribir. Si la capacidad para escribir rápidamente formase parte del constructo medido, entonces no debería de cambiarse el test. Sin embargo, si la finalidad fuese, por ejemplo, medir la rapidez de percepción visual, habría que buscar una forma de respuesta a la prueba más adecuada, pues la artritis estaría influyendo negativamente en el constructo medido.
3. Cuando una discapacidad ajena al constructo medido influye en el rendimiento en el test, deben de llevarse a cabo los ajustes convenientes en la prueba
4. Los usuarios deben de consultar siempre el manual del test y los editores para buscar información sobre las modificaciones de las pruebas y sobre posibles formatos y procedimientos alternativos
5. Los usuarios de tests deben de consultar a las organizaciones correspondientes de discapacitados acerca de las implicaciones de una discapacidad concreta, la documentación sobre la discapacidad, y el tipo de modificaciones que podrían ser convenientes
6. Cualquier tipo de modificación que se haga en el test o en el proceso de aplicación debe de estar rigurosamente documentada, exponiendo claramente las razones que justifican la modificación

DOCUMENTACIÓN

Inglés

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington DC: American Psychological Association.

Bartram, D. (1995). The Development of Standards for the Use of Psychological Tests in Occupational Settings: The Competence Approach. *The Psychologist*, May, 219-223.

Bartram, D. (1996). Test Qualifications and Test Use in the UK: The Competence Approach. *European Journal of Psychological Assessment*, 12, 62-71.

Canadian Psychological Association. (1987). *Guidelines for Educational and Psychological Testing*. Ottawa: Canadian Psychological Association.

Eyde, L. D., Moreland, K. L. & Robertson, G. J. (1988). *Test User Qualifications: A Data-based Approach to Promoting Good Test Use*. Report for the Test User Qualifications Working Group. Washington DC: American Psychological Association.

Eyde, L. D., Robertson, G. J., Krug, S. E. et al (1993). *Responsible Test Use: Case Studies For Assessing Human Behaviour*. Washington DC: American Psychological Association.

Fremer, J., Diamond, E. E. & Camara, W. J. (1989). Developing a Code of Fair Testing Practices in Education. *American Psychologist*, 44, 1062-1067.

Hambleton, R. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229-244.

Joint Committee on Testing Practices. (1988). *Code of Fair Testing Practices in Education*. Washington DC: Joint Committee on Testing Practices.

Kendall, I., Jenkinson, J., De Lemos, M. & Clancy, D. (1997). *Supplement to Guidelines for the use of Psychological Tests*. Australian Psychological Society.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S. & Most, R. B. (1995). Assessment of Test User Qualifications: A Research-Based Measurement Procedure. *American Psychologist*, 50, 14-23.

Schafer, W. D. (1992). *Responsibilities of Users of Standardized Tests: RUST Statement Revised*. Alexandria, VA: American Association for Counseling and Development.

Van de Vijver, F. & Hambleton, R. (1996). Translating tests: some practical guidelines. *European Psychologist , 1*, 89-99.

Castellano

Colegio Oficial de Psicólogos (1987). *Código Deontológico*. Madrid: Colegio Oficial de Psicólogos.

Colegio Oficial de Psicólogos (1999). Página web del COP, Comisión de Tests: http://www.cop.es/tests/

Fernández Ballesteros, R. (1993). Evaluación psicológica en sus contextos de aplicación. *Revista de Historia de la Psicología, 14*, 97-114.

Franca Tarragó, O. (1996). *Ética para psicólogos*. Bilbao: Desclée de Brouwer.

Muñiz, J. (1997). Aspectos éticos y deontológicos de la evaluación psicológica. En A. Cordero (Ed.), *La evaluación psicológica en el año 2000*. Madrid: TEA Ediciones.

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

Rodríguez Sutil, C. (1996). La ética de la devolución en el psicodiagnóstico clínico. *Papeles del Psicólogo, 66*, 91-94.

COMISIÓN DE TESTS DEL COLEGIO OFICIAL DE PSICÓLOGOS

✔ José Muñiz (Coordinador)
Universidad de Oviedo
✔ M. Teresa Anguera
Universidad de Barcelona
✔ Rocío Fernández Ballesteros
Universidad Autónoma de Madrid
✔ José Ramón Fernández Hermida
Universidad de Oviedo

- Manuel García Pérez
  ALBOR-COHS
- Miguel Martínez
  EOS
- Rosario Martínez Arias
  Universidad Complutense de Madrid
- Eduardo Montes
  COP
- Gerardo Prieto
  Universidad de Salamanca
- Carlos Rodríguez Sutil
  Universidad Complutense de Madrid
- Nicolás Seisdedos
  TEA

AGRADECIMIENTOS

Dave Bartram, director del proyecto, expresa su agradecimiento a las siguientes personas, por sus valiosas contribuciones en la elaboración de las directrices:

- Dusica Boben, Produktivnost, Eslovenia.
- Eugene Burke, BPS, Inglaterra.
- Wayne Camara, The College Board, USA.
- Jean-Louis Chabot, ANOP, Francia.
- Iain Coyne, University of Hull, Inglaterra.
- Riet Dekker, Swets and Zeitlinger, Holanda.
- Lorraine Eyde, US Office of Personnel Management, USA.
- Rocío Fernández Ballesteros, EAPA, España.
- Ian Florance, NFER-NELSON, Inglaterra.
- Cheryl Foxcroft, Test Commission of South Africa, Sudáfrica.
- John Fremer, The College Board, USA.
- Kathia Glabeke, Commissie Psychodiagnostiek, Bélgica.
- Ron Hambleton, University of Massachusetts at Amherst, USA.
- Karin Havenga, Test Commission of South Africa, Sudáfrica.
- Jurgen Hogrefe, Hogrefe & Huber Verlagsgruppe, Alemania.
- Ralf Horn, Swets and Zeitlinger, Alemania.
- Leif Ter Laak, Saville and Holdsworth Ltd, Inglaterra.
- Pat Lindley, British Psychological Society, Inglaterra.
- Reginald Lombard, Test Commission of South Africa, Sudáfrica.
- José Muñiz, Colegio Oficial de Psicólogos, España.
- Gill Nyfield, Saville & Holdsworth Ltd, Inglaterra.
- Torleiv Odland, Norsk Psykologforening, Noruega.
- Berit Sander, Danish Psychologists= Association, Dinamarca.
- Francois Stoll, Federation Suisse des Psychologues, Suiza.

Ficha 2.

Cuestionario para la Evaluación de los tests (versión modificada Junio 2013)

1. DESCRIPCIÓN GENERAL DEL TEST[1]

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[2]
(Identifique el área de contenido definido en la publicación. Si no hay una definición clara debe señalarlo en el apartado "Otros", e indicar cuál es el área de contenido más adecuada según la información proporcionada en el manual)

( ) Inteligencia
( ) Aptitudes
( ) Habilidades y Rendimiento académico
( ) Psicomotricidad
( ) Neuropsicología
( ) Personalidad
( ) Motivación
( ) Actitudes
( ) Intereses
( ) Escalas de Desarrollo
( ) Competencia Curricular
( ) Escalas Clínicas
( ) Potencial de Aprendizaje
( ) Calidad de vida/ Bienestar
( ) Estrés/burnout
( ) Estilos cognitivos
( ) Otros (Indique cuál:.....)

1.11. Breve descripción de la variable o variables que pretende medir el test:
(Se trata de hacer una descripción no evaluativa del test entre 200-600 palabras. La descripción debe de proporcionar al lector una idea clara del test, lo que pretende medir y las escalas que lo conforman)

1.12. Área de aplicación[3]
(Identifique el área o áreas de aplicación definidas en la publicación. Si no hay una definición clara debe señalarlo en el apartado "Otros" e indicar cuál es el área de aplicación más adecuada según la información proporcionada en el manual)

[1] Si el test está compuesto de subtests heterogéneos en su formato y características, rellene un cuestionario para cada subtest.
[2] Puede marcar más de una opción.
[3] Puede marcar más de una opción.

( ) Psicología clínica
( ) Psicología educativa
( ) Neuropsicología
( ) Psicología forense
( ) Psicología del trabajo y las organizaciones
( ) Psicología del deporte
( ) Servicios sociales
( ) Salud general y bienestar
( ) Psicología del Tráfico
( ) Otros (Indique cuál:.....)

1.13. Formato de los ítems[4]
( ) Respuesta libre
( ) Respuesta dicotómica (si/no, verdadero/falso, etc)
( ) Elección múltiple
( ) Tipo Likert
( ) Adjetivos bipolares
( ) Otro (Indique cuál:.....)

1.14. Número de ítems[5]

1.15. Soporte[6]
( ) Administración oral
( ) Papel y lápiz
( ) Manipulativo
( ) Informatizado
( ) Otro (Indique cuál:.....)

1.16. Cualificación requerida para el uso del test de acuerdo con la documentación aportada:
( ) Ninguna
( ) Entrenamiento y Acreditación especifica*
( ) Nivel A[7]
( ) Nivel B
( ) Nivel C
( ) Otra (Indique cuál:.....)
**Indique el nombre de la institución que lleva a cabo la acreditación:*

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.)

[4] Puede marcar más de una opción.
[5] Si el test tiene varias escalas, indique el número de ítems de cada una.
[6] Puede marcar más de una opción.
[7] Algunos países han adoptado sistemas para la clasificación de los tests en distintas categorías, en función de la cualificación requerida por los usuarios. Estos sistemas de clasificación proporcionan a los editores de tests un medio para decidir a quién pueden vender los tests. Un sistema muy utilizado es el que divide los tests en tres categorías: Nivel A (tests de rendimiento y conocimientos), Nivel B (tests colectivos de aptitudes e inteligencia) y Nivel C (tests de aplicación individual de inteligencia, personalidad y otros instrumentos complejos).

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*.

1.19. Procedimiento de corrección:
( ) Manual mediante plantilla
( ) Lectora óptica
( ) Automatizada por ordenador
( ) Efectuado exclusivamente por la empresa suministradora
( ) Mediante expertos
( ) Hoja Autocorregible
( ) Otro (Indique cuál:.....).

1.20. Puntuaciones: (Describa el procedimiento para obtener las puntuaciones directas, totales o parciales, corrección de la probabilidad de responder correctamente por azar, etc)

1.21. Transformación de las puntuaciones:
( ) Característica no aplicable para este instrumento
( ) Normalizada (puntuaciones obtenidas mediante normalización aplicada mediante la tabla de la curva normal)
( ) No normalizada (puntuaciones estandarizadas obtenidas mediante transformaciones lineales)

1.22. Escalas utilizadas:
( ) Puntuaciones basadas en percentiles
( ) centiles
( ) quintiles
( ) deciles
( ) Puntuaciones estandarizadas
( ) Puntuaciones típicas
( ) Eneatipos
( ) Decatipos
( ) T (*Media 50 y desviación típica 10)*
( ) S (*Media 50 y desviación típica 20)*
( ) Cocientes de desviación
( ) Puntuaciones directas solamente
( ) Otra (Indique cuál:.....)

1.23. Posibilidad de obtener informes automatizados:
( ) No
( ) Si*

**En caso afirmativo haga una breve descripción no evaluativa del Informe Automatizado, en la que se hagan constar las características fundamentales, tales como tipo de informe, estructura, claridad, estilo, tono, etc.*

1.24. El editor ofrece un servicio para la corrección y/o elaboración de informes:
( ) No
( ) Si

1.25. Tiempo estimado para la aplicación del test (instrucciones, ejemplos y respuestas a los ítems).
En aplicación individual:.....
En aplicación colectiva:.....

1.26. Documentación aportada por el editor:
( ) Manual
( ) Libros o artículos complementarios
( ) Disketes/CD
( ) Información técnica complementaria y actualizaciones
( ) Otra (Indique cuál:.....)

1.27. Precio de un juego completo de la prueba (documentación, test, plantillas de corrección; en el caso de tests informatizados no se incluye el costo del *hardware*):

1.28. Precio y número de ejemplares del paquete de cuadernillos (tests de papel y lápiz):

1.29. Precio y número de ejemplares del paquete de hojas de respuesta (tests de papel y lápiz):

1.30. Precio de la corrección y/o elaboración de informes por parte del editor:

1.31. Bibliografía básica acerca del test aportada en la documentación:

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, etc.)

2.2. Calidad de la documentación aportada:
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción muy clara y completa de las características técnicas, fundamentada en abundantes datos y referencias)

2.3. Fundamentación teórica:
( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción muy clara y documentada del constructo que se pretende medir y del procedimiento de medición)

2.4. Adaptación del test (si el test ha sido traducido y adaptado para su aplicación en España):
( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena

***** ( ) Excelente (Descripción precisa del procedimiento de traducción, de la adaptación de los ítems a la cultura española, de los estudios de equivalencia con la versión original, utilización de la normativa de la International Test Commission, etc.).

2.5. Calidad de las instrucciones:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Claras y precisas. Muy adecuadas para las poblaciones a las que va dirigido el test.

2.6. Facilidad para comprender la tarea:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Suficiente
**** ( ) Buena
***** ( ) Excelente (Los sujetos de las poblaciones a las que va dirigido el test pueden comprender fácilmente la tarea a realizar a partir de las instrucciones proporcionadas).

2.7. Facilidad para registrar las respuestas:

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

2.8. Calidad de los ítems (aspectos formales):

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (La redacción y el diseño son muy apropiados)

2.9. Análisis de los ítems

2.9.1 Datos sobre el análisis de los ítems:

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

2.10. Validez

2.10.1. Validez de contenido[8]

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

2.10.1.1. Calidad de la representación del contenido o dominio:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (En la documentación se presenta una precisa definición del contenido. Los ítems muestrean adecuadamente todas las facetas del contenido)

2.10.1.2. Consultas a expertos[9]

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

2.10.2. Validez de constructo:

2.10.2.1. Diseños empleados[10]

( ) No se aporta información en la documentación
( ) Correlaciones con otros tests
( ) Diferencias entre grupos
( ) Matriz multirasgo-multimétodo
( ) Análisis factorial exploratorio
( ) Análisis factorial confirmatorio
( ) análisis de invarianza/funcionamiento diferencial de items
( ) Diseños experimentales o cuasi experimentales
( ) Otros (Indique cuales:.....).

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

**Describa brevemente el procedimiento de selección.*

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

[10] Puede marcar más de una opción.

2.10.2.4. Mediana de las correlaciones del test con otros tests similares:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.25)
** ( ) Adecuada pero con algunas carencias (0.25≤r<0.40)
*** ( ) Adecuada (0.40≤r<0.50)
**** ( ) Buena (0.50≤r<0.60)
***** ( ) Excelente (r≥0.60)

2.10.2.5. Calidad de los tests empleados como criterio o marcador:

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente

2.10.2.6. Resultados de las diferencias entre grupos (pueden ser grupos naturales o experimentales).

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (se observan diferencias significativas en el sentido esperado)

2.10.2.7. Resultados del análisis de la matriz multirrasgo-multimétodo

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (los resultados apoyan tanto la validez convergente como discriminante)

2.10.2.8. Resultados del análisis factorial

( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (los resultados apoyan la dimensionalidad establecida)

2.10.2.9. Datos sobre el sesgo/funcionamiento diferencial de los ítems:

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuados
** ( ) Adecuados pero con algunas carencias
*** ( ) Adecuados
**** ( ) Buenos
***** ( ) Excelentes (Información detallada sobre diversos estudios acerca del sesgo de los ítems relacionado con el sexo, la lengua materna, etc. Empleo de la metodología apropiada)

2.10.3. Validez predictiva

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

2.10.3.2. Diseño de selección del criterio[11]

( ) Concurrente
( ) Predictivo
( ) Retrospectivo

2.10.3.3. Tamaño de las muestras en la validación predictiva:

( ) 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 y representativa (N≥200)
**** ( ) Varios estudios con muestras representativas de tamaño moderado
****** ( ) Varios estudios con muestras grandes y representativas

2.10.3.4. Procedimiento de selección de las muestras*:

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

*Describa brevemente el procedimiento de selección.

2.10.3.5. Mediana de las correlaciones del test con los criterios:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.20)
** ( ) Suficiente (0.20≤r<0.35)
*** ( ) Buena (0.35≤r<0.45)
**** ( ) Muy buena (0.45≤r<0.55)
***** ( ) Excelente (r≥0.55)

2.10.4. Comentarios sobre la validez en general:

2.11. Fiabilidad

2.11.1. Datos aportados sobre la fiabilidad:

( ) Un único coeficiente de fiabilidad
( ) Un único error típico de medida
( ) Coeficientes de fiabilidad para diferentes grupos de sujetos
( ) Error típico de medida para diferentes grupos de sujetos
( ) Cuantificación del error mediante TRI (Función de información u otros)

2.11.2. Equivalencia (Formas paralelas) (mismas medias, varianzas y correlaciones con otros tests):
2.11.2.1. Tamaño de las muestras en los estudios de equivalencia:

( ) 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

[11] Puede marcar más de una opción.

2.11.2.2. Mediana de los coeficientes de equivalencia:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.50)
** ( ) Adecuada pero con algunas carencias (0.50≤r<0.60)
*** ( ) Adecuada (0.60≤r<0.70)
**** ( ) Buena (0.70≤r<0.80)
***** ( ) Excelente (r≥0.80)

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

2.11.3.2. Coeficientes de consistencia interna presentados

( ) no se aporta información
( ) coeficiente alpha o KR-20
( ) lambda-2
( ) omega (análisis factorial)
( ) theta (análisis factorial)

2.11.3.3. Mediana de los coeficientes de consistencia:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.60)
** ( ) Adecuada pero con algunas carencias (0.60≤r<0.70)
*** ( ) Adecuada (0.70≤r<0.80)
**** ( ) Buena (0.80≤r<0.85)
***** ( ) Excelente (r≥0.85)

2.11.4. Estabilidad (Test-Retest)

2.11.4.1. Tamaño de las muestras en los estudios de estabilidad[12]

( ) 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

2.11.4.2. Mediana de los coeficientes de estabilidad:

( ) No se aporta información en la documentación
* ( ) Inadecuada (r<0.55)
** ( ) Adecuada pero con algunas carencias (0.55≤r<0.65)
*** ( ) Adecuada (0.65≤r<0.75)
**** ( ) Buena (0.75≤r<0.80)
***** ( ) Excelente (r≥0.80)

[12] Número de sujetos con ambas puntuaciones (antes-después).

2.1.5 Cuantificación de la precisión mediante TRI
2.11.5.1. Tamaño de las muestras en los estudios de TRI
Depende del formato de los items y del modelo empleado. Unas recomendaciones generales sobre el tamaño adecuado son 200 sujetos para el modelo de 1 parámetro, 400 para el modelo de dos parámetros y 700 para el de 3 (Parshall, Davey, Spray, & Kalohn, 2001).

( ) No se aporta información en la documentación
* ( ) Un estudio con una muestra pequeña
** ( ) Un estudio con una muestra adecuada
*** ( ) Un estudio con una muestra grande
**** ( ) Varios estudios con muestras de tamaño moderado
***** ( ) Varios estudios con muestras grandes

2.11. 5 Comentarios sobre la fiabilidad en general:

2.12. Baremos
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[13]

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

2.12.3. Procedimiento de selección de las muestras*:

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

*Describa brevemente el procedimiento de selección.

2.12.4. Actualización de los baremos

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

2.12.4. Comentarios sobre los baremos

[13] Si hay varios baremos, clasifique el tamaño promedio

3. VALORACIÓN GLOBAL DEL TEST

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

3.2.A modo de resumen, rellene las Tablas 1 y 2.
La Tabla 1 incluye algunos datos descriptivos del test.

TABLA 1
DESCRIPCIÓN DEL TEST

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

En la Tabla 2 se resume la valoración de las características generales del test. Tome en consideración el promedio de las calificaciones emitidas en los apartados que figuran en la segunda columna de la Tabla 2.

TABLA 2
VALORACIÓN DEL TEST

Característica	Apartados	Valoración
Materiales y documentación	2.1 y 2.2	
Fundamentación teórica	2.3	
Adaptación	2.4	
Análisis de ítems	2.9	
Validez de contenido	2.10.1	
Validez de constructo	2.10.2	
Análisis del sesgo	2.10.2.	
Validez predictiva	2.10.3	
Fiabilidad: equivalencia	2.11.2	
Fiabilidad: consistencia interna	2.11.3	
Fiabilidad: estabilidad	2.11.4	
Baremos	2.12	