FOCAD
Formación Continuada a Distancia
Consejo General de Colegios Oficiales de Psicólogos

1

Sexta Edición Septiembre-Octubre 2009

CONSTRUCCIÓN DE INSTRUMENTOS DE MEDIDA EN PSICOLOGÍA

JOSÉ MUÑIZ Y
EDUARDO FONSECA-PEDRERO
Universidad de Oviedo

FORMACIÓN
continuada a distancia

ISSN 1989-3906

Formación Continuada a Distancia
Consejo General de Colegios Oficiales de Psicólogos

Contenido

FORMACIÓN
continuada a distancia

Documento base.

CONSTRUCCIÓN DE INSTRUMENTOS DE MEDIDA EN PSICOLOGÍA

Introducción

Muchas de las decisiones que toman los psicólogos en su quehacer diario están basadas en datos obtenidos mediante instrumentos de medida tales como tests, escalas o cuestionarios. Si estos instrumentos no tienen unas propiedades psicométricas adecuadas las inferencias hechas a partir de ellos serán incorrectas. Por ello, y dada la trascendencia de las decisiones y las consecuencias que a partir de ellos se derivan, tanto a nivel personal como social, es fundamental disponer de unos instrumentos de evaluación adecuados, construidos y validados mediante un proceso estandarizado, riguroso y objetivo y en función de unos estándares de calidad (Anastasi y Urbina, 1998; Kane, 2006; Messick, 1998; Muñiz, 1997b; Padilla, Gómez, Hidalgo y Muñiz, 2006; Padilla, Gómez, Hidalgo y Muñiz, 2007; Sireci, 2007; Sireci y Parker, 2006; Zumbo, 2007).

Los requisitos técnicos que debe cumplir un instrumento de medida aparecen bien documentados en la literatura psicométrica especializada (*American Educational Research Association, American Psychological Association y National Council on Measurement in Education*, 1999; Carretero-Dios y Pérez, 2005; Clark y Watson, 1995; Downing, 2006; Morales, Urosa y Blanco, 2003; Muñiz, 1996, 1997a, 2000; Nunnally y Bernstein, 1995; Schmeiser y Welch, 2006; Smith, Fischer y Fister, 2003; Wilson, 2005). No obstante, puede resultar de gran utilidad la descripción y síntesis de los pasos concretos que se deben seguir en todo proceso de construcción de un instrumento de medida, y éste será precisamente el objetivo central de este módulo.

Antes de utilizar un instrumento de medida, por muy popular que sea, el profesional debe preguntarse si se construyó siguiendo unos pasos estandarizados, rigurosos y objetivos, que garanticen su utilización en una población y contexto determinados. Se tiende a olvidar con bastante frecuencia que si el proceso de construcción de los instrumentos de evaluación se lleva cabo de forma defectuosa todas las inferencias que se obtengan a partir de las puntuaciones y la toma de decisiones que de ellas se deriven serán equivocadas e infundadas (Elosúa, 2003; Muñiz, 2004; Muñiz, Fidalgo, García-Cueto, Martínez y Moreno, 2005; Schmeiser y Welch, 2006).

La construcción de un instrumento de medida es un proceso complejo que se puede sintetizar en varios pasos, si bien éstos no son automáticos y universales, pudiendo variar en función del propósito del instrumento de medida (selección, diagnóstico, etc.), del modelo psicométrico utilizado (Teoría clásica, Teoría de Respuesta a los Ítems), del tipo de respuesta (selección o construcción), del formato de administración (lápiz y papel o informatizado), o del contexto de evaluación (diagnóstico, evaluación de rendimientos, etc.), por citar sólo algunos casos. Todo el proceso de construcción debe ser definido objetivamente siguiendo unos principios teóricos y métricos para así maximizar la validez de las inferencias hechas a partir de la prueba (Downing, 2006; Smith, 2005). Puede decirse que el proceso de validación ya comienza a fraguarse antes de la propia elaboración empírica del instrumento, pues todas las acciones que realicemos antes, durante y después permitirán recoger datos que ayuden a la interpretación de las puntuaciones (Elosua, 2003; Muñiz, 2004; Zumbo, 2007).

Pasos para la construcción de un instrumento de medida

A continuación se describen los pasos fundamentales a seguir para el desarrollo de un instrumento de medida en Psicología, parte de este documento ha sido tomado del trabajo previo de Muñiz y Fonseca-Pedrero (2008). En la Tabla 1 se recogen de forma esquemática las principales fases que se deben considerar en el proceso de construcción y validación de los instrumentos de medida, y a continuación se comenta cada una de ellas.

1. Marco general del instrumento de medida

Todo proceso de construcción de un instrumento de medida comienza por una justificación detallada y precisa de cuáles son las causas que motivan su construcción. Asimismo, hay que delimitar con claridad cuál es la varia-

ble objeto de medición, cuál va a ser el contexto de aplicación o circunstancias en las que se va a administrar el instrumento de evaluación, el tipo de aplicación (individual, colectiva), el formato de aplicación (lápiz y papel, informática), y qué decisiones se van a tomar a partir de las puntuaciones. Las causas que pueden llevar a la construcción de un instrumento de evaluación son diversas. Por ejemplo, un psicólogo puede decidir construir un instrumento porque no existe ningún otro para medir una determinada variable, porque los instrumentos existentes en el mercado presentan unas pésimas propiedades psicométricas, o simplemente porque no incorporan alguna faceta relevante para analizar dicho constructo. Los responsables de la construcción del instrumento de medida no sólo deben especificar el motivo por el cual quieren desarrollar un instrumento nuevo, 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 como por ejemplo: seleccionar, diagnosticar, clasificar, orientar, evaluar un dominio específico o incluso como método de *screening* (*American Educational Research Association* et al., 1999). Se debe dejar claro que las inferencias que se extraigan de las puntuaciones de un instrumento de medida 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). En suma, un test 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.

Tabla 1
Fases del proceso de construcción de instrumentos de medida

1. Marco general del instrumento de medida
- Justificación y motivación
- Contexto de aplicación
- Uso e interpretación de las puntuaciones

2. Definición operativa de la variable medida
- Definición operativa
- Definición sintáctica y semántica

3. Especificaciones del instrumento de medida
- Requerimientos de administración
- Tipo, número, longitud, formato, contenido y distribución de los ítems
- Especificaciones e instrucciones en la entrega del material
- Aspectos de seguridad

4. Construcción de los ítems
- Directrices para la construcción de ítems de elección múltiple
- Principios generales para la construcción de ítems

5. Producción, base de datos, normas de puntuación y corrección
- Composición
- Edición
- Puntuación y corrección

6. Estudio piloto cualitativo y cuantitativo
- Selección de la muestra piloto (cualitativo y cuantitativo)
- Análisis y resultados del estudio piloto (cualitativo y cuantitativo)
- Depuración, revisión, modificación o construcción de ítems
- Producción de una nueva versión del instrumento de medida

7. Selección de otros instrumentos de medida convergentes
- Obtener información convergente
- Utilizar pruebas ya validadas

8. **Estudio de campo**
- Selección y tamaño de la muestra y tipo de muestreo
- Administración del instrumento de medida
- Control de calidad y seguridad de la base de datos

9. **Estimación de las propiedades psicométricas**
- Análisis de ítems (cualitativo y cuantitativo)
- Dimensionalidad
- Estimación de la fiabilidad
- Obtención de evidencias de validez
- Tipificación

10.Versión definitiva, informe final y manual del instrumento de medida
- Prueba final propuesta
- Manual

2. Definición operativa de la variable medida

El objetivo esencial de esta segunda fase es la definición operativa, semántica y sintáctica de la variable medida, así como las facetas o dimensiones que la componen (*American Educational Research Association* et al., 1999; Carretero-Dios y Pérez, 2005; Lord y Novick, 1968; Wilson, 2005).

El constructo evaluado debe definirse en términos operativos, para que pueda ser medido de forma empírica (Muñiz, 2004). En este sentido, tan interesante puede ser definir cuidadosamente lo que es el constructo como lo que no es. La facilidad o dificultad de la definición operativa depende en cierta medida de la naturaleza de variable objeto de medición. Para llevar a cabo una definición operativa de la variable que nos interesa medir 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 la variable de medición (Calero y Padilla, 2004; Smith, 2005). Hay que evitar dejar fuera alguna característica o dominio relevante del constructo (infra-

estimación), así como ponderar en demasía una faceta o dominio (sobreestimación) (Smith et al., 2003). Una definición operativa y precisa del constructo influye de forma determinante en la posterior obtención de los diferentes tipos de evidencias, ayuda a especificar las conductas más representativas de la variable objeto de medición y facilita el proceso de construcción de ítems (Carretero-Dios y Pérez, 2005; Elosúa, 2003; Muñiz et al., 2005; Sireci, 1998; Smith, 2005).

No sólo es importante una definición operativa de la variable sino que también es preciso identificar y definir las facetas o dominios del mismo (definición semántica) y la relación que se establece entre ellas así como con otras variables de interés (definición sintáctica) (Lord y Novick, 1968). La variable objeto de medición no se encuentra aislada en el mundo, sino que está en relación o interacción (positiva y/o negativa) 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-Dios y Pérez, 2005; Muñiz, 2004; Smith, 2005).

3. Especificaciones del instrumento de medida

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 administración del instrumento de medida, 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 administración del instrumento de medida se refieren a cuál va a ser el soporte de administración (papel o informático), a qué tipo de aplicación se va a realizar (individual 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. Es importante llevar a cabo adaptaciones de acceso en aquellos participantes que no puedan desempeñar la tarea en igualdad de condiciones que el resto, por ejemplo disponer de una versión en *Braille* para una persona con deficiencia visual.

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 van a utilizar. Con respecto a este tema, no existen normas universales, todo dependerá de las circunstancias de aplicación, del propósito del constructor y de otras variables.

4. Construcción de los ítems

La construcción de los ítems constituye una de las etapas más cruciales dentro del proceso de construcción del instrumento de medida (Downing, 2006; Schmeiser y Welch, 2006). Los ítems son la materia prima, los ladrillos, a partir de la cual se forma un instrumento de evaluación, por lo que una construcción deficiente de los mismos, como no puede ser de otro modo, incidirá en las propiedades métricas finales del instrumento de medida y en las inferencias que se extraigan a partir de las puntuaciones (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). Todos los dominios de la variable de interés deben de estar igualmente representados, aproximadamente con el mismo número de ítems, a excepción de que se haya considerando un dominio más relevante dentro de la variable, y que por lo tanto, deba tener un mayor número de ítems, esto es, una mayor representación. Un muestreo erróneo del dominio objeto de evaluación sería una clara limitación a las inferencias que con posterioridad se dibujen a partir de los datos. Los ítems deben de ser heterogéneos y variados para así recoger una mayor variabilidad y representatividad de la variable de medida. Debe primar la claridad y la sencillez, se deben evitar tecnicismos, dobles negaciones, o enunciados excesivamente prolijos o ambiguos (Muñiz et al., 2005). Del mismo modo, los ítems deben ser comprensibles para la población a la cual va dirigido el instrumento de medida, 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 debido a factores espurios o irrelevantes, con la consiguiente merma de validez de la prueba.

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 tests (Balluerka, Gorostiaga, Alonso-Arbiol y Haranburu, 2007; Hambleton, Merenda y Spielberger, 2005; Muñiz y Bartram, 2007). En el caso de ítems originales han de seguirse las

directrices elaboradas para el desarrollo de ítems de elección múltiple (Downing y Haladyna, 2006; Haladyna, 2004; Haladyna et al., 2002; Moreno et al., 2006; Moreno et al., 2004; Muñiz et al., 2005).

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 sólo quedarán aquellos que ofrezcan mejores indicadores o garantías técnicas (sustantivas y métricas). Finalmente, para garantizar la validez de contenido de los ítems (Sireci, 1998) 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. 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.

Véanse en la **Ficha 1** las directrices para la elaboración de ítems de elección múltiple de Haladyna, Downing y Rodríguez (2002), adaptadas por Moreno, Martínez y Muñiz (2004).

5. Producción, base de datos, normas de puntuación y corrección

En esta fase se compone, se edita y se lleva a imprimir la primera versión del instrumento de medida, además de construir la base de datos con la claves de corrección. Este paso ha sido con frecuencia injustamente infraestimado, sin embargo es clave, pues el continente bien podría echar a perder el contenido. Buenos ítems pobremente editados dan como resultado un mal test, igual que los malos barriles pueden echar a perder los buenos vinos. Podemos haber construido un buen banco de ítems que de nada servirá si luego éstos se presentan de forma desorganizada, con errores tipográficos, o en un cuadernillo defectuoso. Uno de los errores más frecuentes entre los constructores de tests aficionados es utilizar fotocopias malamente grapadas, con la excusa de que sólo 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 la validez aparente. Es importante que el instrumento dé la impresión de medir de manera objetiva, rigurosa, fiable y válida la variable de interés. 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.

6. Estudio piloto cualitativo y cuantitativo

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 instrumento de evaluación 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 en voz alta diferentes aspectos relacionados con el instrumento de medida (p. ej., la detección de errores semánticos, gramaticales, el grado de comprensibilidad de los ítems, las posibles incongruencias semánticas, etc.). Los participantes en este pilotaje pueden ser (o no) similares a la población objeto de medición. Por su parte, el estudio piloto cuantitativo permite examinar las propiedades métricas 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 (p. ej., 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 algunos ítems son descartados mientras que otros son modificados. 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 deber ir destinadas a seleccionar los ítems con mayores garantías métricas que maximicen las propiedades finales del instrumento de eva-

luación. 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.

7. Selección de otros instrumentos de medida convergentes

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 evaluación es siempre obtener mayores 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 mediar la misma variable u otras diferentes (lo que anteriormente se había definido como definición sintáctica).

La decisión de qué instrumentos se deben utilizar complementariamente con el nuestro viene afectadas por cuestiones pragmáticas como las exigencias referidas al tiempo y al lugar. Evidentemente las exigencias de tiempo y las razones éticas no permiten administrar 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 sus propiedades psicométricas; b) que sean sencillos y de rápida administración y que conlleven un ahorro de tiempo; c) que tengan "coherencia" sustantiva de cara a establecer relaciones entre las variables.

8. Estudio de campo

En la fase del estudio de campo se incluye la selección de la muestra (tipo, tamaño y procedimiento), la administración del instrumento de medida a los participantes y el control de calidad y seguridad de la base de datos.

La representatividad y generalizabilidad de nuestros 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. Es recomendable que por cada ítem administrado tengamos al menos 5 ó 10 personas, si bien determinadas técnicas estadísticas pueden reclamar incluso más de cara a una buena estimación de los parámetros.

Las actividades relacionadas con la administración y el uso del instrumento de medida son cruciales durante el proceso de validación (Muñiz y Bartram, 2007; Muñiz et al., 2005). Cuando administramos 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. Por todo ello es recomendable elaborar unas pautas o directrices que permitan estandarizar la administración del instrumento de medida.

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. Frecuentemente cuando introducimos las puntuaciones de los participantes en una base de datos se pueden cometer multitud de errores, por ello es altamente recomendable comprobar de forma rigurosa que los datos se han introducido correctamente. 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. No obstante los mejores errores son los que no se cometen, así que hay que poner todos los medios para minimizar los errores a la hora de construir la base de datos.

9. Estimación de las propiedades psicométricas

Una vez administrado el instrumento de medida a la muestra de interés se procede al estudio de las propiedades psi-

cométricas del mismo: análisis de los ítems, estudio de la dimensionalidad, estimación de la fiabilidad, obtención de evidencias de validez y construcción de baremos.

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 de estar correctamente razonadas. En un primer lugar se deben analizar los ítems tanto a nivel cualitativo como a nivel cuantitativo. Para seleccionar los mejores ítems desde el punto de vista psicométrico se pueden 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). No se debe perder de vista que la finalidad del análisis psicométrico de los ítems no debe ser otro que 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 conocer 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 y el análisis de componentes principales son las técnicas más utilizadas para examinar la estructura interna, si bien no son las únicas (Cuesta, 1996). Una vez determinado la dimensionalidad 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 tests como de la teoría de respuesta a los ítems (Muñiz, 1997a, 2000). Posteriormente, y 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 establecen puntos de corte normativos. 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, 2006; Kline, 2005) y el funcionamiento diferencial de los ítems, por citar sólo tres casos (Muñiz et al., 2005).

10. Versión definitiva, informe final y manual del instrumento de medida

En último lugar, se procede a la elaboración la versión definitiva del instrumento de medida, se envía un informe de resultados a las partes interesadas, y se elabora el manual del mismo que permita su utilización a otras personas o instituciones interesadas. El manual de la prueba debe de recoger con todo detalle todas las características relevantes de la prueba. Finalmente y aunque sea la última fase, esto no quiere decir que el proceso de validación concluya aquí, posteriores estudios deberán seguir recogiendo evidencias de validez que permitan tomar decisiones fundadas a partir de las puntuaciones de los individuos. Asimismo conviene llevar a cabo una evaluación rigurosa y sistemática del instrumento elaborado, para lo cual puede utilizarse el Modelo de Evaluación de pruebas elaborado por la *European Federation of Professional Psychologists Associations* (EFPA), adaptado en España por Prieto y Muñiz (2000), publicado en la revista Papeles del Psicólogo. Véase este modelo en la **Ficha 2.**

A modo de conclusión

Se han descrito los diez pasos fundamentales que habría que seguir para desarrollar un instrumento de medida objetivo y riguroso para evaluar variables psicológicas. Estos pasos no se pueden abordar en profundidad desde un punto de vista técnico en un breve documento como éste, no se trata de eso, sino de poner a disposición de los profesionales una guía general que les permita obtener una visión panorámica de las actividades implicadas en el desarrollo de los instrumentos de medida. Se cita además la bibliografía especializada a la que pueden acudir aquellos profesionales interesados en profundizar en esta 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. Con demasiada frecuencia una buena prueba no genera los resultados esperados por hacer un mal uso de ella. Para todo lo relacionado con el uso de las pruebas en contextos nacionales e internacionales pueden consultarse los trabajos de Muñiz (1997b), Prieto y Muñiz (2000) y Muñiz y Bartram (2007).

REFERENCIAS

American Educational Research Association, American Psychological Association, y National Council on Measurement in Education (1999). *Standars for Educational and Psychological Testing*. Washington, DC: Author.

Anastasi, A., y Urbina, S. (1998). *Los tests psicológicos*. México: Prentice Hall.

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*, 124-133.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. 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-Dios, H., y Pérez, C. (2005). Normas para el desarrollo y revisión de estudios instrumentales. *International Journal of Clinical and Health Psychology, 5*, 521-551.

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.

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.

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

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.

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*(3), 309-334.

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

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

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2 ed.). New York: The Guilford Press.

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

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research 45,* 35-44.

Morales, P., Urosa, B., y Blanco, A. B. (2003). *Construcción de escalas de actitudes tipo Likert*. Madrid: La Muralla.

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

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16*(3), 490-497.

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

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

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

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

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

Nunnally, J. C., y Bernstein, I. J. (1995). *Teoría psicométrica*. México: McGraw Hill.

Padilla, J. L., Gómez, J., Hidalgo, M. D., y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema, 19*, 307-312.

Padilla, J. L., Gómez, J., Hidalgo, M. D., y Muñiz, J. (2007). Esquema conceptual y procedimientos para analizar la validez de las consecuencias del uso de los test. *Psicothema, 19*, 173-178.

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.
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.
Sireci, S. G. (1998). Gathering and analyzing content validity data. *Educational Assessment, 5*, 299-321.
Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher 36*, 477-481.
Sireci, S. G., y Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement: Issues and Practice 25*, 27-34.
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.
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.

Ficha 1.

DIRECTRICES PARA LA ELABORACIÓN DE ÍTEMS DE ELECCIÓN MÚLTIPLE

(Haladyna, Downing y Rodríguez, 2002)

Referidas al contenido

1. Cada ítem debería reflejar un contenido específico y una única conducta mental específica, tal como sea requerido en las especificaciones del test.
2. Base cada ítem en un contenido importante para el aprendizaje; evite contenidos triviales
3. Use material novedoso para evaluar el aprendizaje de alto nivel. Cuando los utilice en un ítem, parafrasee el lenguaje de los libros de texto, o el lenguaje utilizado durante la instrucción, para así evitar evaluar el mero recuerdo.
4. Mantenga el contenido de cada ítem independiente del contenido de otros ítems del test.
5. Al escribir ítems de elección múltiple, evite contenidos muy específicos o muy generales.
6. Evite ítems basados en opiniones
7. Evite ítems con trampas
8. Use un vocabulario sencillo para el grupo de estudiantes que están siendo evaluados

Referidas al formato

9. Del formato convencional de elección múltiple utilice la interrogación, completar frases, la mejor respuesta, elección alternativa, verdadero-falso, la verdadero-falso múltiple, emparejamiento, los conjuntos de items y los dependientes de contexto; sin embargo, evite el formato de elección múltiple complejo.
10. Construya el ítem de forma vertical, no horizontal

Referidas al estilo

11. Edite y ensaye los ítems
12. Use gramática, puntuación, mayúsculas y minúsculas y deletreo correctos
13. Minimice la cantidad de lectura en cada ítem

Redacción del enunciado

14. Asegúrese de que el sentido del enunciado resulta muy claro.
15. Incluya la idea central en el enunciado y no en las opciones.
16. Evite adornar el texto en exceso (palabrería excesiva)
17. Exprese el enunciado en términos positivos, y evite negativas tales como NO o EXCEPTO. Si usa términos negativos, hágalo con sumo cuidado y asegúrese que aparecen en mayúsculas o negritas.

Redacción de las opciones

18. Escriba tantas opciones efectivas como pueda, aunque la investigación sugiere que tres es lo adecuado.
19. Asegúrese que sólo una de esas opciones es la respuesta correcta.
20. Varíe la colocación de la respuesta correcta según el número de opciones
21. Coloque las opciones en un orden lógico o numérico
22. Construya las opciones independientes entre sí; no deben solaparse
23. Mantenga a las opciones homogéneas en contenido y estructura gramatical
24. Escriba las opciones con una longitud aproximadamente igual.
25. La opción *Ninguna de las anteriores* debe ser usada con mucho cuidado
26. Evite la opción *Todas las anteriores*.
27. Escriba las opciones en términos positivos; evite negativas tales como NO.

28. Evite dar pistas sobre la respuesta correcta, tales como
    a. Determinantes específicos como siempre, nunca, completamente y absolutamente
    b. Asociaciones por sonido similar y opciones idénticas o parecidas a términos del enunciado
    c. Inconsistencias gramaticales que indiquen al sujeto la elección correcta
    d. Opción correcta destacada
    e. Pares o tríos de opciones que indiquen al sujeto la opción correcta
    f. Opciones claramente absurdas o ridículas
29. Haga plausibles a todos los distractores
30. Use errores usuales de los estudiantes para escribir los distractores
31. Use el humor si es compatible con el profesor y con el ambiente de aprendizaje

Ficha 2.

MODELO PARA LA EVALUACIÓN DE LA CALIDAD DE LOS TESTS

1. Descripción general del test[1]

1.1. Nombre del test:
1.2. Nombre del test en su versión original (si la versión española es una adaptación):
1.3. Autor/es del test original:
1.4. Autor/es de la adaptación española:
1.5. Editor del test en su versión original:
1.6. Editor de la adaptación española:
1.7. Fecha de publicación del test original:
1.8. Fecha de publicación del test en su adaptación española:
1.9. Fecha de la última revisión del test en su adaptación española:
1.10. Clasifique el área general de la o las variables que pretende medir el test[2]

( ) Inteligencia
( ) Aptitudes
( ) Habilidades y Rendimiento académico
( ) Psicomotricidad
( ) Neuropsicología
( ) Personalidad
( ) Motivación
( ) Actitudes
( ) Intereses
( ) Escalas de Desarrollo
( ) Competencia Curricular
( ) Escalas Clínicas
( ) Potencial de Aprendizaje
( ) Otros (Indique cuál:.....)

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

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

1.12. Área de aplicación[3]

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

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

1.13. Formato de los ítems[4]:

( ) Respuesta libre
( ) Respuesta dicotómica (si/no, verdadero/falso, etc)
( ) Elección múltiple
( ) Tipo Likert
( ) Adjetivos bipolares
( ) Otro (Indique cuál:.....)

1.14. Número de ítems[5]:

1.15. Soporte[6]:

( ) Administración oral
( ) Papel y lápiz
( ) Manipulativo
( ) Informatizado
( ) Otro (Indique cuál:.....)

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

( ) Ninguna
( ) Entrenamiento y Acreditación especifica*
( ) Nivel A[7]
( ) Nivel B
( ) Nivel C
( ) Otra (Indique cuál:.....)

**Indique el nombre de la institución que lleva a cabo la acreditación:*

1.17. Descripción de las poblaciones a las que el test es aplicable (especifique el rango de edad, nivel educativo, etc., y si el test es aplicable en ciertas poblaciones específicas: minorías étnicas, discapacitados, grupos clínicos, etc.):

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

1.19. Procedimiento de corrección:

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

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

1.20. Puntuaciones: (Describa el procedimiento para obtener las puntuaciones directas).

1.21. Transformación de las puntuaciones:

( ) Característica no aplicable para este instrumento
( ) Normalizada
( ) No normalizada

1.22. Escalas utilizadas:

( ) Centiles
( ) Puntuaciones típicas
( ) Cocientes de desviación
( ) Eneatipos
( ) Decatipos
( ) T (*Media 50 y desviación típica 10)*
( ) S (*Media 50 y desviación típica 20)*
( ) Otra (Indique cuál:.....)

1.23. Posibilidad de obtener informes automatizados:

( ) No
( ) Si*

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

1.24. El editor ofrece un servicio para la corrección y/o elaboración de informes:

( ) No
( ) Si

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

En aplicación individual:.....

En aplicación colectiva:.....

1.26. Documentación aportada por el editor:

( ) Manual
( ) Libros o artículos complementarios
( ) Disketes/CD
( ) Otra (Indique cuál:.....)

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

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

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

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

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

2. Valoración de las características del test

2.1. Calidad de los materiales del test (objetos, material impreso o *software*):

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Impresión y presentación de gran calidad, software muy atractivo y eficiente, etc.)

2.2. Calidad de la documentación aportada:

* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción muy clara y completa de las características técnicas, fundamentada en abundantes datos y referencias)

2.3. Fundamentación teórica:

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

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

( ) Característica no aplicable para este instrumento
( ) No se aporta información en la documentación
* ( ) Inadecuada
** ( ) Adecuada pero con algunas carencias
*** ( ) Adecuada
**** ( ) Buena
***** ( ) Excelente (Descripción precisa del procedimiento de traducción, de la adaptación de los ítems a la cultura española, de los estudios de equivalencia con la versión original, utilización de la normativa de la International Test Commission, etc.).

2.5. Calidad de las instrucciones:

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

2.6. Facilidad para comprender la tarea:

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

2.7. Facilidad para registrar las respuestas:

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

2.8. Calidad de los ítems (aspectos formales):

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

2.9. Análisis de los ítems

2.9.1 Datos sobre el análisis de los ítems:

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

2.10. Validez

2.10.1. Validez de contenido[8]:

2.10.1.1. Calidad de la representación del contenido o dominio:

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

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

2.10.1.2. Consultas a expertos[9]:

( ) No se aporta información en la documentación

* ( ) No se ha consultado a expertos sobre la representación del contenido

** ( ) Se ha consultado de manera informal a un pequeño número de expertos

*** ( ) Se ha consultado a un pequeño número de expertos mediante un procedimiento sistematizado (N<10)

**** ( ) Se ha consultado a un número moderado de expertos mediante un procedimiento sistematizado (10≤N≤30)

***** ( ) Se ha consultado a un amplio número de expertos mediante un un procedimiento sistematizado (N>30)

2.10.2. Validez de constructo:

2.10.2.1. Diseños empleados[10]:

( ) No se aporta información en la documentación

( ) Correlaciones con otros tests

( ) Diferencias entre grupos

( ) Matriz multirasgo-multimétodo

( ) Análisis factorial exploratorio

( ) Análisis factorial confirmatorio

( ) Diseños experimentales

( ) Otros (Indique cuales:.....).

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

( ) No se aporta información en la documentación

* ( ) Un estudio con una muestra pequeña (N<200)

** ( ) Un estudio con una muestra moderada (200≤N≤500)

*** ( ) Un estudio con una muestra grande (N>500)

**** ( ) Varios estudios con muestras de tamaño moderado

***** ( ) Varios estudios con muestras grandes

2.10.2.3. Procedimiento de selección de las muestras*:

( ) No se aporta información en la documentación

( ) Incidental

( ) Aleatorio

**Describa brevemente el procedimiento de selección.*

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

( ) No se aporta información en la documentación

* ( ) Inadecuada (r<0.25)

** ( ) Adecuada pero con algunas carencias (0.25≤r<0.40)

*** ( ) Adecuada (0.40≤r<0.50)

**** ( ) Buena (0.50≤r<0.60)

***** ( ) Excelente (r≥0.60)

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

[10] Puede marcar más de una opción.

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

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

2.10.2.6. Datos sobre el sesgo de los ítems:

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

2.10.3. Validez predictiva

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

2.10.3.1. Diseño de selección del criterio[11]:

( ) Concurrente
( ) Predictivo
( ) Retrospectivo

2.10.3.2. Tamaño de las muestras en la validación predictiva:

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

2.10.3.3. Procedimiento de selección de las muestras*:

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

*Describa brevemente el procedimiento de selección.

[11] Puede marcar más de una opción.

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

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

2.10.4. Comentarios sobre la validez en general:

2.11. Fiabilidad

2.11.1. Datos aportados sobre la fiabilidad:

( ) Un único coeficiente de fiabilidad
( ) Un único error típico de medida
( ) Coeficientes de fiabilidad para diferentes grupos de sujetos
( ) Error típico de medida para diferentes grupos de sujetos

2.11.2. Equivalencia (Formas paralelas):

2.11.2.1. Tamaño de las muestras en los estudios de equivalencia:

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

2.11.2.2. Mediana de los coeficientes de equivalencia:

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

2.11.3. Consistencia interna

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

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

2.11.3.2. Mediana de los coeficientes de consistencia:

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

2.11.4. Estabilidad (Test-Retest)

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

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

2.11.4.2. Mediana de los coeficientes de estabilidad:

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

2.11. 5 Comentarios sobre la fiabilidad en general:

2.12. Baremos

2.12.1. Calidad de las normas:

( ) No se aporta información en la documentación
* ( ) Un baremo que no es aplicable a la población objetivo
** ( ) Un baremo aplicable a la población objetivo con cierta precaución
*** ( ) Un baremo adecuado para la población objetivo
**** ( ) Varios baremos dirigidos a diversos estratos poblacionales
***** ( ) Amplio rango de baremos en función de la edad, el sexo, el nivel cultural y otras características relevantes

2.12.2. Tamaño de las muestras[13]:

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

2.12.3. Procedimiento de selección de las muestras*:

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

*Describa brevemente el procedimiento de selección.

2.12.4. Comentarios sobre los baremos

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

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

3. Valoración global del test

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

3.21. A modo de resumen, rellene las Tablas 1 y 2.

La Tabla 1 incluye algunos datos descriptivos del test.

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

Tabla 1
Descripción del test

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

Tabla 2
Valoración del test

Característica	Apartados	Valoración
Materiales y documentación	2.1 y 2.2	
Fundamentación teórica	2.3	
Adaptación	2.4	
Análisis de ítems	2.9	
Validez de contenido	2.10.1	
Validez de constructo	2.10.2	
Análisis del sesgo	2.10.2.6	
Validez predictiva	2.10.3	
Fiabilidad: equivalencia	2.11.2	
Fiabilidad: consistencia interna	2.11.3	
Fiabilidad: estabilidad	2.11.4	
Baremos	2.12	