Psicothema

Psicothema 2019, Vol. 31, No. 1, 7-16
doi: 10.7334/psicothema2018.291

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Diez pasos para la construcción de un test

José Muñiz[1] and Eduardo Fonseca-Pedrero[2]
[1] Universidad de Oviedo and [2] Universidad de La Rioja. CIBERSAM

Resumen

Antecedentes: los test son los instrumentos de medida más utilizados por los psicólogos para la obtención de muestras de comportamiento de las personas, tanto en contextos profesionales como en investigación. El objetivo del presente trabajo es sintetizar en diez pasos los aspectos fundamentales que hay que tener en cuenta a la hora de construir un test de forma rigurosa. **Método:** para la elaboración de las diez fases propuestas se revisó la literatura psicométrica especializada y se actualizaron trabajos previos de los autores sobre el tema. **Resultados:** se proponen diez pasos para la construcción objetiva de un test: delimitación del marco general, definición de la variable a medir, especificaciones, construcción de los ítems, edición del test, estudios piloto, selección de otros instrumentos de medida, aplicación de la prueba, propiedades psicométricas y desarrollo de la versión final. **Conclusión:** siguiendo los diez pasos propuestos, se pueden construir test objetivos con propiedades psicométricas adecuadas apoyadas en evidencias empíricas.

Palabras clave: test, fiabilidad, validez, construcción de test.

Abstract

Ten steps for test development. **Background:** Tests are the measurement instruments most used by psychologists to obtain data about people, both in professional and research contexts. The main goal of this paper is to synthesize in ten steps the fundamental aspects that must be taken into account when building a test in a rigorous way. **Method:** For the elaboration of the ten proposed phases, the specialized psychometric literature was revised, and previous works by the authors on the subject were updated. **Results:** Ten steps are proposed for the objective development of a test: delimitation of the general framework, definition of the variable to be measured, specifications, items development, edition of the test, pilot studies, selection of other measurement instruments, test administration, psychometric properties, and development of the final version. **Conclusion:** Following the ten proposed steps, objective tests can be developed with adequate psychometric properties based on empirical evidence.

Keywords: Test, reliability, validity, test development.

Los test son los instrumentos de medida más utilizados por los psicólogos para obtener datos sobre la conducta de las personas. A partir de esos datos los profesionales y los investigadores toman decisiones que pueden tener serias repercusiones sobre la vida de las personas evaluadas. Por tanto, es esencial que los test cumplan unos estrictos estándares científicos de rigor y calidad. Los test no toman decisiones por su cuenta, son los psicólogos quienes las toman, basándose en los datos obtenidos por este u otro procedimiento. Una evaluación rigurosa es la base de un diagnóstico preciso, que a su vez permita una intervención eficaz, basada en evidencias empíricas.

El objetivo central del presente trabajo es presentar de forma sintética los pasos generales que habría que seguir para construir un instrumento de medida con garantías de calidad. No se trata de una exposición exhaustiva, que excede las pretensiones de un artículo como este, pero esperamos que permita al lector extraer una idea cabal de cómo proceder si tuviese que desarrollar un nuevo test, escala o cuestionario. También se indican las referencias especializadas que permitan a los lectores profundizar en temas concretos. Tratamientos exhaustivos pueden verse en los trabajos de Downing y Haladyna (2006), Schmeiser y Welch (2006), Haladyna y Rodríguez (2013), Lane, Raymond y Haladyna (2016), e Irwing (2018), entre otros muchos. Aquí seguiremos en líneas generales los trabajos previos de los autores sobre el tema (Muñiz, 2018; Muñiz y Fonseca-Pedrero, 2008, 2017), por lo que queremos dejar constancia de nuestro agradecimiento a la *Revista de Investigación en Educación*, al Colegio Oficial de Psicólogos y a la Editorial Pirámide.

La construcción de un instrumento de medida es un proceso complejo que aquí vamos a sintetizar en diez pasos, si bien estos no son universales, pudiendo variar en función del propósito del instrumento de medida (selección, diagnóstico, intervención, etc.), del modelo psicométrico utilizado (Teoría clásica, Teoría de Respuesta a los Ítems —TRI–), del tipo de respuesta exigida por los ítems (selección, construcción), del formato de aplicación (lápiz y papel, informatizado), o del contexto de evaluación (clínico, educativo, del trabajo y las organizaciones, etc.), por citar solo algunos casos. Todo el proceso de construcción debe desarrollarse de forma rigurosa y objetiva, siguiendo unos estándares de calidad, para así maximizar la validez de las inferencias hechas a partir de las puntuaciones obtenidas, así como la equidad en la prueba de las personas evaluadas (Dorans y Cook 2016; Downing, 2006; Lane, Raymond y Haladyna, 2016). Puede decirse que el proceso

Received: September 2, 2018 • Accepted: December 12, 2018
Corresponding author: Eduardo Fonseca-Pedrero
Facultad de Letras y de la Educación
Universidad de La Rioja. CIBERSAM
26004 Logroño (Spain)
e-mail: eduardo.fonseca@unirioja.es

de validación ya comienza a fraguarse incluso antes de la propia elaboración empírica del instrumento, pues todas las acciones que se realicen antes, durante y después permitirán recoger evidencias que ayuden a la interpretación de las puntuaciones y a la posterior toma de decisiones (Elosua, 2003; Kane, 2006; Leong et al., 2016; Markus y Borsboom, 2013; Martínez-Arias, 2018; Muñiz, 2004, 2018; Wells y Faulkner-Bond, 2016; Zumbo, 2007).

En la tabla 1 se recogen de forma esquemática los diez pasos que se deben considerar en el proceso de construcción y validación de un test. Este procedimiento a seguir en esencia recoge las recomendaciones de los últimos estándares de la *American Educational Research Association* (AERA), la *American Psychological Association* (APA) y el *National Council on Measurement in Education* (NCME) (2014). Cierto es que otros autores como Downing (2006) y Lane et al. (2016) prefieren establecer doce pasos o fases; por supuesto, no existe un número mágico al respecto, lo esencial queda recogido en los diez propuestos. A continuación se comentan brevemente cada una de ellas.

Pasos para la construcción de un test

Marco general

Todo proceso de construcción de un instrumento de medida comienza por una explicación detallada y precisa de cuáles son las razones que motivan su desarrollo. Un nuevo instrumento no se construye porque sí, hay que justificarlo adecuadamente. Asimismo, hay que delimitar con claridad cuál es la variable objeto de medición, cuál va a ser el contexto de aplicación, las circunstancias en las que se va a aplicar el instrumento, el tipo de aplicación (individual, colectiva), el formato de administración (lápiz y papel, informática), y qué decisiones se van a tomar a partir de las puntuaciones (selección, diagnóstico, etc.). Las causas que pueden llevar a la construcción de un instrumento de medida son lógicamente diversas, por ejemplo, un psicólogo puede decidir construir un test porque no existe ningún otro para medir una determinada variable, porque los instrumentos existentes presentan unas puntuaciones con propiedades psicométricas deficientes, porque no incorporan alguna faceta relevante para analizar dicha variable, o simplemente porque los existentes se han quedado obsoletos. Wilson (2005) detalla y comenta las principales razones para generar nuevos instrumentos de medida.

Los responsables de la construcción del instrumento de medida no solo deben especificar el motivo por el cual quieren desarrollar una nueva herramienta, sino también deben delimitar con claridad cuál es el contexto en el que se va a aplicar, lo que incluye necesariamente la población objeto de medición (pacientes, alumnos, empresas, departamentos, etc.) y las circunstancias de aplicación (lugar, medios de los que se dispone y condiciones de aplicación, individual o colectiva). También debe especificarse de antemano con qué propósito van a ser utilizadas las puntuaciones y qué decisiones se van a tomar a partir de ellas. En este sentido, las puntuaciones en un instrumento de evaluación pueden servir para propósitos varios, tales como seleccionar, diagnosticar, clasificar, orientar, evaluar un dominio específico, o incluso como método de cribado (AERA, APA y NCME, 2014). Se debe dejar claro que las inferencias que se extraigan de las puntuaciones de un instrumento de medida no son universales, son siempre para un uso, contexto y población determinados. Nótese que lo que puede ser válido para un grupo determinado de personas o población, tal vez no lo sea para otra, y lo que pueda ser válido en un contexto de evaluación, no tiene por qué serlo en otro diferente (Zumbo, 2007).

Tabla 1
Fases del proceso de construcción de un test

1. Marco general
2. Definición de la variable medida
3. Especificaciones
4. Construcción de los ítems
5. Edición
6. Estudios piloto
7. Selección de otros instrumentos de medida
8. Aplicación del test
9. Propiedades psicométricas
10. Versión final del test

En suma, un instrumento de medida vale para lo que vale y hay que explicitarlo de forma clara. Ello no es óbice para que una prueba desarrollada originalmente con una determinada finalidad se revele en el futuro, tras distintos procesos de validación, como buena predictora de otros aspectos inicialmente no contemplados. Los usos que se hagan de una prueba deben venir avalados por evidencias empíricas, como bien establece la Norma UNE-ISO 10667 (2013), relativa a la evaluación de personas en entornos laborales y organizacionales. Más aún, como indica nuestro código deontológico en su artículo 17, el psicólogo tiene que estar profesionalmente preparado y especializado en la utilización de métodos, instrumentos, técnicas y procedimientos que adopte en su trabajo y debe reconocer los límites de su competencia y las limitaciones de sus técnicas.

Definición de la variable medida

El objetivo esencial de esta segunda fase es la definición operativa, semántica y sintáctica de la variable medida (AERA, APA y NCME, 2014; Carretero y Pérez, 2005; Wilson, 2005). La variable evaluada debe definirse en términos operativos para que pueda ser medida de forma empírica (Muñiz, 2004). En este sentido, tan interesante puede ser definir cuidadosamente lo que es como lo que no es. La facilidad o dificultad de la definición operativa depende en cierta medida de la naturaleza de la variable objeto de medición. Para llevar a cabo una definición operativa es clave realizar una revisión exhaustiva de la literatura publicada al respecto, así como la consulta a expertos (Clark y Watson, 1995; Wilson, 2005). Ello permite, por un lado, delimitar la variable objeto de medición y considerar todas las dimensiones relevantes de la misma y, por otro, identificar con claridad los comportamientos más representativos de tal variable (Calero y Padilla, 2004; Smith, 2005). Hay que evitar dejar fuera alguna faceta o dominio relevante (infrarrepresentación), así como ponderar en demasía una faceta o dominio (sobrerrepresentación) de la variable (Smith et al., 2003). Asimismo, no se deben incorporar facetas, o ítems, que no tengan relación con la variable objeto de medición (varianza irrelevante). Por ejemplo, si se construye una herramienta para medir el Trastorno por Déficit de Atención e Hiperactividad (TDAH), según los criterios del Manual diagnóstico y estadístico de los trastornos mentales-5 (DSM-5), no tendría sentido evaluar otros componentes con frecuencia asociados al TDAH como pudieran ser los problemas emocionales.

Una definición operativa y precisa de la variable influye de forma determinante en la posterior obtención de los diferentes tipos

de evidencias de validez, ayuda a especificar las conductas más representativas de la variable objeto de medición y facilita el proceso de construcción de los ítems (Carretero y Pérez, 2005; Elosúa, 2003; Muñiz et al., 2005; Sireci, 1998a, 1998b; Smith, 2005). No solo es importante una definición operativa de la variable, sino que también es preciso identificar y definir las facetas o dominios de la misma (definición semántica) y la relación que se establece entre ellas, así como con otras variables de interés (definición sintáctica) (Lord y Novick, 1968). Lógicamente, las diferentes facetas que componen la variable medida se deberían encontrar relacionadas, dado que se supone que miden la misma variable o constructo. Al mismo tiempo hay que establecer la relación con otras variables de interés (red nomológica) dado que la variable objeto de medición no se encuentra aislada en el mundo, sino que está en relación o interacción con otras variables. Es interesante comprender y analizar estas relaciones especificándolas de antemano con el propósito de llevar a cabo posteriores estudios dirigidos a la obtención de evidencias de validez (Carretero y Pérez, 2005; Muñiz, 2004; Smith, 2005).

Especificaciones

Una vez delimitados el propósito de la evaluación y la definición operativa de la variable que interesa medir, se deben llevar a cabo determinadas especificaciones relacionadas con el instrumento de medida. En esta fase se deben describir de forma detallada y precisa aspectos concernientes a los requerimientos de aplicación del instrumento, el tipo, número, longitud, contenido y distribución de los ítems, especificaciones e instrucciones en la entrega del material y aspectos relacionados con la seguridad del mismo.

Los requerimientos de aplicación del instrumento de medida se refieren a cuál va a ser el soporte de administración (papel y/o informático), a qué tipo de aplicación se va a realizar (individual y/o colectiva), y cuándo y en qué lugar se va a administrar el instrumento de medida. Igualmente, se deben especificar los requerimientos cognitivos, de vocabulario y de accesibilidad de los participantes, así como las derivadas del rango de edades al que se pretende aplicar el mismo. Es importante llevar a cabo adaptaciones (acomodaciones) para aquellos participantes que no puedan desempeñar la tarea en igualdad de condiciones que el resto; por ejemplo, disponer de una versión en *Braille* o macrotipos, para una persona con discapacidad visual o con baja visión. También se podría disponer de una versión en lectura fácil para personas con discapacidad intelectual y/o para otros colectivos con dificultades de comprensión lectora. Las adaptaciones que se realicen deben de estar convenientemente avaladas por evidencias empíricas para que no supongan ventajas ni desventajas respecto de la aplicación estándar (Dorans y Cook, 2016; Wells y Faulkner-Bond, 2016).

En relación con los ítems, se debe especificar el tipo, el número, la longitud, el contenido y el orden (disposición) de los mismos, así como el formato de respuesta o el tipo de alternativas que se va a utilizar. Con respecto a este tema, no existen normas universales, todo dependerá de las circunstancias de aplicación, del propósito de la variable objeto de medición y de otras circunstancias. También en esta fase se deberían especificar las medidas a emplear para el control de los sesgos de respuesta, tanto a nivel de redactado de ítems como a nivel de la utilización de procedimientos para el cálculo de las puntuaciones (Wetzel et al., 2016).

Construcción de los ítems

La construcción de los ítems constituye una de las etapas más cruciales dentro del proceso de elaboración del test. Los ítems son la materia prima, los ladrillos a partir de los cuales se conforma un instrumento de evaluación, por lo que una construcción deficiente de los mismos incidirá en las propiedades métricas finales del instrumento de medida y en la validez de las inferencias que se hagan a partir de las puntuaciones (Haladyna y Rodríguez, 2013; Lane, Raymond y Haladyna, 2016; Muñiz et al., 2005; Osterlind, 1998; Schmeiser y Welch, 2006).

Si los ítems provienen de otro instrumento ya existente en otro idioma y cultura deberán seguirse las directrices internacionales para la traducción y adaptación de test (Hambleton, Merenda y Spielberger, 2005; International Test Commission, 2017; Muñiz y Bartram, 2007; Muñiz, Elsoua y Hambleton, 2013). En el caso de ítems originales han de seguirse las directrices elaboradas para su desarrollo (Downing y Haladyna, 2006; Haladyna, 2004; Haladyna et al., 2002, 2013; Moreno et al., 2004, 2006, 2015; Muñiz, 2018; Muñiz et al., 2005).

Los principios básicos que deben regir la construcción de cualquier banco de ítems son: representatividad, relevancia, diversidad, claridad, sencillez y comprensibilidad (Muñiz et al., 2005). Como se comentó anteriormente, todos los dominios de la variable de interés deben de estar igualmente representados, evitando la infra o sobrerrepresentación, aproximadamente con el mismo número de ítems. Existe alguna excepción, por ejemplo, cuando se haya considerado un dominio más relevante dentro de la variable y que, por tanto, deba tener un mayor número de ítems, esto es, una mayor representación. Por ejemplo, en el estudio de la esquizotipia (riesgo de psicosis) en jóvenes, la faceta Anhedonia ha demostrado tener un mayor poder predictivo en la transición al cuadro clínico (Fonseca Pedrero y Debbané, 2017). Considerando el modelo teórico subyacente y la evidencia empírica acumulada, en este caso podría estar justificado representar esta faceta con más ítems. Un muestreo erróneo del dominio objeto de evaluación sería una clara limitación en la obtención de evidencias de validez de contenido y tendrá repercusiones en las inferencias y decisiones que con posterioridad se hagan a partir de las puntuaciones. Los ítems deben de ser heterogéneos y variados para así recoger una mayor variabilidad y representatividad de la variable. Debe primar la claridad y la sencillez, se deben evitar tecnicismos, negaciones, dobles negaciones, o enunciados excesivamente prolijos o ambiguos (Muñiz, 2018; Muñiz et al., 2005). Una práctica habitual es formular los ítems positivamente junto con otros inversos y luego recodificar; sin embargo, esta estrategia no está libre de limitaciones (Suárez et al., 2018). Del mismo modo, los ítems deben ser comprensibles para la población a la cual va dirigido el instrumento de medida, utilizando un lenguaje inclusivo y evitándose en todo momento un lenguaje ofensivo y/o discriminatorio. Ítems con una redacción defectuosa o excesivamente vagos van a incrementar el porcentaje de varianza explicada por factores espurios o irrelevantes, con la consiguiente merma en las evidencias de validez de la prueba. Un aspecto fundamental es garantizar que las personas comprendan perfectamente lo que se está preguntando en el test. Con este fin, y como se verá a continuación, puede ser interesante realizar un estudio piloto donde se analice el grado de comprensión de los ítems por parte de los usuarios a los que va a ir destinado la herramienta. Asumir que una redacción es adecuada para una población o que un concepto o expresión va a ser comprendida puede llevar a

errores graves. Por ejemplificar esta cuestión, en un estudio con jóvenes realizado por nuestro equipo tuvimos que eliminar ítems ya que un porcentaje considerable no comprendía la palabra "superstición" o la expresión "junta de las baldosas".

Durante las fases iniciales de la construcción del banco de ítems se recomienda que el número de ítems inicial sea como mínimo el doble del que finalmente se considera que podrían formar parte de la versión final del instrumento de medida. La razón es bien sencilla: muchos de ellos por motivos diferentes (métricos, comprensibilidad, dificultad, etc.) se acabarán desechando, por lo que solo quedarán aquellos que ofrezcan mejores indicadores o garantías técnicas (sustantivas y métricas). Finalmente, para garantizar la obtención de evidencias de validez basadas en el contenido de los ítems, se ha de recurrir a la consulta de expertos y a la revisión exhaustiva de las fuentes bibliográficas, así como a otros instrumentos similares ya existentes (Sireci, 1998b; Sireci y Faulkner-Bond, 2014). En relación con la valoración de los ítems por parte de los expertos y con la finalidad de una evaluación más precisa y objetiva del conjunto inicial de ítems, se puede pedir a los expertos que juzguen, a partir de un cuestionario, si los ítems están bien redactados para la población de interés, si son o no pertinentes para evaluar una faceta o dominio determinado y si cada ítem representa de forma adecuada la variable o dimensión de interés. Tratamientos exhaustivos sobre el desarrollo y análisis de los ítems pueden verse en Osterlind (1998), Haladyna y Rodríguez (2013), o Lane et al. (2016).

Existe una gran variedad de ítems que se pueden clasificar en distintas categorías en función de los criterios que se tengan en cuenta, tales como su contenido, el formato, o la forma de respuesta exigida, bien sea seleccionar una respuesta entre las presentadas, o desarrollarla (Downing, 2006; Haladyna y Rodríguez, 2013; Magno, 2009; Osterlind, 1998; Osterlind y Merz, 1994; Rauthmann, 2011; Sireci y Zenisky, 2006). Por ejemplo, Scalise y Gifford (2006) establecen siete tipos de ítems en función del formato de respuesta, que van desde la selección pura de la respuesta hasta la construcción completa, pasando por varias posibilidades intermedias. Por su parte, Sireci y Zenisky (2016) añaden todavía otros tipos de tareas. Se han propuesto diversas clasificaciones, tratando de sistematizar y organizar la gran cantidad de tipos de ítems existentes, si bien resultan útiles en la práctica, ninguna de ellas resulta totalmente satisfactoria desde un punto de vista teórico (Moreno, Martínez y Muñiz, 2018). Esta proliferación de los tipos de ítems se ha acentuado en los últimos años debido a las grandes posibilidades que ofrecen las tecnologías de la información y la comunicación, que están influyendo de forma clara en su formulación (Sireci y Zenisky, 2016).

Según Parshall et al. (2010) habría siete dimensiones o aspectos de los ítems en los que se están produciendo las mayores innovaciones, debido a la irrupción de las nuevas tecnologías: a) *estructura*, con la aparición de nuevos formatos facilitados por las nuevas tecnologías y las facilidades que ofrecen las pantallas de los ordenadores para su implementación; b) *complejidad*, al incluirse en los ítems nuevos elementos que han de tenerse en cuenta para responder; c) *fidelidad*, referida a la posibilidad que ofrecen las tecnologías de la información para dar un mayor realismo a los ítems; d) *interactividad*, dado que el ítem puede reaccionar y mutar en función de las respuestas de las personas, volviéndose interactivo; e) *multimedia*, cuando se incluyen en los ítems medios técnicos como audio, vídeo, gráficos, animación, u otros; f) *tipo de respuesta*, habiendo una amplia gama de posibilidades del tipo de tareas que los ítems demandan; y g) *sistemas de puntuación*, pudiendo registrarse además de los clásicos aciertos y errores otros muchos parámetros, tales como tiempos, intentos, estrategias, etc. Un excelente trabajo sobre la influencia de los avances tecnológicos sobre los tests puede verse en Drasgow (2016) y para el tema específico de la generación automática de ítems el lector puede revisar el trabajo de Gierl y Haladyna (2013).

Mención especial requieren los ítems tipo Likert, cuyo nombre proviene del trabajo original del autor (Likert, 1932), tan omnipresente en el ámbito de la medición de las actitudes, opiniones, preferencias, creencias y otros campos afines. La popularidad de este formato proviene de su facilidad de aplicación y de su adaptación a cualquier ámbito de evaluación, de ahí que en la práctica se haya impuesto a otros modelos mejor fundados científicamente, pero de aplicación menos sencilla, como el de las comparaciones binarias de Thurstone (1927a, 1927b, 1928), entre otros. La literatura es abundante sobre su construcción y uso (Dillman et al., 2009; Haladyna y Rodríguez, 2013; Krosnick y Presser, 2010; Suárez et al., 2018), pero excede los objetivos de este trabajo. Para una revisión, véase por ejemplo Muñiz (2018).

También, cabe citar el formato tipo *Ensayo,* en el cual las personas evaluadas deben generar la respuesta, en contraposición con los formatos selectivos, como los de elección múltiple. La gran ventaja de los formatos de construcción frente a los selectivos es que permiten una mayor libertad de expresión de la persona evaluada, pudiendo apreciarse su capacidad de expresión, su creatividad, su estilo y organización, amén de su dominio del tema propuesto. Por estas razones, es un formato muy apreciado entre los educadores, que con cierta frecuencia lo prefieren a los formatos de elección, los cuales más que generar la propia respuesta exigen a la persona evaluada reconocer la alternativa correcta entre las propuestas. Pero todo tiene un precio y el de los formatos de desarrollo es la posible subjetividad a la hora de la corrección y puntuación, lo cual hay que evitar necesariamente, por razones obvias. Para evitar los sesgos es muy importante instruir y entrenar de forma adecuada a los correctores, así como enseñarles a establecer unos criterios claros que les permitan una corrección más analítica y objetiva. Estos criterios de corrección, denominados rúbricas, no solo permiten una mayor objetividad a la hora de corregir, sino que han de hacerse públicos para orientar a las personas evaluadas. Estas directrices o rúbricas no son la panacea, pero ayudan a objetivar la evaluación y a evitar sesgos y subjetivismos. El problema de la construcción y valoración de los ítems de ensayo está ampliamente tratado en la bibliografía psicométrica, pueden consultarse, por ejemplo, las directrices del *Educational Testing Service* (Baldwin et al., 2005; Livingston, 2009), el trabajo de Hogan y Murphy (2007) y un buen resumen en Haladyna y Rodríguez (2013). Tal vez la solución radical a la objetividad de la evaluación de los ensayos venga a través de una vigorosa línea actual de investigación psicométrica sobre la corrección automática mediante programas de ordenador. Puede sorprender al lector la posibilidad de que un ensayo pueda ser corregido por un programa informático, pero los avances en este campo son notables, existiendo ya programas con altas prestaciones (Livingston, 2009; Shermis y Burstein, 2013; Williamson et al., 2006, 2010).

Edición

En esta fase se compone y se imprime la primera versión del test, además de construir la base de datos con las claves de correc-

ción. Este paso ha sido con frecuencia injustamente infraestimado pese a que es esencial, pues el continente bien podría echar a perder el contenido. Buenos ítems pobremente editados dan como resultado un mal test. Podemos haber construido un buen banco de ítems que de nada servirá si luego se presentan de forma desorganizada, con errores tipográficos, o en un cuadernillo defectuoso. Uno de los errores más frecuentes entre los constructores de test aficionados es utilizar fotocopias malamente grapadas, con la excusa de que solo se trata de una versión experimental de la prueba, olvidándose de que para las personas que las responden no existen pruebas experimentales, todas son definitivas. El aspecto físico de la prueba forma parte de su validez aparente. Es importante que el instrumento dé la impresión de medir de manera objetiva, rigurosa, fiable y válida la variable de interés, porque, entre otros aspectos, influye en un punto esencial presente en todo el proceso de evaluación: la motivación y actitud de las personas evaluadas. Por otra parte, en esta fase también se debe construir, si fuera el caso, la base de datos donde posteriormente se van a tabular las puntuaciones y a realizar los análisis estadísticos pertinentes, así como las normas de corrección y puntuación, por ejemplo, si existen ítems que se deben recodificar, si se va a crear una puntuación total o varias puntaciones, etc.

Estudios piloto

La finalidad de cualquier estudio piloto es examinar el funcionamiento general del instrumento de medida en una muestra de participantes con características semejantes a la población objeto de interés. Esta fase es de suma importancia ya que permite detectar, evitar y corregir posibles errores, así como llevar a cabo una primera comprobación del funcionamiento del test en el contexto aplicado. El estudio piloto podría verse como una representación en miniatura de lo que posteriormente va a ser el estudio de campo.

Existen dos tipos fundamentales de estudio piloto: cualitativo y cuantitativo (Wilson, 2005). El estudio piloto cualitativo permite, a partir de grupos de discusión, debatir diferentes aspectos relacionados con el instrumento de medida, por ejemplo, la detección de errores semánticos o gramaticales, el grado de comprensibilidad de los ítems, las posibles incongruencias semánticas, etc. Los participantes en este pilotaje pueden ser similares a la población objeto de medición. Por su parte, el estudio piloto cuantitativo permite examinar las propiedades métricas de la versión preliminar del instrumento de medida y ha de llevarse a cabo con personas similares a las que va dirigida la prueba. En ambos casos se deben anotar de forma detallada todas las posibles incidencias acaecidas durante la aplicación, como, por ejemplo, preguntas o sugerencias de los participantes, grado de comprensión de los ítems, así como posibles errores o problemas detectados en el instrumento.

A continuación, una vez tabulados los datos, se procede a los análisis de la calidad psicométrica de los ítems. En función de criterios sustantivos y estadísticos (p.ej., índice de discriminación, cargas factoriales, funcionamiento diferencial del ítem, etc.), algunos ítems se mantienen, mientras que otros son descartados o modificados. Por ejemplo, en esta fase (al igual que en la fase novena de propiedades psicométricas) se puede examinar la estructura dimensional que subyace a las puntuaciones del instrumento de medida y eliminar aquellos ítems con una carga factorial baja (usualmente inferior a 0,30) o que no se han ajustado al modelo factorial hipotetizado. Es importante que el constructor del instrumento de evaluación deje constancia de qué ítems fueron eliminados o modificados y por qué, además de explicitar con claridad el criterio (cualitativo o cuantitativo) por el cual se eliminaron. En este paso, si se considera conveniente, se pueden incorporar nuevos ítems. Todas las actividades deben ir destinadas a seleccionar los ítems con mayores garantías métricas que maximicen las propiedades finales del instrumento de evaluación. Finalmente, se debe construir una nueva versión del instrumento de medida que es revisada de nuevo por el grupo de expertos y que será la que en última instancia se administre en el estudio final de campo.

Selección de otros instrumentos de medida

La selección adecuada de otros instrumentos de evaluación permite recoger evidencias a favor de la validez de las puntuaciones de los participantes (Elosúa, 2003). Es interesante que no se pierda el norte, la finalidad última de todo proceso de construcción de instrumentos de medida es siempre obtener evidencias de validez. La selección adecuada de otras variables de interés permite aglutinar diferentes tipos de evidencias que conduzcan a una mejor interpretación de las puntuaciones en el instrumento de medida dentro de un contexto y uso particular. En este sentido, se pueden establecer relaciones con un criterio externo, con otros instrumentos de medida que pretendan medir la misma variable u otras diferentes (lo que anteriormente se había denominado definición sintáctica). Las asociaciones entre las variables son la base para la obtención de evidencias de validez de relación con variables externas, que permite la construcción de una red nomológica.

La decisión de qué instrumentos se deben utilizar complementariamente con el desarrollado viene afectada tanto por cuestiones sustantivas como pragmáticas, referidas a exigencias de tiempo y lugar y, cómo no, materiales como la posibilidad de acceso al test, cuestiones económicas, etc. Evidentemente, las exigencias materiales y temporales así como las razones éticas no permiten aplicar todos los instrumentos que quisiéramos, si bien aquí no se trata de pasar cuantos más mejor, sino de seleccionar aquellos de mayor calidad científica, a partir de los cuales se pueda profundizar en el significado de nuestras puntuaciones. Algunas recomendaciones prácticas en la selección de otros instrumentos de medida son: a) que se encuentren validados para la población objeto de interés y se conozcan las propiedades psicométricas de sus puntuaciones; b) que sean sencillos y de rápida aplicación; y c) que tengan coherencia sustantiva de cara a establecer relaciones entre las variables, dentro de su red nomológica.

Aplicación del test

En esta fase de estudio de campo se incluye la selección de la muestra (tipo, tamaño y procedimiento), la aplicación propiamente dicha del instrumento de medida a los participantes y el control de calidad y seguridad de la base de datos. La representatividad y generalizabilidad de los resultados depende en gran medida de que la muestra elegida sea realmente representativa de la población objetivo de estudio. Elegir una muestra pertinente en cuanto a representatividad y tamaño es esencial, si se falla en esto todo lo demás va a quedar invalidado. El muestreo probabilístico siempre es preferible al no probabilístico, para la estimación del tamaño muestral requerido para un determinado error de medida ha de acudirse a los textos especializados, o consultar los expertos en la tecnología de muestreo. Aunque no hay recetas universales y

aún no se dispone de una base sólida para tal afirmación, se suele recomendar que por cada ítem administrado tengamos al menos 5 o 10 personas, o unas 200 observaciones como mínimo (Ferrando y Anguiano, 2010), si bien determinadas técnicas estadísticas pueden reclamar incluso más de cara a una buena estimación de los parámetros, por ejemplo, los modelos de TRI (van der Linden, 2016).

Las actividades relacionadas con la aplicación y el uso del instrumento de medida son cruciales durante el proceso de validación (Muñiz y Bartram, 2007; Muñiz et al., 2005). Cuando aplicamos cualquier instrumento de medida hay que cuidarse de que las condiciones físicas de la aplicación sean las adecuadas (luz, temperatura, ruido, comodidad de los asientos, etc.). Igualmente, las personas encargadas de la administración del instrumento de medida deben establecer una buena relación (*rapport*) con los participantes, estar familiarizados con la administración de este tipo de herramientas, dar las instrucciones a los participantes correctamente, ejemplificar con claridad cómo se resuelven las preguntas, supervisar la administración y minimizar al máximo las posibles fuentes de error. Un aspecto que no se debe descuidar es el referido a las instrucciones en la entrega del material. Las directrices utilizadas (el qué dice y el cómo lo dice) por evaluador en la administración del test pueden modificar drásticamente las respuestas de las personas. Por ejemplo, si se aplica un test que evalúa las experiencias psicóticas atenuadas, es crucial mencionar a los participantes que para responder al mismo no deben considerar aquellas experiencias que se hayan dado bajo los efectos del consumo de sustancias o en estados febriles intensos. Sin esta directriz, las conclusiones extraídas a partir de los datos podrían ser totalmente infundadas. Por todo ello, es recomendable elaborar unas pautas o directrices que permitan estandarizar la administración del instrumento de medida y garanticen la equidad.

El control de calidad de la base de datos es otro tema a veces poco valorado en el proceso de construcción de instrumentos de medida. Por control de calidad nos referimos a una actividad que tiene como intención comprobar que los datos introducidos en la base de datos se correspondan exactamente con las puntuaciones de los participantes en la prueba. En la forma clásica de introducir de forma manual las puntuaciones de los participantes en una base de datos o en una plataforma de una casa editorial se pueden cometer multitud de errores, por ello es altamente recomendable comprobar de forma rigurosa que los datos se han introducido correctamente. En algunos casos una estrategia sencilla que se puede utilizar a posteriori es la de extraer al azar un cierto porcentaje de los participantes y comprobar la correspondencia entre las puntuaciones en la prueba y la base de datos. Actualmente, las nuevas tecnologías y la aplicación de test *on line* permiten soslayar estos problemas, si bien surgen otros nuevos referidos a la calidad de la conexión a internet y de los ordenadores, la seguridad de los datos, por mencionar algunos. No obstante, los mejores errores son los que no se cometen, así que hay que poner todos los medios para minimizarlos a la hora de construir la base de datos.

Propiedades psicométricas

Una vez aplicado el test a la muestra de interés se procede al estudio de las propiedades psicométricas de las puntuaciones del mismo: análisis de los ítems, estimación de la fiabilidad de las puntuaciones, obtención de evidencias de validez (p. ej., estudio de la dimensionalidad, análisis del funcionamiento diferencial de los ítems, relación con variables externas) y construcción de baremos. La fiabilidad se refiere a la precisión de las puntuaciones, mientras que la validez se refiere a la calidad de las inferencias hechas a partir de las puntuaciones (Muñiz, 2018; Prieto y Delgado, 2010). En sentido estricto no es fiable el test, sino las puntuaciones obtenidas en él. Análogamente, un test no es válido, sino que lo son las inferencias hechas a partir de las puntuaciones.

En esta fase debe primar por encima de todo el rigor metodológico. Todos los pasos y decisiones que se tomen se deben describir con claridad y deben estar correctamente razonadas. En un primer lugar, se deben analizar los ítems tanto a nivel cualitativo como cuantitativo. Para seleccionar los mejores ítems desde el punto de vista psicométrico se puede tener en cuenta el índice de dificultad (cuando proceda), el índice de discriminación, las cargas factoriales y/o el funcionamiento diferencial de los ítems (Muñiz et al., 2005). El funcionamiento diferencial de los ítems trata de garantizar la equidad en el proceso de medición. La ausencia de funcionamiento diferencial en un ítem supone que la probabilidad de respuesta correcta depende únicamente del nivel del participante en la variable objeto de medición y no está condicionada por la pertenencia a un grupo determinado o característica, por ejemplo, género, cultura, u otro aspecto cualquiera (Gómez-Benito et al., 2018). No se debe perder de vista que la finalidad del análisis psicométrico de los ítems es maximizar o potenciar las propiedades métricas del instrumento de medida; no obstante, no existen reglas universales y las consideraciones estadísticas no garantizan unos resultados con significación conceptual, por lo que hay que tener presente también los aspectos sustantivos (Muñiz et al., 2005).

Una vez seleccionados los ítems, se procede al estudio de la dimensionalidad del instrumento para obtener evidencias de validez de su estructura interna. En el caso de encontrar una solución esencialmente unidimensional nos podríamos plantear la construcción de una puntuación total, en el caso de una estructura multidimensional deberíamos pensar en un conjunto de escalas o perfil de puntuaciones. El análisis factorial exploratorio y confirmatorio y el análisis de componentes principales son las técnicas multivariantes más utilizadas para examinar la estructura interna que subyace a las puntuaciones de un instrumento de evaluación (Ferrando y Anguiano, 2010), si bien no son las únicas (Cuesta, 1996).

Una vez determinada la dimensionalidad de las puntuaciones del instrumento de medida, se lleva a cabo una estimación de la fiabilidad, para lo cual se pueden seguir diversas estrategias, tanto desde el punto de vista de la teoría clásica de los test como de la TRI (Muñiz, 1996, 1997, 2000, 2018). Cabe mencionar aquí otros procedimientos para estimar la fiabilidad de las puntuaciones que van más allá del clásico coeficiente alpha de Cronbach y que permiten soslayar algunas de sus limitaciones, como, por ejemplo, el coeficiente Omega, el alpha para datos ordinales o la función de la información desde el prisma de la TRI (Elosua y Zumbo, 2008; Muñiz, 2018). Además, no se debe perder de vista la importancia de incluir el error de medida en los informes psicológicos, manuales de test y artículos científicos. Debe quedar meridianamente claro que, en la mayoría de los casos, la puntación obtenida es una estimación y, por lo tanto, conlleva cierto grado de error.

Posteriormente, de cara a obtener evidencias de validez, se debe observar la relación del instrumento de medida con otros instrumentos de evaluación y, finalmente, se lleva a cabo una baremación del instrumento de medida donde se pueden establecer puntos de corte con alguna finalidad práctica o profesional.

Los desarrollos estadísticos y técnicos en este campo son notables, incorporándose cada vez más a menudo los métodos estadísticos robustos (Erceg-Hurn y Mirosevich, 2008), el análisis factorial confirmatorio (Brown, 2015), los tests adaptativos informatizados (Olea, Abad, y Barrada, 2010; Wells y Faulkner-Bond, 2016), o el análisis de redes (Borsboom y Cramer, 2013; Fonseca-Pedrero, 2017), por mencionar algunos.

Versión final del test

En último lugar, se procede a la elaboración la versión definitiva del test, se envía un informe de resultados a las partes legítimamente implicadas en el proceso de evaluación y se elabora el manual que permita su utilización a otras personas o instituciones. El manual debe de recoger con todo detalle las características relevantes de la prueba. Finalmente, y aunque sea la última fase, esto no quiere decir que el proceso de validación concluya aquí, ya que posteriores estudios deberán seguir recogiendo evidencias de validez que permitan tomar decisiones fundadas a partir de las puntuaciones de las personas. Asimismo, conviene llevar a cabo una evaluación rigurosa y sistemática del instrumento elaborado, para lo cual puede utilizarse el *Modelo de Evaluación de Tests* elaborado por la *European Federation of Professional Psychologists Associations* (EFPA), adaptado en España por Hernández, Ponsoda, Muñiz, Prieto y Elosua (2016) (tabla 2). Este modelo es una guía que permite analizar la calidad del instrumento en función de sus características. Dicho modelo se articula a través del Cuestionario para la Evaluación de los Tests, edición Revisada (CET-R), disponible en la página web del Consejo General de la Psicología de España (www.cop.es/uploads/pdf/CET-R.pdf).

Todo fluye: mirando hacia el futuro

Se han descrito los diez pasos fundamentales que habría que seguir para desarrollar un test objetivo y riguroso para evaluar variables en el ámbito de las ciencias sociales y de la salud. Estos pasos no se pueden abordar en profundidad desde un punto de vista técnico en un breve artículo como este, no se trataba de eso, sino de poner a disposición de los estudiantes y profesionales una guía general que les permita obtener una visión panorámica de las actividades implicadas en el desarrollo de los instrumentos de medida. Se aporta además la bibliografía especializada a la que pueden acudir aquellos interesados en profundizar en cada temática. El campo de la elaboración de instrumentos de medida está altamente desarrollado y es necesario acudir a personal cualificado para su desarrollo adecuado, constituyendo una temeridad dejarlo en manos de aficionados bienintencionados. Que un instrumento de evaluación esté adecuadamente construido y reúna las propiedades técnicas adecuadas es condición necesaria, pero no es suficiente, además hay que utilizar la prueba de forma pertinente.

Tabla 2
Ficha resumen del Cuestionario para la Evaluación de los Tests edición revisada (CET-R) (Hernández et al., 2016).

Características	Valoración	Puntuación
Materiales y documentación		
Fundamentación teórica		
Adaptación		
Análisis de los ítems		
Validez: contenido		
Validez: relación con otras variables		
Validez: estructura interna		
Validez: análisis del funcionamiento diferencial de los ítems		
Fiabilidad: equivalencia		
Fiabilidad: consistencia interna		
Fiabilidad: estabilidad		
Fiabilidad: Teoría Respuesta a los Ítems		
Fiabilidad: inter-jueces		
Baremos e interpretación de puntuaciones		

Las diez fases descritas no son estáticas ni inmanentes, la evaluación evoluciona muy rápidamente, influenciada sobre todo por los vertiginosos cambios impulsados por las tecnologías de la información y comunicación, y en especial los avances informáticos, multimedia e Internet. Autores como Bennet (1999, 2006), Breithaupt, Mills y Medican (2006), Drasgow (2016), Drasgow, Luecht y Bennet (2006) o Sireci y Faulkner-Bond (2016), entre otros muchos, consideran que dichas tecnologías están influyendo sobre todos los aspectos de la evaluación psicológica, tales como el diseño de los tests, la construcción y presentación de los ítems, la puntuación de los test y la evaluación a distancia. Emergen nuevas formas de evaluación, aunque no nos engañemos, los test psicométricos seguirán siendo herramientas fundamentales, dada su objetividad y economía de medios y tiempo (Phelps, 2005, 2008). En este contexto de cambio tecnológico surge la llamada Psicología 2.0 (Armayones et al., 2015), que pretende extender la psicología a través de las facilidades que ofrece Internet y las redes sociales. La evaluación no puede estar ajena a estas nuevas tendencias, apareciendo nuevos enfoques psicométricos conectados con el análisis de las grandes bases de datos (*big data*) de las que se dispone actualmente (Markovetz, Blaszkiewicz, Montag, Switala, y Schlaepfer, 2014). Por ejemplo, las ventajas potenciales de usar los teléfonos móviles o la realidad virtual tanto para evaluación como intervención abren nuevas posibilidades para la psicología del futuro (Armayones et al., 2015; Chernyshenko y Stark, 2016; Miller, 2012; Rus-Calafell, Garety, Sason, Craig, y Valmaggia, 2018). El uso de estos dispositivos móviles en salud mental se ha venido a llamar fenotipado digital (Insel, 2017). Además, trabajos como el pionero de Kosinski, Stillwell y Graepel (2013) analizan con éxito la posibilidad de utilizar los "me gusta" de *facebook* como predictores de distintas características humanas, entre ellas los rasgos de la personalidad, lo que hace preguntarse si nuestros rastros en las redes sociales sustituirán algún día no muy lejano a los cuestionarios y test tal como los conocemos ahora.

Otro tema que cobra pujanza es el de la evaluación ambulatoria, la evaluación ecológica momentánea o la metodología de muestreo de experiencias, que si bien tienen rancio abolengo en psicología, están resurgiendo con fuerza en la actualidad impulsada por las tecnologías de la información y comunicación (Chernyshenko y Stark, 2016; Myin-Germeys et al., 2018; Trull y Ebner-Priemer, 2013; van Os, Delespaul, Wigman, Myin-Germeys y Wichers, 2013). Este conjunto de métodos y procedimientos que tratan de estudiar mediante dispositivos móviles (p.ej., *smartphone*, *tablet*) las experiencias de las personas (emociones, sentimientos, pensamientos, síntomas psicológicos, etc.), en su entorno natural y en

la vida diaria. Esta metodología permite evaluar determinadas variables psicológicas desde una perspectiva más dinámica, personalizada, contextual y ecológica. En esencia se trata de captar la naturaleza dinámica, individual y contextual del ser humano, buscando posibles mecanismos causales (van Os et al., 2013). Para ello, habitualmente se realizan evaluaciones varias veces al día (aproximadamente 6-8 por día) durante un período temporal (típicamente una semana) para captar suficientemente la variabilidad de los fenómenos. Las preguntas se activan mediante un *beep* en un marco temporal fijado por el investigador, por ejemplo, entre las diez de la mañana y las diez de la noche. Además, estos *beeps* pueden presentarse de forma aleatoria o en intervalos de tiempo predeterminados, por ejemplo, cada 90 minutos. Todos estos datos se vuelcan a una plataforma para su análisis posterior. Se trata, pues, de un abordaje complementario a los procedimientos tradicionales de evaluación psicométrica de papel y lápiz en contextos más o menos artificiales y de corte más bien transversal y retrospectivo (Fonseca-Pedrero y Muñiz, 2017). La flexibilidad de los nuevos modelos psicométricos de análisis de redes pueden permitir la incorporación y estudio de este tipo de datos (Borsboom y Cramer, 2013; Fonseca-Pedrero, 2017, 2018), así como los modelos procedentes de la teoría de los sistemas dinámicos o la teoría del caos (Nelson, McGorry, Wichers, Wigman, y Hartmann, 2017).

No sabemos nada del futuro, pero se nos representa atractivo y excitante, una lucha sorda de fondo entre nuestra inteligencia de carbono y agua y la artificial del silicio. No sabemos si una de ellas vencerá a la otra, o se producirá la simbiosis, lo que está claro es que el silicio reclama un mayor rol en nuestras vidas y la evaluación psicométrica no es una excepción. Eso sí, la prueba del algodón, el árbitro, siempre será la validez, todas las fantasías sobre los avances tecnológicos pasan por demostrar que aportan mejoras en la medida del constructo evaluado, de lo contrario no dejarán de ser meros fuegos de artificio. Todos estos cambios y otros muchos que afectan a la evaluación obligarán a ir revisando y actualizando los diez pasos descritos, si bien lo esencial permanece: siempre habrá que aportar evidencias empíricas de la fiabilidad y validez, para garantizar que los instrumentos de medida evalúan de forma objetiva y rigurosa.

Agradecimientos

Los autores quieren agradecer los comentarios realizados por los profesores Alicia Pérez de Albéniz y Adriana Diez a una versión preliminar de este trabajo.

Esta investigación ha sido financiada por el Ministerio de Ciencia e Innovación de España (MICINN) (referencias: PSI2014-56114-P, PSI2017-85724-P) y por el Instituto Carlos III, Centro de Investigación Biomédica en Red de Salud Mental (CIBERSAM).

References

American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: Author.

Armayones, M., Boixadós, M., Gómez, B., Guillamón, N., Hernández, E., Nieto, R., Pousada, M., y Sara, B. (2015). Psicología 2.0: oportunidades y retos para el profesional de la psicología en el ámbito de la e-salud. *Papeles del Psicólogo*, *36*, 153-160.

Baldwin, D., Fowles, M., y Livingston, S. (2005). *Guidelines for constructed-response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Balluerka, N., Gorostiaga, A., Alonso-Arbiol, I., y Haranburu, M. (2007). La adaptación de instrumentos de medida de unas culturas a otras: una perspectiva práctica. *Psicothema*, *19*, 124-133.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and Practice, 18*, 5-12.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. En D. Bartram y R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances* (pp. 201-218). Chichester: Wiley.

Borsboom D., y Cramer, A.O.J. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology, 9,* 91-121. doi: 10.1146/annurev-clinpsy-050212-185608

Breithaupt, K. J., Mills, C. N., y Melican, G. J. (2006). Facing the opportunities of the future. En D. Bartram y R. K. Hambleton (Eds.), *Computer-based testing and the Internet* (pp. 219-251). Chichester: John Wiley and Sons.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research (2nd edition)*. New York: Guilford Press.

Calero, D., y Padilla, J. L. (2004). Técnicas psicométricas: los tests. En R. Fernández-Ballesteros (Ed.), *Evaluación psicológica: conceptos, métodos y estudio de casos* (pp. 323-355). Madrid: Pirámide.

Carretero, H., y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5*, 521-551.

Chernyshenko, O. S., y Stark, S. (2016). Mobile psychological assessment. En F. Drasgow (Ed.) (2016). *Technology and testing* (pp. 206-216). Nueva York: Routledge.

Clark, L. A., y Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*, 309-319.

Cuesta, M. (1996). Unidimensionalidad. En J. Muñiz (Ed.), *Psicometría* (pp. 239-292). Madrid: Universitas.

Dillman, D. A., Smyth, J. D., y Christian, L. M. (2009). *Internet, mail and mixed-mode surveys: The tailored design method*. Hoboken, NJ: Wiley.

Dorans N. J., y Cook, L. (2016). *Fairness in educational assessment and measurement*. New York: Taylor & Francis.

Downing, S. M. (2006). Twelve steps for effective test development. En S. M. Downing y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

Downing, S. M., y Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum Associates.

Drasgow, F. (Ed.) (2016). *Technology and testing*. Nueva York: Routledge.

Drasgow, F., Luecht, R. M., y Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement*. Westport, CT: ACE/Praeger.

Elosua, P. (2003). Sobre la validez de los tests. *Psicothema, 15*, 315-321.

Elosua, P., y Zumbo, B. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*, 896-901.

Erceg-Hurn, D. M., y Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*, 591-601. doi: 10.1037/0003-066X.63.7.591

Ferrando, P.J., y Anguiano, C. (2010). El análisis factorial como técnica de investigación en Psicología. *Papeles del Psicólogo, 31*, 18-33.

Fonseca Pedrero, E., y Debbané, M. (2017). Schizotypal traits and psychotic-like experiences during adolescence: An update. *Psicothema, 29*, 5-17. doi: 10.7334/psicothema2016.209

Fonseca-Pedrero, E. (2017). Análisis de redes: ¿una nueva forma de comprender la psicopatología? *Revista de Psiquiatria y Salud Mental*, *10*, 183-224. doi: 10.1016/j.rpsm.2017.06.004

Fonseca-Pedrero, E. (2018). Análisis de redes en psicología. *Papeles del Psicólogo, 39*, 1-12. https://doi.org/10.23923/pap.psicol2018.2852

Fonseca-Pedrero, E., y Muñiz, J. (2017). Quinta evaluación de tests editados en España: mirando hacia atrás, construyendo el futuro. *Papeles del Psicólogo*, *38*, 161-168. https://doi.org/10.23923/pap.psicol2017.2844

Gierl, M. J., y Haladyna, T. M. (Eds.) (2013). *Automatic item generation: Theory and practice*. Nueva York: Routledge.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., y Benítez, I. (2018). Differential Item Functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*, 104-109. doi: 10.7334/psicothema2017.183.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test item (3ª ed.)*. Hillsdale, NJ: LEA.

Haladyna, T. M., Downing, S. M., y Rodríguez, M. C. (2002). A review of multiple-choice item-writing guidelines. *Applied Measurement in Education, 15*, 309-334.

Haladyna, T. M., y Rodríguez, M. C. (2013). *Developing and validating test items*. London: Routledge.

Hambleton, R. K., Merenda, P. F., y Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. London: Lawrence Erlbaum Associates.

Hernández, A., Ponsoda, V., Muñiz, J., Prieto, G., y Elosua, P. (2016). Revisión del modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 37*, 192-197.

Hogan, T. P., y Murphy, G. (2007). Recommendations for preparing and scoring constructed-response items: What the experts say. *Applied Measurement in Education*, *20*, 427-441.

International Test Commission (2017). *The ITC Guidelines for translating and adapting Tests* (Second edition). Recuperado de http://www.InTestCom.org

Irwing, P. (Ed.) (2018). *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*. UK: John Wiley & Sons Ltd.

Insel, T.R. (2017). Digital phenotyping: Technology for a new science of behavior. *JAMA, 318*, 1215-1216. doi: 10.1001/jama.2017.11295

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement (4th edition)* (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kosinski, M., Stillwell, D., y Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behaviour. *Proceedings of the National Academy of Sciences*, *110*, 5802-5805. doi: 10.1073/pnas.1218772110

Krosnick, J. A., y Presser, S. (2010). Question and questionnaire design. En P. V. Marsden y J. D. Wright (Eds.), *Handbook of survey research* (2ª edición) (pp. 263-314). Bingley, Inglaterra: Emerald Group.

Lane, S., Raymond, M.R., y Haladyna, T. M. (2016). *Handbook of test development (2nd edition)*. New York, NY: Routledge.

Leong, T. L., Bartram, D., Cheung, F. M., Geisinger, K. F., e Illiescu, D. (Eds.) (2016). *The ITC International Handbook of Testing and Assessment*. New York: Oxford University Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 22*, 1-55.

Livingston, S. (2009). *Constructed-response test questions: Why we use them, how we score them*. Princeton, NJ: Educational Testing Service.

Lord, F. M., y Novick, M. R. (1968). *Statistical theories of mental test scores*. New York: Addison-Wesley.

Magno, C. (2009). Taxonomy of aptitude test items: A guide for item writers. *The International Journal of Educational and Psychological Assessment, 2*, 39-53.

Markovetz, A., Blaszkiewicz, K., Montag, C., Switala, C., y Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hipotheses*, *82*, 405-411. doi: 10.1016/j.mehy.2013.11.030

Markus, K., y Borsboom, D. (2013). *Frontiers of validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Martínez-Arias, R. (2018). Aproximaciones actuales a la validez de los test. En Academia de Psicología de España (Ed.), *Psicología para un mundo sostenible* (pp. 51-77). Madrid: Pirámide.

Miller, G. (2012). The smartphone psychology manifesto. *Perspectives on Psychological Science*, *7*, 221-237. doi: 10.1177/1745691612441215

Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., y Reininghaus, U. (2018). Experience sampling methodology in mental health research: New insights and technical developments. *World Psychiatry, 17*, 123-132. doi: 10.1002/wps.20513

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16*, 490-497.

Moreno, R., Martínez, R., y Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology, 2*, 65-72.

Moreno, R., Martínez, R., y Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*, 388-394. doi: 10.7334/psicothema2015.110

Moreno, R., Martínez, R. J., y Muñiz, J. (2018). Test item taxonomy based on functional criteria. *Frontiers in Psychology, 9*, 1175, 1-9. doi: 10.3389/fpsyg.2018.01175

Muñiz, J. (Ed.) (1996). *Psicometría*. Madrid: Universitas.

Muñiz, J. (1997). *Introducción a la teoría de respuesta a los ítems*. Madrid: Pirámide.

Muñiz, J. (2000). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, J. (2004). La validación de los tests. *Metodología de las Ciencias del Comportamiento, 5*, 121-141.

Muñiz, J. (2018). *Introducción a la psicometría*. Madrid: Pirámide.

Muñiz, J., y Bartram, D. (2007). Improving international tests and testing. *European Psychologist, 12*, 206-219.

Muñiz, J., Elosua, P., y Hambleton, R.K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. *Psicothema, 25*, 151-157. doi: 10.7334/psicothema2013.24

Muñiz, J., Fidalgo, A. M., García-Cueto, E., Martínez, R., y Moreno, R. (2005). *Análisis de los ítems*. Madrid: La Muralla.

Muñiz, J., y Fonseca, E. (2008). Construcción de instrumentos de medida para la evaluación universitaria. *Revista de Investigación en Educación, 5*, 13-25.

Muñiz, J., y Fonseca, E. (2017). *Construcción de instrumentos de medida en psicología*. Madrid: FOCAD. Consejo General de Psicología de España.

Nelson, B., McGorry, P. D., Wichers, M., Wigman, J. T., y Hartmann, J. A. (2017). Moving from static to dynamic models of the onset of mental disorder. *JAMA Psychiatry, 74*, 528-534. doi: 10.1001/jamapsychiatry.2017.0001

Norma UNE-ISO 10667 (2013). *Prestación de servicios de evaluación. Procedimientos y métodos para la evaluación de personas en entornos laborales y organizacionales* (partes 1 y 2). Madrid: AENOR.

Olea, J., Abad, F., y Barrada, J.R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del Psicólogo*, *31*, 94-107.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and others formats*. Boston: Kluwer Academic Publishers.

Osterlind, S.J., y Merz, W, R. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment, 2*, 133-147.

Parshall, C. G., Harmes, J. C., Davey, T., & Pashley, P. (2010). Innovative items for computerized testing. En W. J. van der Linden y C. A. Glas (Eds.), *Elements of adapting testin*g (pp. 215-230). Londres: Springer.

Phelps, R. (Ed.) (2005). *Defending standardized testing*. Londres: LEA.

Phelps, R. (Ed.) (2008). *Correcting fallacies about educational and psychological testing*. Washington: APA.

Prieto, G., y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo, 31*, 67-74.

Rauthmann, J. (2011). Not only item content but also item formats is important: Taxonomizing item format approaches. *Social Behavior and Personality, 39*, 119-128.

Rus-Calafell, M., Garety, P., Sason, E., Craig, T.J.K., y Valmaggia, L.R. (2018). Virtual reality in the assessment and treatment of psychosis: A systematic review of its utility, acceptability and effectiveness. *Psychological Medicine, 48*, 362-391. doi: 10.1017/S0033291717001945

Scalise, K., y Gifford, B. (2006). Computer-based assessment in e-learning: A framework for constructing "intermediate constraint" questions and tasks for technology platforms. *The Journal of Technology, Learning, and Assesment, 4*(6). Recuperado de http://www.jtla.org

Schmeiser, C. B., y Welch, C. (2006). Test development. En R. L. Brennan (Ed.), *Educational measurement (4th ed.)* (pp. 307-353). Westport, CT: American Council on Education/Praeger.

Shermis, M. D., y Burstein, J. (Eds.) (2013). *Handbook of automated essay evaluation. Current applications and new directions*. Nueva York: Routledge.

Sireci, S. (1998a). The construct of content validity. *Social Indicators Research, 45*, 83-117.

Sireci, S. (1998b). Gathering and analyzing content validity data. *Educational Assessment, 5,* 299-321.

Sireci, S., y Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*, 100-107. doi: 10.7334/psicothema2013.256

Sireci, S., y Zenisky, A. L. (2006). Innovative items format in computer-based testing: In pursuit of construct representation. En S. M. Downing y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 329-348). Hillsdale, NJ: LEA.

Sireci, S., y Zenisky, A. L. (2016). Computerized innovative item formats: Achievement and credentialing. En S. Lane, M. R. Raymond y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 313-334). Nueva York: Routledge.

Smith, G. T., Fischer, S., y Fister, S. M. (2003). Incremental validity principles in test construction. *Psychological Assessment, 15*, 467-477.

Smith, S. T. (2005). On construct validity: Issues of method measurement. *Psychological Assessment, 17*, 396-408.

Suárez, J., Pedrosa, I., Lozano, L., García-Cueto, E., Cuesta, M., y Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*, 149-158. doi: 10.7334/psicothema2018.33

Thurstone, L. L. (1927a). A law of comparative judgement. *Psychological Review, 34*, 273-286.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal Social Psychology, 21*, 384-400.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Trull, T. J., y Ebner-Priemer, U. W. (2013). Ambulatory assessment. *Annual Review of Clinical Psychology, 9*, 151-176. doi: 10.1146/annurev-clinpsy-050212-185510

van der Linden, W. (Ed.) (2016). *Handbook of item response theory* (3 volúmenes). Boca Ratón, FL: Chamman & Hall/CRC.

van Os, J., Delespaul, P., Wigman, J., Myin-Germays, I., y Wichers, M. (2013). Beyond DSM and ICD: Introducing precision diagnosis for psychiatry using momentary assessment technology. *World Psychiatry, 12*, 113-117. doi: 10.1002/wps.20046

Wells, C.S., y Faulkner-Bond, M. (2016). *Educational measurement. From foundations to future*. New York, NY: The Guilford Press.

Wetzel, E., Böhnke, J.R., y Brown, A. (2016). Response Biases. En F.T. Leong et al. (Eds.). *The ITC international handbook of testing and assessment* (pp. 349-363). New York: Oxford University Press.

Williamson, D.M., Bennett, R.E., Lazer, S., Berstein, J., Foltz, P.W., Landauer, T.K., Rubin, D.P., Way, W.P., y Sweeney, K. (2010). *Automated scoring for the assessment of common core standards*. Princeton, NJ: Educational Testing Service.

Williamson, D.M., Mislevy, R.J., y Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: LEA.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. En C. R. Rao y S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45-79). Amsterdam, Netherlands: Elsevier Science.