Papeles del Psicólogo, 2012. Vol. 33(1), pp. 12-21
http://www.cop.es/papeles

TESTS PUBLICADOS EN ESPAÑA: USOS, COSTUMBRES Y ASIGNATURAS PENDIENTES

TESTS PUBLISHED IN SPAIN: USES, CUSTOMS AND PENDING MATTERS

Paula Elosua
Universidad del País Vasco

La psicometría del siglo XXI asume entre sus tareas el desarrollo de modelos formales y el estudio y salvaguarda de las condiciones que garantizan un uso adecuado de los tests. Los progresos en ambas direcciones se conjugan en las directrices redactadas por organizaciones nacionales e internacionales que intentan acercar los últimos avances y reflexiones al profesional. Sin embargo, constatamos la distancia que separa ambos mundos. En este trabajo analizamos la práctica profesional tal y como está reflejada en los manuales de los tests más utilizados en España. Abordamos el tratamiento de la fiabilidad, validez, interpretación de puntuaciones o adaptación tomando como criterio de fuerza las directrices conjuntas de la APA y las directrices de la Comisión Internacional de Tests. Los resultados muestran las lagunas entre los usos y los deberes. A lo largo del trabajo intentamos profundizar en algunas de las razones que puedan explicar esta brecha.
Palabras clave: *Tests, Usos, Directrices, Psicometría.*

The tasks takes up by the XXIst century psychometric include the development of formal models and the study of the conditions that guarantee an appropriate use of the tests. The progresses in both directions are linked by the guidelines edited by national and international organizations that try to bring the last advances and reflections over to the applied professional. Nevertheless, we state the distance that separates both worlds. In this work we analyze the professional practice as it is reflected in the manuals of the tests most used in Spain. We tackle the treatment of the reliability, validity, interpretation of scores or adaptation, taking as a criterion of force the joint guidelines of the APA and the guidelines of the International Test Commission. The results show the gap between the uses and the duties. Along the work we try to study in depth some of the reasons that could explain this breach.
Key words: *Tests, Uses, Standards, Psychometrics.*

USOS Y COSTUMBRES

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 y revisión de los usos relacionados con los tests. Los proyectos amparados por estas organizaciones incluyen estudios sobre las actitudes de los psicólogos hacia los tests (Evers y col., 2011, Muñiz y Fernández-Hermida, 2010), análisis sobre las condiciones que favorecen una correcta evaluación (Muñiz, 2010) o la actualización de las versiones nacionales de los cuestionarios para la revisión de los manuales de tests publicados en Europa (Bartram, 2011; Muñiz, y col., 2011). Las conclusiones emanadas de esta reflexión sobre nuestra praxis permiten dibujar un panorama de situación sobre "usos y costumbres", que es fundamental y previo a cualquier propuesta de mejora con relación a los "deberes".

Las estrategias encaminadas a mejorar el uso de los tests defendidas en España (Muñiz, 2010), abogan por la formación y por la restricción. Los objetivos fijados dentro de la estrategia formativa incluyen la formación del profesional, la diseminación de información sobre la calidad y características de los tests, y el desarrollo de directrices. La estrategia restrictiva limita el uso de tests a personal cualificado. Centrándonos en la primera, comprobamos 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) el conocimiento psicométrico avanza de tal forma que la distancia entre psicometría teórica y práctica aplicada es hoy mayor que nunca, y c) 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 aplican, a pesar de su papel motor sobre la mejora en los usos.

Correspondencia: Paula Elosua. *Universidad del País Vasco. Avda. Tolosa, 70. 20018. San Sebastián. España.*
E-mail:Paula.elosua@ehu.es

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia e Innovación (PSI2011-30256) y por la Universidad del País Vasco (UPV/EHU -GIU 09/22).

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 *Psicothema*. Es innegable, la distancia entre los contenidos de ambas, sin embargo el conocimiento psicométrico está presente y nutre a las dos. Las aportaciones de la psicometría teórica no llegan, en muchas ocasiones, al psicólogo aplicado, más interesado por naturaleza, en cuestiones prácticas y sustantivas alejadas de problemas formales. Una revisión de algunos de los tests más utilizados en España (Muñiz y col., 2011) deja patente que la psicometría aplicada se construye sobre conceptos y usos bien asentados, pero tal vez algo "arcaicos" desde la perspectiva de los avances en teoría psicométrica. Las prácticas más comunes (la utilización del coeficiente alfa de Cronbach como estimador de la fiabilidad, el empleo de la técnica del análisis factorial exploratorio como evidencia de validez interna, la estimación de las relaciones entre el test y medidas convergentes por medio del coeficiente de correlación de Pearson, o la transformación de puntuaciones para la construcción de baremos) se asientan sobre técnicas y procedimientos desarrollados en las primeras décadas del siglo XX (Spearman, 1904; Thurstone, 1932). Su uso generalizado por los profesionales ha exigido la conjunción de dos elementos: la formación, y la disposición de herramientas sencillas de utilizar. El primero queda cubierto por la formación de grado, que de forma universal ofrece conocimientos psicométricos básicos a los estudiantes de psicología. El segundo, está ligado al desarrollo de software amigable que integra módulos de análisis necesarios en el proceso de construcción de tests.

Sin embargo, en los últimos 100 años la psicometría ha avanzado, y lo ha hecho en una doble dirección. Se han desarrollado modelos teóricos más potentes que la Teoría Clásica de Tests (Teoria de Respuesta al ítem; Hambleton y Swaninathan, 1985), y se ha profundizado en la importancia del uso de los tests en la salvaguarda de sus propiedades. Hoy somos más conscientes que nunca de las consecuencias derivadas de un uso incorrecto de los tests (Messick, 1995).

Las asociaciones internacionales relacionadas con el uso de los tests asumen la labor del diseño de planes de formación y planes de actuación para acercar ambos mundos. Este esfuerzo se traduce en la organización de congresos, cursos de formación, y sobre todo, en la elaboración de directrices profesionales y técnicas. Las directrices para el uso de los tests de la AERA, APA y NCME (1999), las directrices para la evaluación en contextos laborales y de organizaciones (EFPA), o las directrices de la Comisión Internacional de Tests relacionadas con la adaptación de tests, 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 se recomiendan principios éticos que garantizan un uso correcto de los tests. Sin embargo, su contenido no siempre se ve reflejado en la práctica profesional.

OBJETIVOS

En este contexto de revisión de usos y costumbres el objetivo de este trabajo es analizar la práctica tal y como está reflejada en la documentación que aportan los tests publicados en España, y contraponerla con algunos de los desarrollos que sin ser de última generación, han tenido un impacto mayor en la teoría/practica psicométricas. Para ello, y con una finalidad descriptiva y formativa, se ha analizado la documentación de los tests más utilizados en España (Muñiz y Fernández-Hermida, 2010), y se ha contrastado con un modelo bien establecido y aceptado; las directrices para el uso de tests publicadas por la AERA, APA y NCME (1999) cuya última revisión está en periodo de discusión. Se abordan los apartados de fiabilidad, validez, normas, administración y adaptación desde una perspectiva conceptual y una aproximación metodológica en la que se refieren las directrices más relevantes y se comprueba el modo de hacerlas operativas.

Fiabilidad. Las directrices definen la fiabilidad en términos de consistencia y en términos de error; de hecho, el título genérico del apartado que aborda este punto es "Fiabilidad y Errores de Medida" (Directriz 2.1. *"For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors and standard errors of measurement or test information functions should be reported "*). Esta dualidad no es la norma en los manuales analizados.

Desde la aparición del artículo 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 inter-

pretar (Nunnally, 1978) y no es complicado mejorar su valor. Prácticamente todos los manuales analizados aportan información sobre este indicador, que se ha convertido en el estadístico más conocido de la Teoría Clásica de Tests.

Error de medida. Sin embargo, 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 o *Se*) 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 *Se* podrían construirse enunciados como: "Con un 95% de probabilidad la puntuación de la persona X se sitúa entre los valores 34 y 48".

La necesidad de aportar información sobre la consistencia interna y el error de medida afecta a todas y a cada una de las escalas parciales que miden áreas o aspectos bien diferenciados de la conducta dentro de un test (Elosua, 2008).

La importancia de informar sobre la (in)certidumbre de una puntuación aumenta en las situaciones que exigen el establecimiento de puntos de corte en la interpretación. A pesar de que uno de los principios básicos de la teoría clásica de tests asume la constancia del error típico de medida a lo largo del continuo de puntuaciones, su cumplimiento es violado de forma habitual (Hambleton y Swaminathan, 1985). Si existen indicios de tal violación, y además el test ofrece intervalos de puntuación con criterios diagnósticos o de selección, es importante estimar el efecto del error de medida alrededor de las puntuaciones críticas. De igual manera, si en el test se aportan criterios de interpretación diferentes en función del sexo, edad u otra variable relevante con relación al test, situación común en la práctica profesional , sería conveniente estimar el error típico de medida de modo diferenciado para cada uno de los grupos considerados.

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 al uso no ofrecen información sobre este índice entre sus estandarizadas salidas.

Consistencia desde un modelo. A pesar del extendido uso del coeficiente alfa como estimador de la fiabilidad, un número cada vez mayor de psicómetras lo desaconseja, y propone alternativas de estimación y de definición de la fiabilidad construídas sobre modelos alternativos de medida a la teoría clásica de tests (McDonald , 1981,1999; Jöreskog, 1971; Raykov, 2001).

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 la perspectiva factorial ayudaría a eliminar a) el uso incorrecto del coeficiente alfa como indicador de unidimensionalidad (Hattie, 1985), 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 (Zumbo, Gadermann y Zeisser, 2007), y c) permitirían profundizar en el estudio de la estructura interna del test. Son varios los procedimientos disponibles para el estudio de la homogeneidad de las medidas desde los modelos de ecuaciones estructurales para cuya estimación puede utilizarse software como Mplus (Muthén y Muthén, 2001), FACTOR (Lorenzo-Seva y Ferrando, 2007), LISREL (Jöreskog y Sörbom, 1996), EQS (Bentler, 1995), AMOS, o R (Ihaka y Gentleman, 1996; ver códigos en Elosua y Zumbo, 2007).

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 ellos 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.

Validez. Si hay un punto sobre el que no hay discusión entre teóricos y profesionales es el referente a la importancia de la validez y del proceso de validación de las puntuaciones en la construcción y uso de los tests. *"Validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests"* (APA, AERA y NCME, 1999, pág. 9).

Desde una perspectiva histórica la definición y evolución del concepto de validez ha quedado reflejada en las sucesivas publicaciones de las directrices conjuntas de la APA (1954, 1966, 1974, 1985, 1999; ver Elosua, 2003). La diferenciación entre validez de constructo, validez predictiva y validez de contenido ha impregnado las 4 primeras ediciones. La edición de 1985 defiende por primera vez una concepción unitaria de la validez, aunque diferencia entre tres tipos de evidencia. En la edición de 1999 se define la validez como concepto unitario, se postulan cinco fuentes de evidencia, 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 directrices aconsejan la utilización de cinco fuentes de evidencias: evidencias de contenido, evidencias basadas en la estructura interna, evidencias basadas en las relaciones con otras variables, evidencias sobre el proceso de respuesta y evidencias basadas en las consecuencias del test.

La nueva teoría sobre la validez (validación) sitúa el foco de atención sobre la validación de la interpretación propuesta; no hablamos ya de validez del test y en este contexto carece de sentido hablar de validez de contenido, validez de criterio y validez de constructo. El objetivo es justificar una interpretación de las puntuaciones basada en razones, argumentos, que se recogen durante el proceso de validación (Kane, 1992, 2006). Esta idea, que está recogida en las directrices de 1999 *("validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the proposed use"*. APA, AERA y NCME, 1999, pág. 9), persigue justificar el uso de las puntuaciones como representación del constructo que se quiere medir, o cuando así proceda, justificar su utilidad en la predicción de conductas.

Las definiciones de validez ofrecidas en los años 30 que diferencian tres tipos de validez, son las más comúnmente adoptadas en los manuales analizados. Reflejan la idea de que un test es válido para aquello con lo que correlaciona (Kelley, 1927; Guilford, 1946), o equiparan la validez del test al grado en que el test mide lo que pretende medir. La operacionalización de la primera acepción es harto sencilla; basta estimar la correlación entre un test y un criterio; es decir basta estimar un coeficiente de validez. La segunda acepción más acorde o cercana a los modelos factoriales, se materializa habitualmente por medio del análisis factorial exploratorio. Aunque ambas definiciones han sido tachadas como incorrectas desde hace décadas (Anastasi, 1954; Rulon, 1946), los procedimientos ligados y derivados de ellas siguen vigentes en la práctica actual. Las técnicas factoriales y los estudios de correlación (regresión) están presentes en todos los manuales analizados. Son técnicas diseñadas en los albores de la psicometría (Spearman, 1904, Thurstone, 1932) que han influenciado las primeras definiciones de validez, están integradas en todos 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.

Sin embargo, 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 y de software amigable para su estimación (Bentler, 1980, 1986; Bollen, 1989; Jöreskog y Sörbom, 1996), que no aparece reflejado en los manuales de los tests analizados. 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. Los modelos de regresión y correlación no contemplan este hecho. La aplicabilidad de las técnicas SEM está bien documentada tanto para estudios correlacionales como para estudios de carácter experimental. Sus ventajas en los estudios cross-seccionales o longitudinales, entre los que cobran especial relevancia los diseños de curvas de crecimiento, es evidente, en tanto en cuanto favorecen e impulsan el rol de la teoría en la investigación aplicada y permiten contrastar y evaluar modelos explicativos alternativos (Kline, 2010; Millsap y Maydeu-Olivares, 2009). 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 simultaneo para diferentes grupos.

El software para la aplicación de modelos SEM es variado; LISREL (Jöreskog y Sörbom, 1996) , EQS (Bentler, 1995) AMOS, Mx (ahora integrado en R) o los paquetes *sem* y *lavaan* de R (Elosua, 2009; Ihaka y Gentleman, 1996).

Escalas, baremos y comparación entre puntuaciones. La transformación de las puntuaciones brutas en puntuaciones derivadas (puntuaciones típicas, percentiles, escalas de grado…) cuya finalidad es favorecer la interpretación, o la definición de puntos de corte que discriminan entre categorías diagnósticas, o niveles de rendimiento, y el establecimiento de criterios para la selección, ocupan un apartado específico en las directrices. En ellas se remarca la importancia de diferenciar entre interpretaciones normativas e interpretaciones criteriales de las puntuaciones. En el primer caso, que es el más habitual en los manuales analizados, las puntuaciones se leen con referencia a la distribución estadística de la muestra de baremación; una puntuación se interpreta en función de cómo ha realizado el grupo normativo el test, y para ello se utilizan transformaciones en escalas percentiles, puntuaciones z, puntuaciones T , escalas de grado u otro tipo de puntuación derivada. La interpretación criterial es diferente, y su utilización exige la definición de un referente externo respecto al cual se comparan los niveles de ejecución. En este sentido, es posible informar 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 distinción teórica (norma-criterio), no es mutuamente excluyente, pero su utilización conjunta ha de estar basada en una documentación que justifique tanto una como otra. Son varios los manuales que crean categorías sustantivas en base a distribuciones normativas, convirtiendo un percentil en una clase diagnóstica. Sin embargo, en este punto las directrices son claras (Directriz 4.9. *" When raw score or derived score scales are designed for criterion-referenced interpretation including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained"*. Directriz 4.19. *"When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be clearly documented"*.).

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). Esta práctica que es común en la evaluación educativa a gran escala o en los programas internacionales como PISA o TIMMS no está presente en los manuales de los tests analizados.

La lectura criterial de los resultados adquiere una relevancia mayor cuando el objetivo de la evaluación es un diagnostico diferencial. (Directriz 12.6 *"When differential diagnosis is needed the professional should choose, if possible, a test for which there is evidence of the tests's ability to distinguished between the two or more diagnostic groups of concern rather than merely to distinguished abnormal cases from the general population"*). En los contextos evaluativos los argumentos aportados en el proceso de validación habrían de recoger información sobre la plausibilidad de las inferencias con relación a un diagnóstico. No constituye un argumento válido la descripción de las medias aritméticas obtenidas en diferentes grupos. Los argumentos habrían de incluir intervalos de confianza, tamaños del efecto, o tablas que muestren el grado de superposición de la distribución entre muestras diagnósticas, análisis discriminantes, u otras técnicas derivadas de la minería de datos que estimen funciones de clasificación/predicción (Bully y Elosua, 2011).

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 de examen óptimas 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. Los métodos de acomodación que son comunes en la evaluación educativa todavía no están recogidos en nuestra praxis.

La importancia de una adecuada elaboración de los informes de evaluación está siendo reconocida de forma unánime por la comunidad psicométrica (Hattie, 2009; Hambleton y Zenisky, en prensa). La información sumativa, diagnostica 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. A este respecto recuerda la directriz 5.10. *". …the interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used."*

Adaptación de tests. La edición de las directrices conjuntas de la APA sobre el uso de tests no tiene un apartado dedicado a la adaptación de tests, carencia que es cubierta por las directrices elaboradas por la Comisión Internacional de Tests. Son un modelo inmejorable como criterio de referencia. Su importancia en el contexto que tratamos es clara si tenemos en cuenta que de los 10 tests más utilizados en Europa (Elosua y Iliescu, en prensa) 9 fueron construidos en lengua inglesa, y de los 25 tests más utilizados en España 17 son adaptaciones de versiones construías en otro idioma.

Las directrices para la adaptación de tests fueron publicadas por primera vez el año 1994 (Muñiz y Hambleton, 1996), y en su segunda edición (Bartram, Gregoire, Hambleton, y van de Vijver, 2011; Elosua y Hambleton, 2011) aportan un modelo en el cual se describen y operativizan de forma clara los puntos a considerar en la adaptación de tests. Las 20 directrices están estructuradas en 6 categorías con el fin de cubrir todas las etapas implicadas en el proceso de adaptación; directrices previas, directrices de desarrollo, directrices de confirmación, directrices de administración, directrices de puntuación, y directrices referidas a la documentación.

Una de las más importantes y que puede resumir el contenido de todas ellas recuerda la importancia de ofrecer información empírica sobre la equivalencia de constructo, equivalencia de método y equivalencia entre los ítems en todas las poblaciones a las que va destinado el test (segunda directriz de confirmación). Sintetiza la relevancia del análisis de la equivalencia entre el objeto de medida en distintas poblaciones y el modo en que este es medido. Las aproximaciones para su estudio pueden ser cualitativas (procedimientos de juicio, análisis de contenido, entrevistas) y cuantitativas. Desde esta última podrían adoptarse acercamientos exploratorios clásicos como el índice de congruencia de Tucker (1951), si bien es recomendable la utilización de modelos de invarianza derivados de los modelos de respuesta al ítem o de los modelos de ecuaciones estructurales (Elosua y Muñiz, 2010).

PROCEDIMIENTO

La última edición del estudio sobre las opiniones de los psicólogos llevado a cabo en España (Muñiz y Fernández-Hermida, 2010) ha revelado entre otros puntos, los tests más utilizados en la práctica profesional (tabla 1). En la relación de los 10 tests más utilizados por especialidad (Clínica, Educativa y Trabajo) comprobamos que de los 25 cuestionarios, 24 son instrumentos estandarizados de medición. De ellos 15 fueron construidos en Estados Unidos, 1 en el Reino Unido, 1 en Suiza, 1 en Francia, 1 en Italia, y 6 son de origen español. Estos datos reflejan que el 76% de los tests más usados en España son adaptaciones, y de ellas, 15 (79%) provienen de Estados Unidos.

Los 10 tests más usados en España con independencia de la especialidad profesional (ver tabla 1) son adaptaciones; de ellos 9 fueron construidos originalmente en inglés y sus primeras versiones fueron publicadas hace décadas. El test de matrices progresivas de RAVEN es el más antiguo; su primera edición se publicó en 1938. Es curioso constatar que el año medio de publicación de los *top ten* nos sitúa en el año 1960; lo cual es un indicador de la fortaleza y actualidad de estos tests. De ellos algunos están construidos sobre sólidos modelos formales. Tal es el caso de las escalas cognoscitivas (WAIS, WISC) o de las medidas de la personalidad de Cattell que se construyeron sobre modelos factoriales, bien de la inteligencia bien de la personalidad. Otras escalas tienen una naturaleza más ecléctica y surgen a partir de consideraciones prácticas o aplicadas, como el MMPI, SCL o BDI.

La clasificación de los tests en función de su dominio psicológico permite atestiguar que de los 25 tests, 12 son de naturaleza cognitiva (WAIS, WISC, Raven, BADYG, TALE, MSCA, PROLEC, BENDER, ITPA, TAMAI, DAT, IGF), 6 son cuestionarios de personalidad (16PF, NEO PI-R, PAPI, TPT, IPV, BFQ) y 6 tienen carácter clínico (MCMI, MMPI, SCL-90, BDI, STAI, MMSE).

De los 25 tests que componen el universo del estudio, se han analizado 22. El Rorschach se excluyó por su naturaleza proyectiva; el BDI es un cuestionario "fantasma" en el sentido de que si bien figura en la lista de los tests más utilizados en la práctica profesional, no existía cuando se elaboró este estudio una versión oficial de él en Español; finalmente no se estudió el PAPI por no poder acceder a él.

RESULTADOS

Fiabilidad. El tratamiento otorgado al tema de la fiabilidad por los manuales se corresponde mayoritariamente con una concepción clásica en la que se equipara la fiabilidad con la consistencia interna. En este orden, el coeficiente alfa de Cronbach es el indicador preferido. Un total de 15 manuales ofrecen esta información. La estabilidad temporal de las puntuaciones analizada por medio de un test-restest es abordada por un total de 6 manuales. Desafortunadamente la información sobre el error típico de medida no está presente en todos los manuales, sólo 5 manuales aportan esta información. Si se

considera la conveniencia de ofrecer información sobre los errores de medida condicionales a las puntuaciones o a las muestras de baremación, este número decrece. Uno de los manuales analizados ofrece información sobre el ETM condicional a la puntuación en el marco de la Teoría de Respuesta al Item, y 4 tests estiman este estadístico por muestra.

Tratamiento de la validez. La perspectiva mayoritaria adoptada en los manuales analizados no se adecúa al marco teórico defendido desde las directrices de la AERA; APA y NCMEA (1999). Incluso si se considera la concepción trinitaria de la validez (contenido, criterio, constructo), solo 11 de los manuales hacen referencia a las tres formas de validez. Evidencias sobre el contenido son recogidas en 8 manuales; informan sobre la estructura interna en 16 manuales, y sobre las relaciones entre el test y otras variables en 17. Sólo se ha encontrado una escueta referencia al proceso de respuesta o las consecuencias en 1 de los manuales.

Metodología de validación. La mayoría de los tests apoya sus evidencias en estudios correlacionales que son utilizados en 18 manuales. El modelo de regresión es utilizado en 2 de los manuales. El análisis factorial exploratorio es estimado en 11 manuales, y se aporta información sobre modelos confirmatorios en 6 manuales.

Interpretación de las puntuaciones. En el proceso de interpretación de las puntuaciones la mayoría de los manuales analizados proponen interpretaciones normativas basadas en la distribución de las puntuaciones o en normas de grado. Además, algunos de ellos, 7, ofrecen interpretaciones criteriales. Puntos de corte entre categorías son ofrecidos en 14 manuales. Sin embargo, justificaciones o evidencias para la definición de esos puntos de corte solo son ofrecidas en 8 manuales.

Los manuales ofrecen tablas o baremos para la interpretación de las puntuaciones. Se han utilizado transformaciones no-lineales en 20 manuales y transformaciones lineales en 19. La información sobre la distribución de las puntuaciones ofrecidas en los 7 manuales que la incorporan es escueta. La mayoría de los manuales no añaden información basada en el error de típico de medida en la lectura de los resultados.

Adaptación. Aunque la mayoría de los tests analizados son adaptaciones, el proceso de adaptación no aparece

TABLA 1
TESTS MÁS UTILIZADOS EN ESPAÑA

	Nombre castellano	Ámbito	1ª edición	Origen	Adaptación
MCMI-III*	Inventario Clínico Multiaxial de Millon-III	C-T	1977	EE.UU.	2007
16PF-5*	16PF-5, Cuestionario Factorial de Personalidad	C-E-T	1949	EE.UU.	1995/2005
MMPI-2-RF*	Inventario de Personalidad Multifásico de Minnesota-2-Reestructurado	C-T	1943	EE.UU.	2009
BDI-II*	Beck Depression Inventory-II	C	1961	EE.UU.	--
WISC-IV*	Escala de Inteligencia de Wechsler para Niños-IV	C	1949	EE.UU.	2005
WAIS-III*	Escala de Inteligencia de Wechsler para Adultos-III	C	1955	EE.UU.	1999
STAI*	Cuestionario de Ansiedad Estado-Rasgo	C	1970	EE.UU.	1994
RORSCHACH*	Rorschach	C	1921	Suiza	
SCL-90-R*	Cuestionario de 90 sintomas Revisado	C	1975	EE.UU.	2001
MMSE	Examen Cognoscitivo Mini-Mental	C	1975	EE.UU.	2002
BADYG	Batería de Aptitudes Diferenciales y Generales	E	1989	España	--
TALE	Test de Análisis de Lecto-Escritura	E	1980	España	--
MSCA	Escalas McCarthy de Aptitudes y psicomotricidad para Niños	E	1972	EE.UU.	1977/2006
RAVEN*	Raven Matrices progresivas	E	1938	Reino Unido	2001
PROLEC-R	Batería de Evaluación de los Procesos Lectores Revisada	E	1996	España	--
BENDER	Test gestaltico visomotor de Bender	E	1938	EE.UU.	1993
ITPA	Test Illinois de Aptitudes Psicolingüiesticas	E	1968	EE.UU.	1984/2004
TAMAI	Test Autoevaluativo Multifactorial de Adaptación Infantil	E	1983	España	No
PAPI	Inventario de personalidad y preferencias	T	1960	EE.UU.	
DAT-5	Test de Aptitudes Diferenciales-5	T	1947	EE.UU.	1960 / 2002
TPT	Test de Personalidad de TEA	T	2002	España	--
IPV	Inventario de Personalidad para Vendedores	T	1977	Francia	2005
IGF	Inteligencia General y Factorial Renovado	T	1991	España	--
BFQ	Cuestionario "Big Five"	T	1993	Italia	1995 /2007
NEO-PI-R	Inventario de Personalidad NEO Revisado	T	1978	EE.UU.	1999/2008

(C-Clínica, E-Educación, T-Trabajo; * 10 tests más usados independientemente de la especialidad)

bien documentado. Solo 8 tests discuten problemas relacionados con la equivalencia lingüística. El coeficiente de congruencia de Tucker aparece en 4 manuales. No se han encontrado pruebas sobre estudios de equivalencia estructural o métrica basada en modelos SEM ni en el estudio del funcionamiento diferencial del ítem.

DISCUSIÓN

Los tests constituyen el aspecto más conocido y de mayor impacto social relacionado con la investigación psicométrica. Desde los albores del siglo XX han servido a la investigación psicológica, y han ayudado en la toma de decisiones en los ámbitos educativo, social, jurídico o clínico. A lo largo de su historia su utilización y consiguiente consideración social han pasado por etapas de apogeo y crisis en cuyo origen encontramos usos abusivos e incorrectos.

En la práctica profesional los test psicológicos son herramientas de apoyo en el diagnostico, en el diseño de planes de intervención y evaluación, y en la selección profesional. Independientemente de la modalidad del test (cognitivo, neuropsicológico, adaptativo, social, conductual, de personalidad o vocacional), y de su propósito, es un instrumento que tiene que ser construido siguiendo principios que garanticen su calidad técnica (Wilson, 2005), y tiene que ser utilizado de acuerdo a criterios que permitan salvaguardar éstas. Sólo un uso apropiado y conforme a los propósitos para los que el test fue construido garantizará la validez de sus interpretaciones.

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 asientas sobre la inercia marcada por años de tradición y usos, y se retroalimenta con respecto a las prácticas establecidas. Pero la psicometría ha evolucionado, y la distancia entre teoría psicométrica y práctica profesional es hoy mayor que nunca. El desarrollo de la psicometría 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.

Las directrices elaboradas por las organizaciones profesionales 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. En este trabajo, que sigue la línea formativa y divulgadora iniciada desde el CGCOP , se pretendía mostrar el reflejo de dos de los referentes más importantes relacionados con el uso de los tests, las directrices conjuntas de la APA, y las directrices redactadas desde la Comisión Internacional de Tests. El resultado muestra los usos y costumbres asentados en nuestra práctica, pero también las asignaturas pendientes.

Las conclusiones más importantes referidas a cada uno de los puntos analizados se resumirían en la necesidad

FIGURA 1
APROXIMACIONES UTILIZADAS EN LOS MANUALES ANALIZADOS (N=22)

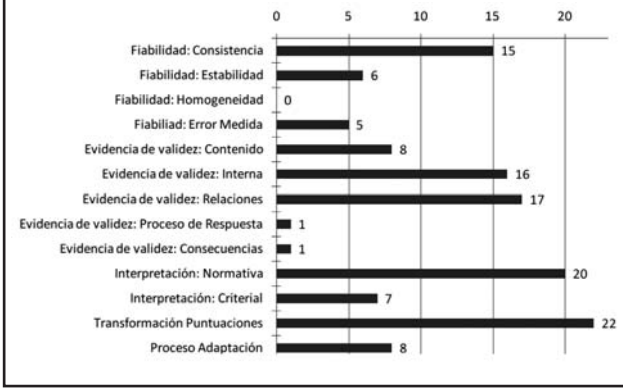
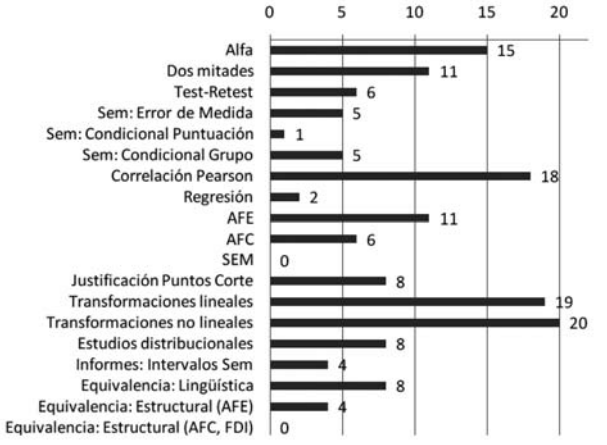


FIGURA 2
METODOLOGÍA UTILIZADA EN LOS MANUALES ANALIZADOS (N=22)



de aportar información sobre el error típico de medida, la necesidad de actualizar la idea de validez hacia una concepción más dinámica y argumental, la conveniencia de utilizar modelos confirmatorios y explicativos en el estudio de las relaciones entre variables, el interés en justificar las interpretaciones criteriales y de utilizar modelos adecuados para estimar la verosimilitud de un diagnóstico, o la importancia de garantizar la equivalencia en la adaptación de tests por medio de estudios de invarianza. Ninguno de los puntos es nuevo en la investigación psicométrica, y el trabajo de Muñiz-Fernández-Hermida (2010) muestra que los profesionales son conscientes de algunas de estas deficiencias (p.e., la no utilización del error de medida es reconocida como un problema por los profesionales). Sin embargo, tras más de una década de publicación de las directrices su contenido no está reflejado en la documentación analizada. Las directrices marcar un camino y la responsabilidad sobre la transición desde las costumbres hacia los deberes es compartida entre todos; profesionales, editores, colegios profesionales y profesores.

REFERENCIAS

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

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

Bartram, D. (1998). The need for international guidelines on standards for test use: A review of European and international initiatives. *European Psychologist*, 2, 155-163.

Bartram, D. (2011, Julio). The EFPA Test Review Model: Time for an Update? Symposium organizado en el *12th European Congress of Psychology*, Estambul, Turquía.

Bartram, D., Gregoire, J., Hambleton, R. K., y van de Vijver, F. (2011, Julio). International Test Commission Guidelines for adapting educational and psychological tests (2nd edition). Sesión especial en el *7th Conference of the International Test Commission*, Honk Kong, China.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software

Bentler, P.M. (1980). Multivariante analysis with latent variables: casual modeling. *Annual Review of Psychology*, 31, 419-456.

Bentler, P.M. (1986). Structural modeling and psychometrika: an historical perspective on growth and achievements, *Psychometrika,* 51, 31-51.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley

Bully, P. y Elosua, P. (2011, Julio). Classification procedures and cut-score definition in psychological testing: A review. Comunicación presentada en el *11th European Conference on Psychological Assessment*. Ritga.

Cizek, G. J., y Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

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.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J.R., Frans, O., Gintiliené, G., Hagemeister, C., Halama, P., Iliescu, D., Jaworowska, A., Jimenez, P., Manthouli, M., Matesic, K., Schittekatte, M., Sümer, C., & Urbánek, T. (in press). Testing practices in the 21th century. Developments and Europeas Psychologist's opinions. *European Psychologists.*

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(3), 497-503.

Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema* 21,4, 652-655.

Elosua, P. y Hambleton, R.K. (2011, Julio). Nuevas directrices de la ITC para la adaptación de tests. Trabajo presentado en el *XII Congreso de Metodología de las Ciencias Sociales y de la Salud*, San Sebastián.

Elosua, P. y Iliescu, D. (2011). Psychological test validity. Where we are an where we should to go. Comunicación presentada en el *12th European Congress of Psychology*, Estambul, Turquía.

Elosua, P. y Iliescu, D. (en prensa). Test in Europe. Where we are an where we should to go. *International Journal of Testing.*

Elosua, P. y Muñiz, J. (2010). Exploring the factorial structure of the Self-Concept: A sequential approach using CFA, MIMIC and MACS models, across gender and two languages. *European Psychologist* 15, 58-67.

Elosua, P. y Zumbo, B. (2008). Reliability coefficients for ordinal response scales. *Psicothema,* 20, 896-901.

Guilford, J.P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.

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. (en prensa). *Reporting Test Scores in More Meaningful Ways:A Research-Based Approach to Score Report Design*

Hattie, J. (1985). Methodology review: assessing unidimensionality of tests and items. *Applied Psychological Measurement* 9, 139–164.

Hattie, J. (2009, April) *Visibly learning from reports: The validity of score reports*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Ihaka, R., y Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5, 299–31

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, 36, 109-133.

Jöreskog, K.G. y Sörbom, D. (1996). *LISREL8. User's Reference Guide*. Chicago:Sci Software Int.

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.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling* (3rd Edition).The Guilford Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lorenzo-Seva, U. y Ferrando, P. J. (2007). *FACTOR: A computer program to fit the exploratory factor analysis model*. University Rovira y Virgilli

McDonald, R.P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology* 34,110–117.

McDonald, R. P. (1999). *Test theory. A unified treatment*. Mahwah, NJ, Lawrence Erlbaum Associates.

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. (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. 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.

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

Muthén, L. K. y Muthén, B. O. (2001). Mplus user's guide. Los Angeles: Muthén y Muthén.

Nunnally, J.C. (1978). *Psychometric theory* , New York: McGraw-Hill.

Raykov, T. (2001). Estimation of Congeneric Scale Reliability via Covariance Structure Analysis with Nonlinear, *British Journal of Mathematical and Statistical Psychology,54*, 315-323.

Rulon, P. J. (1946). On the validity of educational tests.*Harvard Educational Review*, 16,290-296.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology* 15, 201-293.

Thurstone, L. L. (1924/1973).*The Nature of Intelligence*. London: Routledge.

Tucker, L.R. (1951). *A method for synthesis of factor analysis studies*. Personnel Research Section Report, 984. Washington, D. C.: Department of the Arm

Van der Linden, W. y Hambleton, R.K. (1997). *Handbook of modern ítem 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.