© International Journal of Clinical and Health Psychology ISSN 1697-2600
2007, Vol. 7, No. 3, pp. 863-882

Standards for the development and review of instrumental studies: Considerations about test selection in psychological research

Hugo Carretero-Dios[1] y Cristino Pérez *(Universidad de Granada, España)*

(Recibido 7 de mayo 2007 / Received May 7, 2007)
(Aceptado 11 de junio 2007 / Accepted June 7, 2007)

RESUMEN. En este trabajo se discuten los criterios a tener en cuenta a la hora de seleccionar tests de evaluación psicológica en un contexto de investigación. Tradicionalmente la atención se ha centrado y se centra sobre las fases que deben regir todo proceso de construcción/adaptación de tests. Estándares internacionalmente aceptados sirven para dirigir este trabajo, y la comunidad científica coincide en la importancia de éstos. No obstante, y más allá de cualquier proceso de construcción/adaptación, el hecho es que el uso de tests es una constante en la investigación psicológica, y una adecuada selección de las pruebas resulta un asunto de vital importancia. Por ello, y esquematizando en primer lugar los criterios que deben guiar la construcción/adaptación de test, en este estudio teórico se desarrollan unas directrices generales a tener en cuenta a la hora de seleccionar tests para efectuar una investigación psicológica. La información va a presentarse organizada en un total de seis apartados, cada uno de los cuales corresponde a una fase distinta dentro del proceso de creación de tests: a) delimitación conceptual del constructo objeto de evaluación; b) información sobre la construcción y evaluación cualitativa de ítems; c) resultados del análisis estadístico de los ítems; d) evidencias empíricas de la estructura interna de la prueba; e) resultados de la estimación de la fiabilidad; y f) evidencias externas de la validez de la puntuaciones. Se finaliza el trabajo reflexionando sobre el alcance de las directrices propuestas y sobre la importancia de seleccionar bajo criterios claros los tests que vayan a usarse en una investigación.

PALABRAS CLAVE: Normas para la revisión de estudios instrumentales. Construcción de tests. Adaptación de tests. Selección de tests. Estudio Teórico.

ABSTRACT. This paper discusses the criteria that should be considered when selecting psychological assessment tests in a research context. Traditionally attention has focused – and still does – on the stages that must govern any process of test construction/adaptation. This work is guided by internationally accepted standards, whose scientific importance is agreed by the scientific community. However, beyond any construction/adaptation process, the use of tests is a constant feature of psychological research, so it is of vital importance to select the tests adequately. For this reason, in this theoretical study we provide a summary of the criteria that should guide test construction/adaptation as well as some general guidelines to consider when selecting tests for psychological research. The information

[1] Correspondencia: Facultad de Psicología. Universidad de Granada. Campus Cartuja. 18071 Granada (España). E-mail: hugocd@ugr.es

presented is organized into six sections, each of which corresponds to a different stage in the process of test creation: a) conceptual definition of the construct to assess; b) information about item construction and qualitative assessment; c) results of the statistical analysis of the items; d) empirical evidence of the internal structure of the test; e) results of the reliability estimation; and f) external evidence of score validity. The study ends with a reflection on the scope of the proposed guidelines and the importance of using clear criteria to select the tests used in research.

KEY WORDS. Standards for the review of instrumental studies. Test construction. Test adaptation. Test selection. Theoretical study.

RESUMO. Neste trabalho discutem-se os critérios a considerar na hora de seleccionar os testes de avaliação psicológica num contexto de investigação. Tradicionalmente a atenção tem-se centrado e centra-se sobre as fases que devem orientar todo o processo de construção / adaptação de testes. Critérios standards internacionalmente aceites servem para dirigir este trabalho, e a comunidade científica coincide na importância que lhes atribui. No entanto, e para além de qualquer processo de construção/adaptação, o facto é que o uso de testes é uma constante na investigação psicológica, e uma selecção adequada das provas torna-se num assunto de grande importância. Por isso, e esquematizando em primeiro lugar os criterios que devem guiar a construção / adaptação de testes, neste estudo teórico desenvolvem-se algumas directrizes gerais a ter em consideração na altura de seleccionar testes para efectuar una investigação psicológica. A informação apresentada está organizada num total de seis pontos, cada um dos quais corresponde a uma fase distinta dentro do processo de criação de testes: a) delimitação conceptual do construto objecto de avaliação; b) informação sobre a construção e avaliação qualitativa dos itens; c) resultados da análise estatística dos itens; d) evidências empíricas da estrutura interna da prova; e) resultados da estimação da fiabilidade; f) evidências externas da validade das pontuações. O trabalho termina com reflexões sobre o alcance das directrizes propostas e sobre a importância de seleccionar sob critérios claros os testes que venham a usar-se numa investigação.

PALAVRAS CHAVE. Normas para a revisão de estudos instrumentais. Construção de testes. Adaptação de testes. Selecção de testes. Estudo teórico.

Introducción

En la investigación psicológica actual, el uso de instrumentos o herramientas, tales como las computadoras, sistemas de registro, instrumentos de medida, etc. supone una característica definitoria de la propia investigación. De hecho, sin dichos instrumentos, la investigación científica, tal y como actualmente se conoce, sería imposible, requiriéndose un análisis cuidadoso y cíclico de éstos y de su influencia sobre los resultados de investigación (Sturm y Ash, 2005). Dentro de los múltiples y variados instrumentos que pueden ser empleados en un contexto de investigación psicológica, la utilización de tests de evaluación es algo más que frecuente, sin olvidar igualmente lo generalizado que está el uso de éstos dentro de la práctica profesional que genera la Psicología como disciplina (Muñiz *et al.*, 2001).

El hecho es que los psicólogos trabajan con fenómenos no directamente observables, los cuales pretenden medirse, y para lo que se usan aproximaciones indirectas.

De esta forma, su medición está condicionada a la obtención de indicadores observables, y es aquí donde cabría resaltar la importancia de las respuestas generadas ante un test como material esencial para los psicólogos. Estas respuestas sirven para generar puntuaciones que finalmente sirven para múltiples objetivos, tales como la puesta a prueba de teorías, la toma de decisiones acerca de la efectividad de un tratamiento psicológico, la verificación experimental del impacto de una o varias variables independientes, etc. Así pues, las puntuaciones que se obtienen a partir de los tests tienen implicaciones de suma importancia sobre el resultado final de cualquier investigación que haga uso de ellos, al igual que sobre las consecuencias aplicadas que se derivan de la actividad de los profesionales, y que en su día a día toman decisiones en función del resultado generado por dichos tests (Padilla, Gómez, Hidalgo y Muñiz, 2006, 2007).

Los *Standars for Educational and Psychological Testing* (AERA, APA y NCME, 1999) intentan dar respuestas a las problemáticas que se generan en el proceso de creación/adaptación y uso de tests. Los investigadores que dirigen sus esfuerzos a la creación/adaptación de tests tienen en estos estándares una referencia que guía su trabajo y unifica criterios de valoración. De hecho, el debate sobre los estándares está continuamente abierto (Linn, 2006) y las sugerencias para su perfeccionamiento y mejora son objeto de publicaciones (Koretz, 2006; Wise, 2006), lo que lleva a contar con unas directrices que responden a las exigencias de cada momento y que son una fuente de indudable valor para el perfeccionamiento del trabajo llevado a cabo por los investigadores. Sin embargo, y a pesar de la importancia de estos estándares, su uso está más relacionado a los investigadores que centran sus esfuerzos en los denominados estudios instrumentales, consistentes en el "desarrollo de pruebas y aparatos, incluyendo tanto el diseño (o adaptación) como el estudio de las propiedades psicométricas de los mismos" (Montero y León, 2005, p. 124). Esto no significa, no obstante, que de los estándares no puedan derivarse importantes implicaciones para aquellos investigadores que hacen uso de tests para objetivos no vinculados a los que son propios de los estudios instrumentales.

En la actualidad, todo investigador que se disponga a hacer un estudio para el que requiera hacer uso de tests, cuenta, en la mayoría de las ocasiones, con varias alternativas posibles con objetivos similares. En estos casos, y dada la influencia directa que el uso de un instrumento u otro va a tener sobre los resultados finales, la selección razonada de los tests debe ser un criterio necesario a no obviar, salvando pues justificaciones centradas, por ejemplo, en el acceso más fácil a un test que a otro, o cualquier otra razón que se aleje de lo que se supone un esquema de acción científica. Ocurre, no obstante, que la importancia supuesta de trabajar con un instrumento u otro parece no tener su reflejo en las publicaciones. Así, más que poder concluir que la selección de tests está gobernada por criterios no científicos, habría que decir que en muchos de los casos existe una ausencia de información sobre las razones que han llevado a emplearlos. Por ejemplo, Hogan y Agnello (2004) pusieron de manifiesto que sólo el 55% de 696 publicaciones científicas donde se hacía uso de tests proporcionaba alguna evidencia sobre la validez de las puntuaciones generadas por los instrumentos usados. Además, y tal y como puede comprobarse fácilmente, una gran mayoría de autores justifican su uso refugiándose en la mera notificación de los valores numéricos relativos a los coeficientes de fiabilidad y validez de los mismos. Con esta forma de proceder se salva toda responsabilidad en relación con la selección y utilización de las pruebas, aún a sabiendas de que al final de todo proceso de investigación, la responsabilidad de los resultados obtenidos no es de los creadores de las pruebas, sino de los autores de estas investigaciones.

Lo indicado anteriormente se agrava aún más por el hecho incontestable de que la mayoría de las pruebas publicadas -tanto en revistas de toda índole como por empresas especializadas en su construcción y comercialización-, adolecen de los mínimos exigidos en los *Standars for Educational and Psychological Testing* (AERA, APA y NCME, 1999). Se construyen muchas pruebas y muy diversas, a veces por especialistas y, otras muchas, por investigadores muy alejados de este ámbito. Un alto porcentaje de ellas se desarrollan a colación de necesidades de investigación muy específicas, siendo mínimo el conocimiento psicométrico al que puede llegarse a través de su uso. Así, al revisar muchos de los tests publicados, se puede comprobar que tras intuir, no sin dificultad, la definición de la variable objeto de evaluación, de inmediato se observan los valores de los coeficientes de fiabilidad y validez, no encontrándose la información necesaria sobre el procedimiento seguido, sobre su justificación, o acerca de otros aspectos a tener en cuenta al enjuiciar la calidad de cualquier test.

Suponer que un test, por el hecho de estar publicado, cumple con los requisitos científicos mínimos, resulta, cuanto menos, arriesgado. Así, y en relación con las pruebas publicadas, se quiere llamar la atención acerca de cómo, en general, se ofrece nada o muy poca información sobre el proceso de edición de los ítems, la justificación del número de ítems necesario para representar al constructo evaluado, la adecuada representación de las dimensiones a través de los ítems considerados, etc. Además, y en este caso destacando su ausencia más si cabe, hay que notar como existe una carencia casi absoluta de datos en cuanto a los controles aplicados, tanto cualitativos como cuantitativos, para garantizar a la calidad de los ítems, y que hagan referencia a los criterios de eliminación, sustitución, modificación, etc. de éstos.

Son también numerosos los problemas que se aprecian en lo referente a la idoneidad de los procedimientos seguidos para el cálculo de la fiabilidad, o en lo que concierne a las estrategias empleadas para mostrar las evidencias de validez de las puntuaciones de la prueba. Valga a modo de ejemplo alguno de los resultados asociados al ya clásico test *PMA: Aptitudes Mentales Primarias* de Thurstone y Thurstone, en su adaptación española (TEA, 1996). Cuando se ofrece el coeficiente de fiabilidad del factor numérico, el dato es de 0,99. Este resultado, alarmante a todas luces, debería ser una llamada de atención para los investigadores y psicólogos aplicados que eligen dicha prueba para su uso. Así, si se indaga sobre las razones de este inesperado e inaceptable valor del coeficiente de fiabilidad, se puede apreciar como fue el uso de un procedimiento inadecuado el que elevó hasta la cuantía comentada el valor de dicho coeficiente: se trata de ítems de rapidez y para el cálculo del mencionado coeficiente se siguió la estrategia de la división del test en dos mitades (TEA, 1996, p. 13).

En cuanto a las evidencias de validez de las puntuaciones de las pruebas en sí, es imprescindible que los autores de las mismas destaquen y justifiquen una definición sintáctica en la que relacionen, con más o menos firmeza, las conexiones del constructo objeto de medida con otros constructos constitutivos de una red conceptual bien asentada o, en última instancia, con indicadores empíricos que posibiliten la posterior puesta en marcha de las pertinentes estrategias de confirmación. Sin estos previos, las evidencias de validez de las distintas pruebas no dejan de ser resultados estadísticos aislados, sin forma de poder asignarles u otorgarles un significado o utilidad, y que finalmente sólo sirven para ocultar las deficiencias de un proceso de construcción deficiente.

Lo hasta ahora apuntado, que duda cabe, debe resultar alarmante, teniendo en cuenta la importancia que tiene el uso de tests en la investigación psicológica. Además de la influencia directa sobre los resultados, habría que hablar del alcance o uso generalizado de

los tests en la mayoría de las publicaciones. Por ejemplo, en un medio como el presente, el *International Journal of Clinical and Health Psychology*, el 100% de los estudios originales publicados durante 2007 han hecho uso de tests para el desarrollo de la investigación. Por ello, se hace necesario tener en cuenta ciertos criterios para la selección de los tests antes de proceder a su uso, considerando pues que la mera publicación de un test no garantiza su calidad. No obstante, y dentro de un contexto delimitado por los artículos científicos, el debate no estaría centrado en la calidad científica de las medidas usadas, ya que en el ámbito comentado se entiende que dicha calidad estaría presente como necesidad básica de toda investigación. La discusión sería otra: ¿los tests usados en las investigaciones publicadas han sido seleccionados bajos unos criterios de decisión objetivos?, ¿se han considerado los aspectos diferenciales que presentan instrumentos distintos construidos bajo objetivos de evaluación similares?, ¿los criterios empleados permiten una mayor seguridad a la hora de concluir que la herramienta empleada es la mejor opción de entre todas las disponibles?

El objetivo del presente trabajo es proponer unas directrices generales que guíen la selección de tests en un contexto de investigación, aunque sin olvidar que muchos de los criterios propuestos deberían ser igualmente tenidos en cuenta por los profesionales aplicados. Como resulta lógico, dicha selección debe estar regida por el hecho de poder garantizar que el instrumento utilizado cumple con unas propiedades científicas mínimas, y esto significaría que se han seguido las normas internacionalmente aceptadas para la construcción de tests (AERA *et al.*, 1999). Recientemente se discutieron dichas normas, y se propusieron unas pautas básicas para el desarrollo y revisión de estudios instrumentales (Carretero-Dios y Pérez, 2005). Sobre dichas pautas (Anexo 1) se asienta el presente trabajo, haciendo ahora hincapié en el ejercicio responsable de toma de decisiones que debe hacer todo investigador que pretenda hacer uso de tests ya disponibles, y por lo tanto sometidos a análisis científico previo. Este estudio se inserta dentro de una marco más general que se ocupa de la estandarización de los procederes científicos presentes en sus distintos ámbitos de acción (Blanton y Jaccard, 2006; Botella y Gambara, 2006; Ramos-Álvarez, Valdés-Conroy y Catena, 2006).

Criterios para la selección de tests

Las directrices que van a presentarse a continuación van a tener un contexto de aplicación concreto, y que no es otro que aquel donde se haga necesario el uso de instrumentos objetivos de medida, ya sea en un área aplicada o de investigación, e independientemente de la categoría donde puedan encuadrarse dichos instrumentos: autoinformes, cuestionarios, tests psicológicos en general, etc. Lo que vendría a defenderse es que siempre que para el desarrollo de un trabajo se necesite evaluar un constructo a partir de una prueba desarrollada para tal fin, resultaría conveniente hacer uso de unas directrices generales para una óptima selección de entre los instrumentos disponibles, así como para detectar alguna deficiencia en los mismos. Téngase en cuenta, no obstante, que el contenido del presente trabajo va a estar influido por el medio donde es publicado, y por la intención de que su contenido sea significativo desde el inicio para la audiencia que dicho medio tiene. Esto tendrá consecuencias sobre los ejemplos que se usen y sobre las publicaciones a las que se acuda para ejemplificar algunas cuestiones.

Para este trabajo, el término *constructo* se entiende como “el concepto, atributo o variable objeto de medición. Los constructos pueden diferir en su grado de especificidad desde un nivel molar, con variables latentes tales la responsabilidad, hasta un nivel

molecular con variables que requieren un menor nivel de inferencia tales como la ingesta de alcohol o la agresión física" (Haynes, Richard y Kubany, 1995, p. 239). A pesar de esta definición, debe tenerse en cuenta que las variables objeto de evaluación dentro de la Psicología son fundamentalmente constructos que hacen referencia a atributos de carácter general de las personas evaluadas, y para los que se requiere una aproximación a su definición que permita tratar con un nivel de especificidad del constructo mucho más concreto. Esto, tal y como se verá a continuación, tiene importantes implicaciones para la selección de los tests, y en concreto para la fase en la que se tiene que analizar la definición aportada de los constructos evaluados.

A continuación se presentan las recomendaciones a tener en cuenta para la selección de tests. Estas recomendaciones van a desarrollarse en seis apartados, cada uno de los cuales corresponde a una etapa crucial dentro del proceso de construcción/adaptación de tests (véase su desarrollo en Carretero-Dios y Pérez, 2005 o un resumen en Anexo 1). Por ello, el investigador debería hacer un análisis de dichas etapas, y delimitar cómo éstas quedan reflejadas en los instrumentos con los que pretenda trabajar. La estructuración del trabajo parte del supuesto de que la persona encargada de la selección de un test ha considerado, y en primer lugar, el objetivo de evaluación y el para qué de ésta. Por ello, la exposición se centra a partir del momento en el que el involucrado en la selección de un test se encuentra con distintas alternativas posibles para un mismo objetivo de evaluación y uso previsto de las puntuaciones. De esta forma, los apartados que van a guiar la presentación son: a) delimitación conceptual del constructo objeto de evaluación; b) información sobre la construcción y evaluación cualitativa de ítems; c) resultados del análisis estadístico de los ítems; d) evidencias empíricas de la estructura interna de la prueba; e) resultados de la estimación de la fiabilidad; y f) evidencias externas de la validez de la puntuaciones.

Delimitación conceptual del constructo objeto de evaluación

Resulta obvio apuntar que a la hora de seleccionar un test, el interesado debe tener claro qué se evalúa. La respuesta a la pregunta qué es lo que se evalúa no puede contentarse con la corroboración de que aparece una etiqueta indicativa de su objetivo insertada en el nombre que defina al test, como por ejemplo depresión, ansiedad social, búsqueda de sensaciones, etc. Téngase en cuenta que la parte más importante para la construcción de un instrumento que acabe presentando las adecuadas garantías psicométricas es partir de una definición completa y exhaustiva del constructo evaluado (Nunnally y Berstein, 1995). De hecho, de una definición ambigua e inespecífica se derivan ítems ambiguos e inespecíficos, y por ende puntuaciones no concretas y cuyo significado final resultaría difícil de concretar.

En la actualidad existe gran cantidad de tests que tienen como objetivo de evaluación una misma etiqueta, lo que no significa que un mismo concepto. Detrás de una misma etiqueta se esconden aproximaciones conceptuales distintas, definiciones distintas y, por lo tanto, objetivos de medición distintos, aunque no siempre explícitos. A la hora de decidir qué test seleccionar se debe consultar, en el caso de que se encuentre disponible, la definición ofrecida sobre el constructo evaluado. El investigador que se disponga a realizar un estudio para el que le resulte esencial trabajar con unos tests concretos tendrá unos objetivos específicos de investigación y, por ello, para cubrir éstos deberá cerciorarse de que los instrumentos que escoja se centran en sus conceptos de interés más allá de una etiqueta común a través de instrumentos.

Al adoptar este procedimiento, es decir, analizar las definiciones ofrecidas por los creadores de pruebas, el encargado de esta selección podrá constatar que resulta más común de lo que cabría esperar encontrar estudios donde se presenta una escala que se asienta sobre una delimitación conceptual inespecífica del constructo evaluado. La definición se

suele basar en una afirmación genérica de lo que el constructo es y que a su vez se basa en otros constructos igualmente no delimitados. Sin embargo, esta forma de proceder se aleja de las recomendaciones presentes en los trabajos especializados (véase Murphy y Davidshofer, 1994; Walsh, 1995).

Al autor o autores de una prueba debe exigírsele que proporcionen una delimitación concreta de los componentes o facetas que definen su constructo objeto de evaluación y que a su vez concreten operacionalmente a lo que se refiere cada uno de estos componentes, es decir, se debe facilitar lo que ha venido a denominarse como definición semántica de la variable (Lord y Novick, 1968). Debido a la complejidad de los constructos psicológicos, la presentación pormenorizada y justificada de esta definición sobrepasaría lo que son los límites al uso de un artículo de investigación. A pesar de esto, al menos en el trabajo debe aparecer una referencia que permita consultar de manera detallada la definición ofrecida, y donde el espacio no sea una limitación (por ejemplo, el manual del test, un libro centrado en el constructo evaluado, etc.). La persona encargada de seleccionar un test debe tener como principio que la prueba que no presente claramente los elementos diferenciadores del constructo evaluado, que no recoja la variedad de manifestaciones operativas de éste, o que no concrete claramente sus componentes, va a provocar un proceso de construcción/adaptación impreciso y caracterizado por unas deficientes evidencias de validez de contenido (Downing y Haladyna, 2004; Haynes *et al.*, 1995; Smith, 2005).

A la hora de seleccionar un test, se tendrían mayores garantías acerca de que se ha efectuado una adecuada definición operativa del constructo si se pusiera de manifiesto que los autores han seguido las recomendaciones existentes sobre cómo presentar dicha definición, y en concreto, que hacen uso de una tabla de especificaciones del test donde se inserte toda la información de interés del constructo evaluado (Osterlind, 1989). Así, junto a la presentación pormenorizada de la definición del constructo, debe corroborarse si dicha definición ha sido sometida a una revisión por parte de expertos antes de la creación de ítems propiamente dicha (véase Carretero-Dios, Pérez y Buela-Casal, 2006). Aunque es común no hacer uso de esta valoración a través de expertos, ésta ha sido planteada como un elemento esencial para proporcionar evidencias teóricas de validez de contenido (Rubio, Berg-Weger, Tebb, Lee y Rauch, 2003) y posibilita que desde el inicio o primeras fases de construcción de una prueba se facilite la representatividad de los ítems que se construyan para el constructo de interés. De esta forma, es una vez que se concluye con el juicio de expertos de la definición cuando se concreta definitivamente la tabla de especificaciones del test (Spaan, 2006), tabla donde se debería encontrar qué constructo se va a evaluar, cuáles son sus componentes y cómo deberían verse representados éstos en el instrumento final según su importancia diferencial.

Nótese como el hecho de disponer de la tabla de especificaciones del test sería un aspecto crucial para facilitar los procesos de adaptación de las escalas a distintas culturas (Balluerka, Gorostiaga, Alonso-Arbiol y Aramburu, 2007), proporcionándose una herramienta esencial para conseguir que las adaptaciones guarden equivalencia conceptual con las escalas origen. De hecho, en las adaptaciones lo relevante no es exclusivamente mostrar evidencias de una posible equivalencia lingüística entre el instrumento origen y el adaptado, aspecto éste que parece ser el único que preocupa en la mayoría de las ocasiones a los autores de las adaptaciones. Por contra, la clave es poner de manifiesto que las adaptaciones son equivalentes conceptualmente hablando, y en este sentido, contar con la tabla de especificaciones del test sería un elemento a considerar a la hora de poder establecer la conexión conceptual obligada. Así pues, ya sea a la hora de valorar escalas

originales, o bien sus posibles adaptaciones, al seleccionar un test debería tenerse en cuenta si se parte de la denominada tabla de especificaciones del test (Spaan, 2006).

Lord y Novick (1968) también subrayaron la relevancia de especificar una vez operativizado el construco, la definición sintáctica de la variable o relaciones esperadas entre el constructo evaluado y otros constructos o indicadores. Al seleccionar un instrumento se debe tener en cuenta que lo que finalmente le va a dar significado o utilidad a unas puntuaciones es el entramado de relaciones contrastadas. Por ello, dichas relaciones deben plantearse a modo de hipótesis a verificar, lo que finalmente posibilitará obtener las evidencias externas de validez del instrumento, elemento esencial de su validez de constructo (Smith, 2005).

Para resumir este apartado se insiste en que el autor o autores de una investigación que han hecho uso de un determinado test deben dejar constancia de que a la hora de seleccionar éste han atendido a la definición operativa del constructo de interés y a cómo se ha llegado a ésta, considerando además que dicha definición está insertada en un entramado teórico de relaciones, el cual permite asignarle significado al trabajo que se haga con la escala.

Información sobre la construcción y evaluación cualitativa de ítems

Es inusual encontrar en los trabajos donde se presentan los datos referidos a la creación/adaptación de un test, información acerca de los criterios usados para la creación de ítems, justificación sobre las opciones de respuesta, etc. Existen trabajos que sirven para guiar este proceso (Martínez, Moreno y Muñiz, 2005; Moreno, Martínez y Muñiz, 2006; Osterlind, 1989) y a la hora de seleccionar un instrumento la elección debe inclinarse a favor de aquellos donde se deje constancia de al menos los criterios de referencia empleados. Este aspecto resulta esencial ya que los ítems no son ni más ni menos que la concreción operativa de los componentes a evaluar. Así, de ítems inadecuados surge siempre una delimitación operativa errónea, y por lo tanto unos resultados finales alejados de los propósitos iniciales.

Al seleccionar un instrumento, el encargado debe tener claro qué respuestas referentes a un constructo le interesan, y comprobar cuál es la prueba que se ajusta mejor a eso. Por ejemplo, en algunos casos, y para algunos trastornos psicológicos, se puede tener interés por su frecuencia de ocurrencia, pero en otros quizá se quiera evaluar su intensidad en el momento actual. Según este ejemplo, en función de cuál sea el objetivo se debería atender a que los ítems y su formato de respuesta se centrasen en intensidad o en frecuencia.

A los creadores/adaptadores de tests se les debe exigir el uso de la denominada tabla de especificaciones de los ítems (Osterlind, 1989; Spaan, 2006), y al menos insertarla en el manual del test o en una publicación similar. En esta tabla, y de manera resumida, aparecen todos los elementos referentes a los ítems generados (formato, escala de respuesta, proporción dentro de la escala, ejemplos redactados, etc.). A través de esta tabla se garantiza una creación dirigida y estandarizada de los ítems por parte de los encargados, mejorándose así la calidad de los mismos. Constatar la presencia de una tabla de especificaciones de los ítems como elemento que ha guiado la creación de ítems debería ser un elemento a considerar para la selección de un instrumento u otro.

Téngase en cuenta lo ya comentado en el apartado anterior en cuanto a la importancia de la tabla de especificaciones del test para los procesos de adaptación, y que sería aplicable a la tabla de especificaciones de los ítems. No obstante, hay que subrayar que para los casos en los que los instrumentos a elegir son adaptaciones y, por lo tanto, los ítems en muchos casos suelen ser traducciones de los originales, debe corroborarse que se han

seguido las recomendaciones existentes sobre este proceso de traducción (Balluerka *et al.*, 2007; Hambleton, 1994, 1996; Hambleton y Jong, 2003; Gordon, 2004), sin olvidar la necesaria equivalencia conceptual entre los ítems originales y traducidos.

Deberá observarse si creados los ítems, las instrucciones de la escala y demás aspectos formales del futuro instrumento, los autores sometieron a evaluación dichos aspectos con la intención de detectar fallos en las instrucciones, ítems mal redactados, etc. Además, téngase en cuenta que al usar un test debemos tener datos que permitan concluir que sus ítems resultan relevantes desde un punto de vista teórico para los componentes del constructo (Clark y Watson, 2003). Por ello, debería estudiarse si el test sobre el que se está interesado proporciona información que asegure que los ítems creados son teóricamente pertinentes para cada componente, y si éstos están representados por una proporción de ítems adecuada, es decir, si los autores del instrumentos proporcionan resultados sobre la validez de contenido del test (Armstrong, Cohen, Eriksen y Cleeland, 2005; Haynes *et al.*, 1995). En este proceso de valoración de los aspectos formales del tests, y de la relevancia teórica de los ítems, normalmente se produce una eliminación determinada de elementos. A la hora de seleccionar un test es importante corroborar que los autores informan sobre qué se ha eliminado y porqué, ya que da información valiosa sobre lo que se queda y sobre la estrategia seguida.

Resultados del análisis estadístico de los ítems

Un aspecto esencial a considerar dentro del proceso que finalmente lleve a seleccionar un test frente a otro, está relacionado con las propiedades métricas de los ítems del instrumento. Una vez constatado que se partió de una batería de ítems claramente superior a los necesitados, y que éstos han pasado los filtros formales y teóricos pertinentes, se debe atender a los resultados encontrados al analizar estadísticamente los ítems. Se debe prestar atención a que en la construcción de la escala se haya procedido efectuando un primer análisis estadístico de éstos a modo de estudio piloto, y donde los criterios de eliminación de los ítems hayan sido claramente especificados. Es conveniente que los resultados de este estudio piloto se hayan visto corroborados con una muestra mayor, y que para ambos casos la muestra de participantes sea de unas características semejantes a las posteriormente usadas para la aplicación de la escala.

En cuanto a la inspección de los estadísticos de los ítems, el investigador debe tener claro para qué va a usar la escala, y así estimar si los estadísticos encontrados le permiten concluir que el test es adecuado para sus intereses. De esta forma, hay que subrayar que no hay criterios estadísticos universales y que deban aplicarse para todos los ítems sea cual sea la escala de la que formen parte. Por ello, y quizá lo más importante al revisar los análisis de ítems asociados a un instrumento, es observar si la decisión de eliminar o conservar un ítem estuvo basada exclusivamente en la aplicación irreflexiva de ciertos índices numéricos, o si se consideraron dichos criterios a la luz de la definición del constructo inicial y de los objetivos de aplicación (para una revisión de los estadísticos más comúnmente usados y cómo valorarlos véase Muñiz, Hidalgo, García-Cueto, Martínez y Moreno, 2005).

Evidencias empíricas de la estructura interna de la prueba

Al analizar la dimensionalidad de una prueba se busca estimar “el grado en el que los ítems y los componentes del test conforman el constructo que se quiere medir y sobre el que se basarán las interpretaciones” (Elosua, 2003, p. 317). Así, las conclusiones sobre si la estructura interna de un instrumento representa fielmente a los componentes o dimensiones del constructo no pueden basarse en las suposiciones teóricas de los autores de la prueba o

en la coherencia aparente que presentan los ítems. Para poder concluir esto, se hace necesario constatar que se ha usado algún procedimiento que permita evaluar empíricamente la estructura interna de la escala. Si los autores del instrumento parten de una definición clara del constructo y de sus componentes, al inspeccionar el test se debe corroborar que se ha usado una estrategia dirigida a contrastar la hipótesis del investigador basada en cómo deben de agruparse los ítems.

Tradicionalmente, y desde un punto de vista empírico, la tarea dirigida a explorar la estructura interna de un test ha sido llevada a cabo a través de la aplicación del análisis factorial (Floyd y Widaman, 1995). A la hora de seleccionar una prueba, se debería al menos estar familiarizado con los pormenores de esta técnica. Hay varios trabajos que han tratado el uso inadecuado e irreflexivo de análisis factorial o temáticas relacionadas (e.g. Batista-Foguet, Coenders y Alonso, 2004; Elosua, 2005; Ferrando, 1996) siendo un clásico el apartado "cómo engañarse a uno mismo con el análisis factorial" (Nunnally y Bernstein, 1995 pp. 599-601). Así, a la hora de seleccionar una prueba debería tenerse en cuenta esta información, y hacer una valoración crítica tanto de los resultados encontrados como del proceso de aplicación seguido.

En otro lugar (Carretero-Dios y Pérez, 2005) ya insistimos en subrayar que el análisis factorial exploratorio no entiende de Psicología. El análisis sólo "agrupa" correlaciones similares, pero conviene resaltar que la agrupación puede ser debida a más elementos que los propiamente conceptuales, como por ejemplo el formato y tipo de ítems. "Se recuerda que la técnica debe estar sometida a los intereses conceptuales, y que un agrupamiento de ítems es sólo eso, un agrupamiento, y que aunque empíricamente relevante, puede carecer de significado psicológico. Los factores "no psicológicos" que pueden hacer que unos ítems aparezcan juntos son tantos, que la aplicación de esta técnica de análisis en el vacío teórico es totalmente improductiva e ineficaz" (Carretero-Dios y Pérez, 2005 p. 536). Por lo anterior, al seleccionar un test debe apreciarse que la aplicación de los análisis factoriales ha estado sujeta a premisas teóricas sobre la dimensionalidad subyacente a los ítems usados. Además, debería apreciarse una contrastación de la dimensionalidad de la prueba a través de muestras distintas (Elosua, 2005).

Resultados de la estimación de la fiabilidad

La fiabilidad de las puntuaciones de una prueba supone otro criterio esencial a tener en cuenta a la hora de seleccionar un test u otro. De hecho, normalmente es uno de los tópicos que se tratan en primer lugar en las monografías especializadas sobre la construcción de tests, y es el valor al que únicamente se suele recurrir para justificar la selección de una prueba. A pesar de esto, en la presentación que se ha seguido se ha optado por tratar la fiabilidad una vez que se han discutido los aspectos que se consideran previos, cronológicamente hablando, en el proceso que constituye la construcción de un instrumento de evaluación. De hecho, no es hasta que se cuenta con un agrupamiento "definitivo" de ítems por componente, cuando se cuenta con el test "definitivo", cuando la estimación de la fiabilidad adquiere su mayor alcance. No obstante, en muchos trabajos se recurre a presentar estimaciones de la fiabilidad dentro de la fase de análisis de ítems, y se incluye normalmente el alfa de Cronbach como un indicador más del análisis de ítems. Sin embargo, se quiere resaltar que a la hora de seleccionar un instrumento, y al revisar la información disponible, se debe apreciar que se ofrecen estimaciones de la fiabilidad de las puntuaciones obtenidas a partir de la forma definitiva o publicada del test, y no de versiones previas o experimentales.

De nuevo, y para evaluar la adecuación de un test en cuanto a la fiabilidad de sus puntuaciones, el investigador interesado debe preguntarse por ciertos aspectos que se relacionan estrechamente con el juicio final que se emita. Por ejemplo, para qué van a ser usadas las puntuaciones, si los participantes con los que se va a trabajar tienen características semejantes a los usados para estudiar la prueba, si las condiciones de evaluación van a ser similares, etc. Junto a estas preguntas, no debe perderse de vista el concepto propio de fiabilidad, además de las características que presentan los distintos estimadores. Así, la valoración de nuevo debe ser razonada y no sólo guiada por la aplicación de reglas estándar.

Sabiendo que son tres los métodos habituales para obtener estimaciones del coeficiente de fiabilidad (Traub, 1994), tales como el método de formas paralelas, el basado en el test-retest y el método centrado en una única aplicación de la prueba, al seleccionar un instrumento u otro debe hacerse un análisis del método aplicado, las razones de su aplicación y su idoneidad para el test en concreto. Desde un punto de vista teórico, y si la prueba y otros aspectos relevantes lo permiten (Muñiz, 1998), la aplicación repetida del test en dos momentos temporales distintos sería el método de preferencia. Si el investigador interesado en seleccionar un test observara que han sido usadas formas paralelas, debería atender a los problemas relacionados con este procedimiento, entre los que destaca la verificación de que realmente se cuenta con formas paralelas del test en cuestión. Sea como fuere, tanto el test-retest como el método de formas paralelas se enfrentan a problemas más generales y que deben considerarse para juzgar las estimaciones de fiabilidad ofrecidas. Entre éstos podrían destacarse el efecto de la experiencia o práctica de la primera evaluación sobre la segunda, los cambios “reales” que se producen en el constructo evaluado y el intervalo de tiempo usado para llevar a cabo la nueva administración del test o de una forma paralela de éste (Muñiz, 1998).

En general, al revisar los tests publicados, puede constatarse que los constructores/adaptadores de un test tienden a estimar la fiabilidad a partir de una única administración del instrumento, recurriendo a los procedimientos basados en el cálculo de la consistencia interna (Osburn, 2000). En el caso de ítems con una escala tipo Likert, el índice de consistencia interna más usado es el alfa de Cronbach, el cual en muchas ocasiones se aplica incumpliendo las recomendaciones sobre su uso (Cortina, 1993). Ya se propusieron varios ejemplos (Carretero-Dios y Pérez, 2005) para evidenciar ciertos problemas asociados a la aplicación indiscriminada del alfa de Cronbach o a la interpretación superficial de los resultados que facilita. No obstante, la frecuencia con la que puede observarse un uso deficiente de este índice, hace que se le dedique más extensión a este contenido.

El encargado de la selección de un tests deberá cerciorarse de que las estimaciones sobre la fiabilidad a través de un índice de consistencia interna son calculadas para las puntuaciones de cada uno de los componentes supuestos del constructo evaluado. Normalmente, los constructos se ven delimitados por varias facetas o componentes que se postulan como elementos a considerar aisladamente. Por ello, la consistencia interna debería ser estimada para cada faceta del constructo.

El juicio sobre la fiabilidad obtenida a través del alfa de Cronbach debe estar muy conectado con el formato de los ítems o con algunas propiedades métricas de éstos que se encuentran muy relacionadas con el resultado final del alfa de Cronbach, tal y como por ejemplo la dificultad de los ítems. Así, en algunos autoinformes, al usar ítems que consisten en preguntas o afirmaciones muy inespecíficas, con un formato muy semejante entre ellos, y con opciones de respuesta comunes, se puede provocar que la respuesta de los participantes

sea "consistente" a través de los ítems, pero que lo que refleje este resultado sea una consistencia a través de ítems que se conectan más con un factor denominado "formato del instrumento", que con el concepto subyacente teóricamente supuesto. Además, esta problemática podría vincularse a los valores "artificialmente" altos que pueden encontrarse a través del alfa de Cronbach, que los investigadores suelen juzgar como algo muy positivo, y que sin embargo servirían para poner de manifiesto un grave problema de representación del constructo por parte de los ítems (consultar la ya clásica problemática de la denominada paradoja de la atenuación, Loevinger, 1957). "En psicología, valores de consistencia interna entorno a 0,95 pondrían de manifiesto más un problema de infra-representación del constructo y validez deficiente, que de adecuada fiabilidad" (Carretero-Dios y Pérez, 2005 p. 541).

Usando valores que puedan servir de guía, que no de constatación irreflexiva, se podría afirmar que índices de fiabilidad situados alrededor de 0,70 resultarían adecuados si el objetivo de la escala es la investigación. Cuando el objetivo del test es el diagnóstico o clasificación, el valor mínimo aconsejado debe situarse entorno a 0,80 (Nunnally y Bernstein, 1995).

Evidencias externas de la validez de las puntuaciones

Las evidencias externas de validez se basan en el análisis de las relaciones entre la puntuación o puntuaciones ofrecidas por el test y: a) un criterio que se esperaba fuera predicho; b) otros tests con el mismo objetivo de medición o con otros constructos con los que se esperaría relación; c) otras variables o constructos con los que se esperaría ausencia de relación, o una relación menor que la esperada con otras variables (AERA *et al.*, 1999).

Al iniciar la presentación de las directrices a seguir a la hora de seleccionar un test de evaluación, se insistió en que el constructo objetivo debía definirse operativamente (semánticamente) pero también ofrecer una definición conceptual delimitada por las relaciones esperadas con otros constructos (sintáctica), o lo que es lo mismo, ubicar al constructo en un entramado de relaciones teóricas. Por parte de la persona que busca seleccionar un test, de lo que se trataría es de establecer hasta qué punto, usando las puntuaciones del test, se han obtenido evidencias que confirman las relaciones esperadas. Es la inspección de los resultados encontrados en este sentido lo que le facilitaría al investigador la información referente a la utilidad o significado de las puntuaciones del test.

El interesado en llevar a cabo la selección de un test en particular, debe tener presente que no existe una estrategia metodológica o técnica de análisis estadístico al uso que sea exclusiva de los estudios que se hayan encargado de obtener evidencias externas de validez. Los resultados podrían derivarse de usar estrategias experimentales, cuasi-experimentales o no experimentales, y por lo tanto las técnicas de análisis podrían apreciarse como diversas. Por ello, en este contexto, lo realmente relevante es apreciar si los autores de un test concreto han justificado las relaciones aportadas a partir de las teorías de interés o resultados de investigación previos, y que en su momento se deberían haber reflejado en la definición sintáctica de la variable. Por supuesto, se deberá verificar si en función de los objetivo de análisis específicos, se ha usado la metodología de estudio más afín a éstos, y los procedimientos de análisis más convenientes, hecho no obstante que es generalizable a la revisión científica de cualquier estudio publicado. Además, habría que recordar que las puntuaciones de un test no "consiguen" evidencias que denoten que ya está fijada su validez de una vez y para siempre. La obtención de evidencias de validez conlleva un proceso inacabado por definición, en continua revisión, y sensible a la evolución del conocimiento sobre el constructo medido, aspectos a los que debe ser igualmente sensible el responsable de la selección de un test.

Conclusiones

El uso de un test u otro para ser usado en una investigación resulta una problemática de suma importancia. Con la intención de discutir sobre las posibles dificultades que pueden aparecer en este proceso de selección de tests, y con la idea de clarificar algunas directrices que ayuden a realizar dicha selección, se ha escrito este trabajo. Sin embargo, las directrices propuestas más que convertirse en una guía esquemática y concreta de aplicación, buscan ser una herramienta que conduzca a la reflexión sobre ciertos elementos y que hagan sopesar de manera más mesurada algunas de las decisiones. Un investigador jamás podrá llegar a unas conclusiones rigurosas si la materia prima que usa para plantear éstas son puntuaciones ofrecidas por instrumentos deficientes. De igual forma, y por la propia ética que define la actividad científica, el responsable de un estudio no puede contentarse con el hecho de haber usado un test con cierto respaldo psicométrico y con unas garantías científicas suficientes. Por el contrario, debe haber una información de base que asegure que ha usado la mejor opción posible de entre todas las que estaban en su conocimiento.

Un informe de investigación de una revista científica al uso tiene un espacio restringido. La justificación de por qué un instrumento y no otro rebasaría este espacio, y por ende resultaría inviable el tratamiento exhaustivo de las razones que han llevado a trabajar con un test y no con otro. Sin embargo, esto no es óbice para que el autor o autores de un trabajo de investigación hagan uso del esquema presentado o de cualquier otro que garantice una selección científica de los tests. Así, y al igual que en otros apartados de un informe se obvia información para simplemente señalar un procedimiento seguido o estrategia empleada, en este campo de la selección de tests los editores de las publicaciones científicas y los revisores deberían insistir en que los autores de un trabajo indiquen al menos los criterios seguidos para seleccionar los instrumentos y dónde estos criterios pueden ser tratados con más detenimiento. Resulta sorprendente encontrar en muchas revistas científicas, y dentro del apartado instrumentos, un mero listado de escalas, y para las que se informa como mucho de su fiabilidad y de algunas referencias donde éstas se han aplicado para ser estudiadas. Junto a la indicación de las escalas empleadas se reclama que deba haber una pregunta que sirva de hilo conductor del apartado citado: ¿por qué estos tests y no otros? Pregunta que debería poder ser contestada por los autores de cualquier trabajo científico donde se haga uso de tests de evaluación psicológica.

Referencias

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

Armstrong, T.S., Cohen, M.Z., Eriksen, L. y Cleeland, C. (2005). Content validity of self-report measurement instruments: An illustration from the development of the Brain Tumor Module of the M.D. Anderson Symptom Inventory. *Oncology Nursing Forum, 32,* 669-676.

Batista-Foguet, J.M., Coenders, G. y Alonso, J. (2004). Análisis factorial confirmatorio. Su utilidad en la validación de cuestionarios relacionados con la salud. *Medicina Clínica, 122,* 21-27.

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.

Blanton, H. y Jaccard, J. (2006). Arbitrary metrics in Psychology. *American Psychologist, 61,* 27-41.

Botella, J. y Gambara, H. (2006). Doing and reporting a meta-analysis. *Internacional Journal of Clinical and Health Psychology, 6,* 425-440.

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.

Carretero-Dios, H., Pérez, C. y Buela-Casal, G. (2006). Dimensiones de la apreciación del humor. *Psicothema, 18,* 465-470.

Clark, L.A. y Watson, D. (2003). Constructing validity: Basic issues in objective scale development. En A.E. Kazdin (Ed.), *Methodological issues & strategies in clinical research (3ª ed.)* (pp. 207-231). Washington, D.C.: APA.

Cortina, J.M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78,* 98-104.

Downing, S.M. y Haladyna, T.M. (2004). Validity tretas: overcoming interferente with proponed interpretations of assessment data. *Medical Education, 38,* 327-333.

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

Elosua, P. (2005). Evaluación progresiva de la invarianza factorial entre las versiones original y adaptada de una escala de autoconcepto. *Psicothema, 17,* 356-362.

Ferrando, P.J. (1996). Evaluación de la unidimensionalidad de los ítems mediante análisis factorial. *Psicothema, 8,* 397-410.

Floyd, F.J. y Widaman, K.F. (1995). Factor análisis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299.

Gordon, J. (2004). Developing and improving assessment instruments. *Assessment in Education: Principles, Policy and Practice, 11,* 243-245.

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229-240.

Hambleton, R.K. (1996). Adaptación de tests para su uso en diferentes idiomas y culturas: fuentes de error, posibles soluciones y directrices prácticas. En J. Muñiz (Ed.), *Psicometría* (pp. 203-238). Madrid: Universitas.

Hambleton, R.K. y Jong, J.H. (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20,* 127-134.

Haynes, S.N., Richard, D.C.S. y Kubany, E.S. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment, 7,* 238-247.

Hogan, T.P. y Agnello, J. (2004). An empirical study of reporting practices concerning measurement. *Educational and Psychological Measurement, 64,* 802-812.

Koretz, D. (2006). Steps toward more effective implementation of the Standards for Educational and Psychological Testing. *Educational Measurement: Issues & Practice, 25,* 46-50.

Linn, R.L. (2006). Following the Standards: Is it time for another revisions? *Educational Measurement: Issues & Practice, 25,* 54-56.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635-694.

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

Martínez, R.L., Moreno, R. y Muñiz, J. (2005). Construcción de ítems. En J. Muñiz, A.M. Hidalgo, E. García-Cueto, R. Martínez, y R. Moreno, *Análisis de ítems* (pp. 9-52). Madrid: La Muralla.

Montero, I. y León, O. (2005). Sistema de clasificación del método en los informes de investigación en Psicología. *International Journal of Clinical and Health Psychology, 5,* 115-127.

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

Muñiz, J. (1998). *Teoría clásica de los tests.* Madrid: Pirámide.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R. y Zaal, J.N. (2001). Testing practices in european countries. *European Journal of Psychological Assessment, 17,* 201-211.

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

Murphy, K.R. y Davidshofer, C.O. (1994). *Psychological testing: Principles and applications* (3ª ed.). Englewood Cliffs, NJ: Prentice-Hall.

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

Osburn, H.G. (2000). Coefficient alpha and related internal consistency reliability coefficients. *Psychological Methods, 5,* 343-355.

Osterlind, S.J. (1989). *Constructing Test Items.* Londres: Kluwer Academic Publishers.

Padilla, J.L., Gómez, J., Hidalgo, M.D. y Muñiz, J. (2006). La evaluación de las consecuancias del uso de los tests en la teoría de la validez. *Psicothema, 18,* 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 tests. *Psicothema, 19,* 173-178.

Ramos-Álvarez, M.M., Valdés-Conroy, B. y Catena, A. (2006). Criterios para el proceso de revisión de cara a la publicación de investigaciones experimentales y cuasi-experimentales en Psicología. *International Journal of Clinical and Health Psychology, 6,* 773-787.

Rubio, D.M., Berg-Weger, M., Tebb, S.S., Lee, E.S. y Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27,* 94-104.

Smith, G.T. (2005). On Construct Validity: Issues of Method and Measurement. *Psychological Assessment, 17,* 396-408.

Spaan, M. (2006). Test and item specifications development. *Language Assessment Quarterly, 3,* 71-79.

Sturm, T. y Ash, M.G. (2005). Roles of instruments in psychological research. *History of Psychology, 8,* 3-34.

TEA (1996). *PMA: Aptitudes Mentales Primarias* (9ª edición). Madrid: TEA.

Traub, R.E. (1994). *Reliability for the social sciences: Theory and applications.* Londres: Sage.

Walsh, W.B. (1995). *Tests and assessment.* Nueva York: Prentice-Hall.

Wise, L.L. (2006). Encouraging and supporting compliance with Standards for Educational Tests. *Educational Measurement: Issues & Practice, 25,* 51-53.

ANEXO 1. Normas para el desarrollo y revisión de estudios instrumentales (tomado de Carretero-Dios y Pérez, 2005).

(A) Justificación del estudio.

		Sí	Dudoso	No
A.1.	**Los antecedentes sobre los que se asienta la justificación son relevantes.**			
A.2.	**La creación/adaptación del instrumento va a suponer una aportación significativa para al área de estudio.**			
A.3.	**El objetivo general de evaluación del test está claramente especificado.**			
A.4.	**Se concreta la población a la que irá destinado el test.**			
A.5.	**Se especifica el propósito o finalidad para el que van a usarse las puntuaciones del test.**			
A.6.	**El proceso de creación/adaptación resulta viable.**			

(B) Delimitación conceptual del constructo a evaluar.

		Sí	Dudoso	No
B.1.	**Aparecen claramente especificados los intentos de conceptuación más relevantes del constructo de interés.**			
B.2.	**Las distintas propuestas conceptuales se presentan integradas en uno o varios marcos teóricos de referencia.**			
B.3.	**Se hace una revisión de los principales instrumentos de evaluación encargados de evaluar a éste o a constructos relacionados.**			
B.4.	**Tras la revisión se realiza una propuesta operativa de las facetas o componentes operativos del constructo a evaluar, la cual es sometida a evaluación a través de expertos.**			
B.5.	**Se presenta detalladamente la información relacionada con el juicio de expertos (selección de expertos, material utilizado, forma de evaluar, etc.)**			
B.6.	**Considerando los resultados de la evaluación de los expertos, los datos de investigación, y los marcos teóricos de referencia, se concreta definitivamente la definición operativa del constructo.**			
B.7.	**Teniendo en cuenta la definición adoptada del constructo, se concretan las relaciones esperadas entre éste y otras variables.**			
B.8.	**Las relaciones predichas para la puntuación total en el constructo están adecuadamente justificadas.**			
B.9.	**En el caso de que el constructo esté compuesto por distintas facetas o componentes, también son establecidas las relaciones esperadas para cada uno de estos componentes.**			
B.10.	**Las relaciones predichas se presentan claras, especificándose cuando el constructo va ser variable predictora, cuando predicha y cuando covariado.**			

(C) Construcción y evaluación cualitativa de ítems.

		Sí	Dudoso	No
C.1.	**La información que justifica el tipo de ítems a construir (incluyendo formato, tipo de redacción, escala de respuesta, etc.) es presentada con claridad.**			
C.2.	**El autor hace uso de una tabla de especificaciones de los ítems para guiar la elaboración de éstos.**			
C.3.	**La tabla de especificaciones de los ítems recoge toda la información necesaria para la construcción de éstos.**			
C.4.	**Se justifica adecuadamente el número de ítems final de la**			

	escala a crear/adaptar.			
C.5.	**La batería de ítems inicial está compuesta por al menos el doble de ítems por componente de los que finalmente pretenden usarse.**			
C.6.	**En caso de traducir los ítems, se ha usado una estrategia que asegura la equivalencia conceptual entre los originales y los traducidos.**			
C.7.	**En caso de haber traducido los ítems, el autor proporciona nuevos ítems vinculados a los componentes del constructo a evaluar.**			
C.8.	**Se presentan las evidencias de validez de contenido proporcionadas por la valoración de un grupo de jueces acerca de la batería inicial de ítems.**			
C.9.	**Aparece toda la información relacionada con el procedimiento seguido para la valoración de los ítems por parte de un grupo de jueces.**			
C.10.	**La valoración de los ítems por parte de un grupo de jueces ha sido llevada a cabo adecuadamente.**			
C.11.	**Los ítems eliminados una vez terminado el proceso de valoración llevado a cabo por un grupo de jueces están claramente especificados.**			
C.12.	**Los ítems conservados una vez terminado el proceso de valoración llevado a cabo por un grupo de jueces están claramente especificados.**			

(D) Análisis estadístico de los ítems.

		Sí	**Dudoso**	**No**
D.1.	**La delimitación del trabajo es clara (primer estudio de los ítems, estudio piloto o validación cruzada)**			
D.2.	**Los objetivos del análisis aparecen claramente especificados (homogeneidad y consistencia de la escala *frente a* validez de criterio).**			
D.3.	**Es facilitada toda la información referente a los ítems, instrucciones a los participantes, contexto de aplicación ,etc.**			
D.4.	**La muestra de estudio tiene características similares a la de la población objetivo del test.**			
D.5.	**El tamaño de la muestra es adecuado para los objetivos del estudio.**			
D.6.	**El procedimiento de evaluación es similar al que se tiene planificado para la escala definitiva (muestreo).**			
D.7.	**Se especifican con claridad los criterios a considerar para la selección-eliminación de los ítems.**			
D.8.	**Los cálculos estadísticos efectuados resultan pertinentes.**			
D.9.	**Los resultados (cualitativos y cuantitativos) se discuten con claridad.**			
D.10.	**Las decisiones sobre los ítems tienen en cuenta cuestiones**			

	teóricas			
D.11.	**Se especifica claramente que ítems son eliminados y por qué.**			
D.12.	**Los ítems seleccionados quedan claramente delimitados.**			

(E) Estudio de la dimensionalidad del instrumento (estructura interna).

		Sí	Dudoso	No
E.1.	**La delimitación del trabajo es clara (primer estudio de dimensionalidad de la escala o validación cruzada de resultados previos)**			
E.2.	**Los objetivos del análisis aparecen claramente especificados (estudio exploratorio *frente a* análisis confirmatorio, o ambos).**			
E.3.	**La información presentada sirve para justificar con claridad los objetivos propuestos.**			
E.4.	**Es facilitada toda la información necesaria para que el lector conozca los antecedentes que justifican la escala y la dimensionalidad esperada de ésta.**			
E.5.	**Información sobre la muestra es completa y pertinente.**			
E.6.	**La muestra de estudio tiene características similares a la de la población objetivo del test.**			
E.7.	**El tamaño de la muestra es adecuado para los objetivos del estudio.**			
E.8.	**El procedimiento de muestreo seguido es correcto para los objetivos del estudio.**			
E.9.	**En el caso de usarse un procedimiento exploratorio de análisis factorial, aparece justificada su necesidad.**			
E.10.	**Se razona con claridad el por qué ha decidido usarse un tipo concreto de análisis factorial exploratorio y no otro.**			
E.11.	**Con anterioridad a la aplicación del análisis factorial exploratorio el autor informa sobre la adecuación de la matriz de correlaciones (esfericidad de Barlett e índice de Kaiser-Meyer-Olkin)**			
E.12.	**La interpretación de la dimensionalidad de la escala es efectuada sobre la solución factorial rotada.**			
E.13.	**El procedimiento de rotación factorial usado es justificado correctamente.**			
E.14.	**El procedimiento de rotación factorial usado es adecuado.**			
E.15.	**La información facilitada sobre la solución factorial resultante es la adecuada (número de factores, saturaciones factoriales relevantes de los ítems que los integran, porcentaje de varianza explicada y comunalidad).**			
E.16.	**Los procedimientos estadísticos usados para discutir cuáles son los factores relevantes a tener en cuenta son adecuados.**			
E.17.	**La discusión sobre los factores a tener en cuenta es enmarcada en la investigación teórica y empírica previa.**			
E.18.	**En el caso de aplicarse un procedimiento basado en el análisis factorial confirmatorio, el modelo de medida (forma de distribuirse los ítems) a analizar es claramente delimitado.**			
E.19.	**En el estudio, junto al modelo de referencia, se someten a diagnóstico comparativo propuestas alternativas.**			
E.20.	**Se justifica el procedimiento de estimación usado.**			
E.21.	**El procedimiento de estimación elegido en el estudio resulta adecuado.**			
E.22.	**Para el diagnóstico del modelo el autor usa simultáneamente varios índices.**			
E.23.	**En el trabajo se informa sobre el por qué de los índices seleccionados y cuáles van a ser los valores de corte a**			

	considerar para estimar la bondad de ajuste del modelo.			
E.24.	**En el trabajo se presentan con claridad los resultados para los distintos índices de bondad de ajuste.**			
E.25.	**Si el autor hace modificaciones para mejorar el ajuste, las decisiones están claramente fundamentadas (teóricas y empíricamente), y aparecen con claridad en el estudio.**			
E.26.	**El autor presenta el diagrama (path diagram) donde aparece la distribución de los ítems por factor, el "grado" en el que cada uno de éstos es predicho por el factor de pertenencia, y en general todos los parámetros considerados relevantes en la especificación inicial del modelo.**			

(F) Estimación de la fiabilidad.

		Sí	**Dudoso**	**No**
F.1.	**En el trabajo se justifica el procedimiento de estimación de la fiabilidad a usar (adecuación teórica).**			
F.2.	**El método de estimación de la fiabilidad empleado se considera adecuado.**			
F.3.	**Si en el informe se usa el método *test-retest*, son proporcionados y discutidos los aspectos más significativos que afectan a este cálculo aparte de las cuestiones teóricas (intervalo temporal, condiciones de evaluación, correspondencia muestral, etc.).**			
F.4.	**Teniendo en cuenta los aspectos más significativos que afectan a la aplicación del método test-retest (intervalo temporal, condiciones de evaluación, correspondencia muestral, etc.), ésta se considera adecuada.**			
F.5.	**Si en el informe se usa el método de *formas paralelas*, son proporcionados y discutidos los aspectos más significativos que afectan a este cálculo aparte de las cuestiones teóricas (datos sobre la equivalencia de las pruebas, además de la información común al test-retest, como intervalo temporal, condiciones de evaluación, correspondencia muestral, etc.)**			
F.6.	**Teniendo en cuenta los aspectos más significativos que afectan a la aplicación de las formas paralelas (equivalencia de las pruebas, intervalo temporal, condiciones de evaluación, correspondencia muestral, etc.), ésta se considera adecuada.**			
F.7.	**Si en el informe se usa el índice *alpha de Cronbach* basado en la consistencia interna, son proporcionados y discutidos los aspectos más significativos que afectan a este cálculo aparte de las cuestiones teóricas (número de ítems por componente del constructo, y formato de éstos).**			
F.8.	**Teniendo en cuenta los aspectos más significativos que afectan a la aplicación del *alpha de Cronbach* (número de ítems por componente del constructo y formato de éstos), ésta se considera adecuada.**			
F.9.	**Si en el informe se usa un procedimiento basado en la obtención de *dos mitades de un test* para el cálculo de la consistencia interna, son proporcionados y discutidos los aspectos más significativos que afectan a este cálculo aparte de las cuestiones teóricas (procedimiento para obtener las dos**			

	partes y número de ítems que las integran).			
F.10.	**Teniendo en cuenta los aspectos más significativos que afectan a la aplicación del procedimiento basado en la obtención de *dos mitades de un test* (número de ítems y formato de éstos), ésta se considera adecuada.**			
F.11.	**El tamaño de la muestra de estudio es adecuado para los objetivos de la investigación.**			
F.12.	**Las características de los participantes son adecuadas en función de los objetivos del test y finalidad de las puntuaciones**			
F.13.	**El procedimiento de evaluación utilizado es adecuado en función de las características de la prueba.**			
F.14.	**Los resultados derivados de la estimación de la fiabilidad se muestran con claridad.**			
F.15.	**La discusión de los resultados se hace teniendo en cuenta tanto aspectos metodológicos como teóricos.**			
F.16.	**En el caso de obtenerse unos datos deficientes de fiabilidad, en el trabajo son discutidas las estrategias a adoptar.**			