© International Journal of Clinical and Health Psychology

ISSN 1697-2600
2005, Vol. 5, Nº 3, pp. 521-551

Normas para el desarrollo y revisión de estudios instrumentales

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

(Recibido 5 de abril 2005 / Received April 5, 2005)
(Aceptado 7 de junio 2005 / Accepted June 7, 2005)

RESUMEN. En este estudio teórico se proponen los criterios más importantes a tener en cuenta para el desarrollo y revisión de estudios que tengan como objetivo crear o adaptar tests referidos a la norma. En concreto, este artículo va a ocuparse de la creación/adaptación de los tests que evalúan algún constructo relacionado con el campo de la Psicología Clínica y de la Salud. La información va a presentarse organizada en un total de siete apartados, cada uno de los cuales corresponde a una fase distinta dentro del proceso de creación/adaptación de tests. Estos apartados son: justificación del estudio, delimitación conceptual del constructo a evaluar, construcción y evaluación cualitativa de ítems, análisis estadístico de los ítems, estudio de la dimensionalidad del instrumento, estimación de la fiabilidad y obtención de evidencias externas de validez. El trabajo finaliza con un resumen de las reglas básicas a considerar, insistiéndose que éstas sean tenidas en cuenta no sólo para desarrollar o revisar estudios cuyo objetivo sea crear/adaptar tests, sino también para decidir sobre el uso de un instrumento de medida en una investigación.

PALABRAS CLAVE. Normas para la revisión de estudios instrumentales. Construcción de tests. Adaptación de tests. Estudio teórico.

ABSTRACT. The more important criterions to development and to review studies whose aim is to create or adapt norm referred tests are proposed in this theoretical study. Thus, this manuscript is focused on the creation/adaptation of tests used on

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

clinical and health psychology fields. The content is structured on seven principal points, which are centred in each one of distinct phases of the test creation/adaptation process. These are: study justification, conceptual and theoretical framework related to construct to be assessed, items contruction and item theoretical assessment, item analysis, internal structure study, reliability, and studies to obtain external evidences of validity. An abstract of basic rules to be considered is presented. Finally it is noted the relevance to not consider this rules only to create or adapt tests but to assess the adequacy about using a particular assessment instrument.

KEY WORDS. Norms to review instrumental studies. Test construction. Tests adaptation. Theoretical study.

RESUMO. Neste estudo teórico propõem-se os critérios mais importantes a ter em conta para o desenvolvimento e revisão de estudos que tenham como objectivo criar ou adaptar testes referidos a normas. Concretamente, este artigo ocupa-se da criação / adaptação dos testes que avaliam algum construto relacionado com o campo da Psicologia Clínica e da saúde. A informação apresentada será organizada num total de sete pontos, cada um dos quais corresponde a uma fase distinta dentro do processo de criação / adaptação de testes. Estes pontos são: justificação do estudo, delimitação conceptual do construto a avaliar, construção e avaliação qualitativa de itens, análise estatística dos itens, estudo da dimensionalidade do instrumento, estimação da fiabilidade e obtenção de evidências externas de validade. O trabalho finaliza com um resumo das regras básicas a considerar, insistindo-se que estas sejam tidas em conta não só para desenvolver ou rever estudos cujo objectivo seja criar/adaptar testes, mas também para decidir sobre o uso de um instrumento de medida numa investigação.

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

Introducción

Dentro de la gran variedad de trabajos de investigación que surgen desde la Psicología, los centrados en la construcción o adaptación de tests ocupan un lugar destacado, lo que podría ponerse de manifiesto a través de distintos ejemplos. Así, la base de datos PsycINFO, con las palabras clave *test construction*, *test adaptation* y *test translation* dentro del campo de búsqueda *materia* (unidas con el nexo "or"), proporciona para los últimos cinco años (2000-2004) un total de 2752 publicaciones. De igual modo, en un estudio donde fue analizada la proporción de trabajos publicados según su metodología de estudio en las revistas de Psicología en castellano con factor de impacto durante los años 2000-2001 (Buela-Casal, Carretero-Dios y De los Santos-Roig, 2002), se constató como el 22% de los artículos publicados en la revista Psicothema, el 13% de los aparecidos en la Revista Latinoamericana de Psicología y el 25% de los presentes en la Revista Mexicana de Psicología tenían como objetivo analizar las propiedades psicométricas de algún instrumento de evaluación. Si a esto se le añade que en la mayoría de trabajos de investigación llevados a cabo en Psicología se hace uso de tests,

la conclusión resulta evidente, y no es otra que la relevancia de la medida de lo psicológico para investigar en Psicología.

El hecho es que la Psicología frecuentemente estudia fenómenos no directamente observables, los cuales pretenden medirse, y para lo que se usan aproximaciones indirectas. La depresión, ansiedad, calidad de vida, neuroticismo, etc. son ejemplos de constructos empleados para referirse a este tipo de fenómenos y que supuestamente son parte de un sujeto determinado. Su medición suele conllevar el uso de indicadores observables, como por ejemplo las respuestas de una persona a las preguntas de un cuestionario, y la bondad de ésta va a depender de las garantías científicas de dicho cuestionario. Así, y ni que decir tiene, los estudios dedicados a investigar dichas garantías son de vital importancia para la Psicología. De hecho, y como señal de esta relevancia, una clasificación de las metodologías de investigación en Psicología presentada por Montero y León (2002, 2005), viene a proponer a los estudios instrumentales como categoría independiente, entendiendo a éstos como los encargados del "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). Propiedades psicométricas que igualmente deben ponerse de manifiesto siempre que cualquier investigador necesite utilizar algún instrumento para medir sus variables de estudio.

Dada la incuestionable importancia de la construcción, adaptación y uso de tests, resulta más que adecuado reconsiderar periódicamente el proceso y secuenciación de las tareas que se dan entorno a estas actividades, persiguiendo con esto mejorar la labor que supone la medida de lo psicológico a través de tests. El objetivo de este artículo, el cual puede clasificarse como estudio teórico (Montero y León, 2005), es presentar algunos principios básicos que deben tenerse en cuenta en todo proceso de construcción/adaptación de un test, a la vez que destacar la información y formato que deben presentar los artículos científicos dedicados a difundir dicho proceso. Estos principios, claro está, han sido ya tratados y analizados en diversas fuentes (AERA, APA, NCME, 1999; Clark y Watson, 1995; Hambleton y Jong, 2003; Haynes, Richard y Kubany, 1995; Muñiz y Hambleton, 1996, 2000; Nunnally y Bernstein, 1995; Smith, Fischer y Fister, 2003; Walsh, 1995, etc.). Sin embargo, una revisión reciente de los trabajos ocupados de esta temática (Clark y Watson, 2003), pone de manifiesto la falta de coherencia entre los distintos artículos publicados, y que se pondría de manifiesto tanto en el formato de presentación de este tipo de estudios, como en el rigor y disparidad de datos facilitados para trabajos con objetivos similares.

En nuestro entorno cultural ha sido señalada la falta de similitud formal y de contenido que en la actualidad puede observarse entre artículos con objetivos y metodología similares (Buela-Casal *et al.,* 2002), y que provoca que sea difícil la puesta en marcha de distintos meta-análisis, la comparación directa entre éstos para, por ejemplo, analizar metodológicamente los trabajos, o simplemente la mera comunicación científica. Con la intención de paliar esta situación, en el medio donde este trabajo se publica, el *Internacional Journal of Clinical and Health Psychology,* han aparecido distintos artículos con el propósito de consensuar ciertas normas para la publicación en general (Bobenrieth, 2002), o para situaciones más particulares como podrían ser la redacción de casos clínicos (Buela-Casal y Sierra, 2002) o la presentación de estudios experimentales (Ramos-Álvarez y Catena, 2004).

En este estudio van a proponerse las normas a tener en cuenta para la elaboración y revisión de investigaciones instrumentales (véase el Anexo 1). Su finalidad es la de servir como referencia para el interesado en la publicación o revisión de trabajos de estas características, a la vez que se llama la atención al "usuario" de tests, es decir, se insiste en que estas normas sean igualmente tenidas en cuenta a la hora de seleccionar un instrumento de medida ya disponible para su aplicación. La intención no es que los criterios que van a especificarse sean definitivos e indiscutibles. Éstos, a través de la comunicación e intercambio entre profesionales, junto con el inevitable avance del conocimiento, y considerando las observaciones, reflexiones y sugerencias que puedan emitirse desde distintos puntos de vista, deberán ir retocándose y adaptándose.

Ámbito de aplicación de la propuesta y consideraciones previas

Las normas que van a presentarse en este trabajo no pueden aplicarse a cualquier investigación encargada de desarrollar o adaptar un test. Así, para enmarcar debidamente el alcance de este estudio habría que concretar que: a) el foco de atención va a recaer exclusivamente sobre los auto-informes, los cuales ocupan el 67% de las publicaciones de carácter instrumental en Psicología (Clark y Watson, 2003), con una escala de respuesta tipo Likert común para todos los ítems, señalado igualmente como el procedimiento más comúnmente usado (Smith y McCarthy, 1995); b) la información presentada va a ocuparse de los auto-informes que tienen como objetivo la evaluación de constructos relacionados con la Psicología Clínica y de la Salud y que no puedan ser enmarcados en el campo de las aptitudes; c) dentro de este grupo, este artículo va a centrarse en los tests referidos a las normas, entendidos como aquellos que tienen como objetivo evaluar una cualidad o "rasgo" latente del sujeto (depresión, ansiedad, estilo de afrontamiento, etc.), y donde la finalidad es poner de relieve las diferencias interindividuales en dicha cualidad o rasgo; con esta aproximación, el "lugar" que ocupa una persona en un continuo imaginario encargado de representar, por ejemplo, al neuroticismo, va a depender del grupo de comparación o normativo; d) los criterios especificados van a sustentarse en la Teoría clásica de los tests, no tratándose información relacionada con otras aproximaciones dirigidas a la construcción de tests, como la Teoría de respuesta a los ítems o la Teoría de la generalizabilidad.

Proponer unas normas sobre la elaboración y revisión de artículos instrumentales para su aplicación a cualquier trabajo con las características anteriormente señaladas resulta imposible. La labor dirigida a crear o adaptar un test ocupa un espacio prolongado de tiempo, lo que suele ir acompañado de un amplio bagaje de resultados. Así, es impensable que un investigador presente en un solo artículo toda la información derivada del proceso llevado a cabo.

En las revistas especializadas es normal encontrar trabajos dedicados exclusivamente a los análisis estadísticos de los ítems, a estudiar la estructura interna de una escala, explorar las evidencias de validez convergente o divergente, etc. Esto significa que la propuesta de unas normas para la publicación de artículos instrumentales va a ir más allá de cualquier estudio individual, pero éste debe verse representado en dichas normas. Así, se advierte que quizá un investigador solo deba considerar parte de la

información que va a presentarse en este estudio, lo que va a depender del objetivo de su trabajo en cuestión (por ejemplo, estudiar la estructura interna de una escala). Para facilitar esta tarea, los elementos a tener en cuenta en la labor que supone la construcción o adaptación de un test van a ser agrupados en categorías de interés. Estas categorías corresponden a cada uno de los pasos que deben darse a la hora de construir o adaptar un test, y aparecerán ordenadas según el lugar que ocupan dentro de este proceso (AERA *et al.,* 1999).

En España, lo normal suele ser la "importación" de instrumentos de evaluación (Buela-Casal, Sierra, Carretero-Dios y De los Santos-Roig, 2002). Esto quiere decir que la mayoría de los investigadores están más ocupados en *adaptar* que en *crear*, a pesar de las reiteradas advertencias sobre la falta de funcionalidad de muchas de estas adaptaciones, o sobre la ausencia de relevancia cultural que de ellas puede derivarse (Pelechano, 1997, 2002). Sea como fuere, la adaptación de instrumentos de evaluación sigue siendo lo habitual en nuestro país, por lo que unas normas para la elaboración y revisión de artículos instrumentales deben recoger este aspecto. Es en este momento donde no puede dejar de advertirse que para la adaptación de un test deben seguirse los mismos pasos que para su creación original.

Algunos investigadores opinan que si un determinado instrumento ha mostrado ciertas garantías científicas en el entorno donde fue creado, éste puede ser extrapolado sin más a otro contexto cultural, si acaso proporcionando algún dato sobre fiabilidad o estructura factorial. Lo que el adaptador "obtiene" de la escala original es la concreción de partida del autor de la prueba sobre el constructo a evaluar, es decir, la delimitación conceptual de éste. La adaptación supone partir desde esa conceptuación, y repetir de nuevo todos los pasos necesarios para que el instrumento se adapte adecuadamente al nuevo entorno. Por ello, las normas centradas en las etapas de construcción de una prueba, son igualmente las que recogen la adaptación de ésta.

Otra práctica habitual es la de traducir una escala a una lengua distinta a la usada para su creación, tomando a esta traducción como algo distinto a una adaptación. Son muchos los trabajos donde han sido propuestas las directrices a seguir a la hora de traducir un test a otra lengua (Hambleton, 1996; Hambleton y Jong, 2003; Muñiz y Hambleton, 1996, 2000); sin embargo, hay que recordar que el seguimiento exclusivo de las normas sobre cómo llevar a cabo el proceso mismo de traducción para que los indicadores operativos (ítems) del constructo evaluado se vean reflejados en la escala traducida, no significa que una vez traducidos éstos pueda concluirse que la nueva versión es por sí misma adecuada, o que ya no tiene que pasar por los filtros a los que la escala original tuvo que someterse, o debería haberse sometido. Los autores de este trabajo mantienen que la traducción es parte de la adaptación y que, por lo tanto, es una posible elección a la hora de adaptarla. Así, en vez de crear de nuevo una batería de ítems para el constructo evaluado teniendo en cuenta la conceptuación original, se opta por seleccionar directamente los que componen la escala que interesa. Con este procedimiento se supondría que traduciendo correctamente los ítems, se contaría con los elementos necesarios para que la nueva versión "funcione", y que cuando sea analizada los resultados serán similares a los encontrados con el instrumento original. En la mayoría de las ocasiones la mera traducción no conlleva adecuación conceptual ni relevancia cultural para el nuevo entorno, y cuando los análisis pertinentes son llevados

a cabo los resultados suelen alejarse de los encontrados con la escala original. Así, distintos autores han destacado como la traducción suele ser una elección rápida y arriesgada que conduce a resultados inadecuados (Osterlind, 1989; Smith *et al.,* 2003), y que finalmente provoca, en el mejor de los casos, un paso atrás en la investigación para una nueva formulación de ítems, y en el peor, un abandono del proceso y un dato más sin conexión ni utilidad. Por ello, hay que insistir en que la creación y adaptación de tests es cuestión de reflexión y no de premura, y que se adopte la elección que se adopte, deben seguirse las recomendaciones generales sobre este difícil proceso (AERA *et al.,* 1999). Dado que la traducción es una práctica habitual entre los investigadores, las normas que van a presentarse van a ocuparse de este proceso, aunque entendido como uno de los posibles pasos a dar dentro de una adaptación, y sin dejar de recomendar al lector que acuda a trabajos especializados como los anteriormente citados.

Normas para la elaboración y revisión de artículos instrumentales

A continuación se presentan las recomendaciones a tener en cuenta para la elaboración y revisión de artículos ocupados en alguna de las fases que comprenden la construcción o adaptación de un test (se recuerda de nuevo que el usuario final del test debe igualmente valorar las reglas que van a especificarse). Los criterios van a desarrollarse dentro de siete apartados, cada uno de los cuales corresponde a una etapa distinta dentro del proceso que nos ocupa; en éstos son analizados los aspectos más relevantes a considerar dentro de la fase en cuestión de construcción/adaptación de la prueba. Una vez terminada la presentación, podrá verse en el Anexo 1 un resumen de las seis primeras fases de construcción/adaptación de un test previamente analizadas, en donde podrá apreciarse la concreción de las reflexiones y comentarios efectuados a lo largo del trabajo. Así, se presentarán distintas afirmaciones referentes a determinados aspectos que deben tenerse en cuenta a la hora de presentar o revisar una investigación instrumental. A cada afirmación le sigue tres posibilidades de respuesta ("sí", "no" o "dudoso"), siendo lo adecuado una respuesta afirmativa. Como podrá observarse no se inserta información correspondiente a la séptima y última fase del proceso de construcción/adaptación de un test, lo cual será justificado dentro del apartado correspondiente (véase en Obtención de evidencias externas de validez). La Tabla 1 muestra las fases de una investigación encargada de la construcción/adaptación de un test con las particularidades apuntadas anteriormente.

TABLA 1. Fases de la construcción/adaptación de un test.

A. Justificación del estudio.
B. Delimitación conceptual del constructo a evaluar.
C. Construcción y evaluación cualitativa de ítems.
D. Análisis estadístico de los ítems.
E. Estudio de la dimensionalidad del instrumento (estructura interna).
F. Estimación de la fiabilidad.
G. Obtención de evidencias externas de validez.

Justificación del estudio (A)

Para llegar a proponer como objetivo de investigación crear un nuevo instrumento de evaluación o adaptar una herramienta ya existente a otro contexto de aplicación resulta fundamental justificar las razones de este hecho, presentando información coherente y relevante sobre qué aportaría la nueva escala o adaptación con respecto a medidas ya existentes y cuáles son las condiciones que hacen que la investigación propuesta resulte, además de pertinente, viable. Junto a lo anterior hay que resaltar que el primer paso a tener en cuenta en todo proceso de creación/adaptación de un test es delimitar qué se quiere evaluar, a quién y para qué van ser usadas las puntuaciones obtenidas con el test. La respuesta a qué es lo que se va a evaluar va a afectar al tipo de información que va a tenerse en cuenta o qué conjunto de teorías va a ser consultado para acabar proponiendo una conceptuación determinada del constructo de interés, al modelo de medida elegido, sin olvidar que dicha respuesta también va a influir en el procedimiento de evaluación que va a ser seleccionado, y que por lo que aquí respecta ya ha sido reducido a los auto-informes de uso en Psicología Clínica y de la Salud y no relacionados con las aptitudes. Con respecto a quién quiere evaluarse, habría que decir que igualmente es de vital importancia señalar si, por ejemplo, la población objetivo son los estudiantes universitarios, los pacientes diagnosticados de una determinada psicopatología, la población general, etc. Esta concreción va a afectar a fases esenciales de la construcción/adaptación del test, y que irían desde el tipo de ítems a redactar, selección de muestras, diseño de los estudios de validez, etc. Resulta sumamente importante conocer para qué van a ser usadas las puntuaciones que se obtengan con el test. En función de la finalidad perseguida van a adoptarse posturas diferenciadas dentro de algunas de las fases de creación/adaptación de la prueba. Así, no es lo mismo usar las puntuaciones para poner a prueba distintas teorías científicas, que para llevar a cabo un diagnóstico clínico. En este caso, por ejemplo, las exigencias de fiabilidad son distintas, y los criterios para la selección de ítems pueden ser igualmente dispares. Consúltese el trabajo de Navas (2001) o Barbero (2003) para una clasificación de los posibles usos de las puntuaciones proporcionadas por un test.

Delimitación conceptual del constructo a evaluar (B)

Un primer paso indiscutible a la hora de emprender cualquier proyecto dirigido a la creación/adaptación de un nuevo instrumento de evaluación es definir cuidadosamente el constructo que quiere evaluarse. Aunque la insistencia sobre este aspecto no es reciente, algunas investigaciones adolecen de una adecuada conceptuación del constructo, que suele ser la consecuencia de una deficiente revisión bibliográfica, y que finalmente acaba repercutiendo sobre la calidad del instrumento creado. Sobre cómo realizar la concreción conceptual del constructo existen varios trabajos que pueden ser de ayuda (Haynes *et al*., 1995; Murphy y Davidshofer, 1994; Walsh, 1995). Algunos de los elementos a considerar pueden ser comunes a toda investigación (una adecuada revisión bibliográfica), pero otros son actividades más específicas del proceso de creación/adaptación de un test. Así, se quiere destacar la importancia que en la conceptuación del constructo tiene la concreción inicial de las facetas o componentes operativos de éste, denominada clásicamente como definición semántica de la variable (Lord y Novick,

1968), junto con la evaluación a través de expertos de dicha definición. La definición semántica de un constructo que no presente claramente sus elementos diferenciadores, 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 ambiguo, impreciso y tendente a proporcionar unas deficientes evidencias de validez de contenido (Nunnally y Berstein, 1995). Así, desde los propios estándares para la creación de tests psicológicos y educativos (AERA *et al.,* 1999) se insiste en la necesidad de hacer explícita la definición semántica del constructo que va a servir de referencia a la hora de crear/adaptar el instrumento de interés. De esta forma, existen varias propuestas sobre cómo presentar dicha definición, aunque coincidentes en lo adecuado de usar una tabla donde se inserte toda la información de interés (Osterlind, 1989). Esto permitiría detectar fácilmente los componentes o facetas operativas del constructo y, aunque el texto del trabajo debe justificar la información insertada en la tabla, un investigador interesado en "ver" la propuesta general de definición podría hacerlo accediendo directamente a dicha tabla. Aparte de los intereses de comunicación científica, el objetivo de usar una tabla para la concreción de la definición operativa del constructo es básicamente el de la evaluación que debe llevarse a cabo de dicha definición por parte de un grupo de expertos en la temática (AERA *et al.,* 1999); aunque es común obviarla (Smith *et al.,* 2003), ésta ha sido planteada como un elemento esencial para proporcionar evidencias teóricas de validez (Osterlind, 1989; Rubio, Berg-Weger, Tebb, Lee y Rauch, 2003). Por ello, el hecho de evaluar en las primeras fases de construcción/adaptación de una prueba si un componente seleccionado como pertinente para un constructo lo es realmente resulta más que importante. Además, a través de esta evaluación se busca recoger sugerencias y recomendaciones acerca de la definición adoptada, lo que puede conducir a una matización, mejora o reconsideración de ésta, lo que sería mucho más costoso en fases posteriores (véase Grant y Davis, 1997 para un ejemplo de aplicación). Finalizado el juicio de expertos de la definición semántica debe realizarse una propuesta operativa definitiva. Ésta, también presentada en una tabla (Smith *et al.,* 2003), debe contener toda la información referente a la definición semántica del constructo, pasando a ser la tabla de especificaciones del test, puesto que a través de dicha tabla debe saberse qué constructo va a ser evaluado, cuáles son sus componentes y cuál es la importancia diferencial de cada uno de éstos. Así, el contenido de la prueba que se cree o adapte va a tener que reflejar la información recogida en la tabla encargada de presentar la definición semántica de la variable; de ahí la importancia de insertar ésta en los trabajos instrumentales. Finalmente, y para acabar con la fase de delimitación conceptual, hay que insistir en que junto con la definición operativa o semántica del constructo, en esta fase igualmente deben hacerse explícitas las relaciones esperadas para el constructo evaluado, es decir, debe proporcionarse la definición sintáctica de la variable (Lord y Novick, 1968). La especificación de las relaciones esperadas va a ser un factor de suma importancia para los posteriores estudios dirigidos a obtener las evidencias externas de validez del instrumento. Así, en función de la revisión bibliográfica hecha, y de los modelos teóricos de referencia, debe proponerse un entramado significativo de relaciones para el constructo, y que deben posteriormente corroborarse a través de las puntuaciones derivadas de la prueba. Son

estas relaciones las que acaban dando significado al valor que la escala facilite, siendo esencial e ineludible su justificación y especificación.

Construcción y evaluación cualitativa de ítems (C)

En el momento que se cuenta con un constructo claramente delimitado en cuanto a sus facetas o componentes operativos, e igualmente ha sido establecida la red de relaciones esperadas tanto para el constructo en general como para cada una de sus facetas, puede emprenderse la tarea de construcción de ítems. La elaboración de los ítems de la prueba va a suponer una etapa crucial dentro del proceso de construcción/adaptación de ésta, y no conviene olvidar que "el uso de los refinados procedimientos empíricos para analizar y seleccionar los ítems no permitirá construir un test de calidad si la materia prima es deficiente" (Prieto y Delgado, 1996, p. 108). En esta tarea debe tenerse en cuenta a quién se quiere evaluar (acomodando los ítems a su nivel cultural, edad, lengua, etc.). Además, la respuesta sobre a quién quiere evaluarse, junto con la consideración de otros factores externos que van a estar siempre presentes, va a afectar, por ejemplo, al tiempo que va a poder dedicarse a la evaluación, a cómo va a ser la aplicación, individual o colectiva, o a cuál va a ser el modelo de medida adoptado. Antes de crear los ítems propiamente dichos debe reflexionarse sobre todas las posibles variables de influencia, las cuales deben afectar al proceso de creación de ítems. En la mayoría de las publicaciones encargadas de presentar los datos referidos a la creación/adaptación de un test, no se presenta información sobre las razones que han provocado que los ítems sean redactados de una forma determinada, que se use una escala de respuesta específica con un número de opciones de respuesta concreto o por qué se ha decidido, por ejemplo, asociar estas opciones a etiquetas verbales referidas a un criterio temporal (Nunca, Siempre, A veces, etc.), de intensidad (Poco, Bastante, etc.), o de adhesión (Totalmente de acuerdo, Nada de acuerdo, etc.). Estas decisiones deben ser tomadas en función de las características del constructo a evaluar, los modelos teóricos adoptados, objetivo de evaluación, población de interés y exigencias de la realidad. Así, no puede olvidarse que los ítems son la concreción operativa de los componentes a evaluar y que de ítems inadecuados surge una delimitación operativa errónea, es decir, una deficiente validez de contenido (Rubio *et al.,* 2003). Todo ello conlleva la necesidad de proponer una tabla de especificaciones de los ítems (Osterlind, 1989), donde aparezcan todos los elementos necesarios para poder elaborar éstos (formato de ítems, escala de respuesta, proporción dentro de la escala o, incluso, un ejemplo redactado). La tabla de especificaciones de los ítems debe permitir que una persona experta no involucrada en la construcción/adaptación del test, teniendo en cuenta la información que allí se facilita, pueda generar ítems. La delimitación de la tabla de especificaciones de los ítems es algo que en la actualidad brilla por su ausencia, a pesar de resaltarse sus ventajas, tanto por el hecho de facilitar que se obtengan ítems más relacionados con los intereses de partida, como por el hecho de posibilitar le creación de ítems por profesionales distintos, aumentando la cantidad y variedad de éstos, y posibilitando en mayor medida la obtención de una adecuada validez de contenido (Osterlind, 1989). El iniciar la tarea de construcción de ítems igualmente conlleva preguntarse por cuántos ítems son suficientes. El autor de la prueba debe tener planificado y justificado el

número de ítems deseado de su instrumento en función, fundamentalmente del número y disparidad de dimensiones recogidas en la tabla de especificaciones del test y del tiempo disponible para la evaluación, hecho este último muy influido, como ya se ha dicho, por quién va a evaluar, cómo, dónde, etc. Una vez decidido el número de ítems que la escala definitiva debe tener, en esta primera etapa de construcción de ítems es deseable elaborar bastantes más elementos de los estipulados como adecuados para dicha escala. Téngase en cuenta que esos ítems van a tener que pasar por distintos filtros, lo que provocará que muchos de ellos tengan que descartarse. Así, cuando se parte de un número reducido de elementos puede ocurrir que una vez finalizados los análisis pertinentes no se tenga el número necesario como para cubrir las necesidades teóricas y psicométricas fijadas. Así, los autores recomendamos construir al menos el doble de ítems que los estipulados como adecuados para cada uno de los componentes del instrumento final.

Cuando para la construcción de ítems se decide llevarse a cabo la traducción de los ítems de una prueba para así emprender su adaptación, esto debe estar motivado por una labor de reflexión del investigador encargado de dicha tarea y no por resultar la opción más "fácil" (además de asegurarse de que la prueba original presenta las necesarias garantías psicométricas). Esto significa que del análisis de los ítems de la escala que quiere adaptarse original debe concluirse su adecuación teórica-práctica para los propósitos de la adaptación, hecho que conduce finalmente a su traducción (el autor de la adaptación tendría que añadir nuevos ítems a los ya traducidos evitando así los posibles problemas ya señalados, para lo que beberá tener en cuenta los aspectos comentados en los párrafos anteriores). Para la traducción deben seguirse ciertos procedimientos que aseguren la equivalencia entre los originales y los traducidos (Gordon, 2004; Hambleton, 1994, 1996; Hambleton y Jong, 2003). En general, pueden adoptarse dos estrategias de las que el adaptador debe dar completa información. Una de ellas es la traducción hacia delante o directa, donde un grupo de traductores traduce los ítems de la escala original al nuevo idioma, para que a continuación otro grupo de traductores juzgue su equivalencia. La otra estrategia se denomina traducción inversa. En este caso, también un grupo de traductores traduce los ítems al idioma requerido, aunque una vez hecha esta tarea, ahora otro grupo de traductores lo vuelve a traducir a la lengua original, y es esta nueva versión la que se compara con la original (véase Hambleton, 1996 para más detalle).

No puede olvidarse que el objetivo esencial de esta fase es conseguir una muestra de ítems relevante para cada uno de los componentes del constructo (Clark y Watson, 2003), teniéndose pues que facilitar la evidencia necesaria que asegure que cada componente esté bien representado por los ítems elaborados y en la proporción adecuada en función de su importancia dentro de la definición adoptada. Debido a este objetivo, no puede considerarse que una vez construida la batería inicial de ítems pueda darse por finalizada esta etapa. Aún deben obtenerse las necesarias evidencias cualitativas de validez de contenido (Smith *et al.,* 2003).

La validez de contenido viene siendo estudiada como una parte integrante de la validez de constructo. Los autores de este trabajo apoyan la idea de considerar lo que tradicionalmente se entiende como validez de contenido como una evidencia de que la

definición semántica quedó bien recogida en los ítems formulados. En este sentido, dentro de la fase que nos ocupa, el propósito es proporcionar evidencias a favor de que los ítems construidos son relevantes para el constructo y representan adecuadamente a cada uno de los componentes propuestos en la definición semántica (Sireci, 1998). En los estándares para la creación de tests psicológicos y educativos (AERA *et al.,* 1999) se subraya la necesidad de someter la batería de ítems a una evaluación por parte de jueces seleccionados por tener unas características similares a la población objetivo o por ser expertos en la temática. Lynn (1986) sugiere un mínimo de 3 jueces, aunque esta cifra no está consensuada (Gable y Wolf, 1993) y va a depender de los intereses del investigador y de la complejidad del constructo. A los jueces seleccionados se les debe facilitar la definición operativa del constructo a evaluar y la batería de ítems creada. Tienen que estimar si los ítems son pertinentes para la faceta para la que han sido creados, a la vez que indicar si el número de ítems por componente refleja adecuadamente la importancia atribuida en la definición. Además, se aconseja recoger información sobre si los ítems están redactados de manera clara. Esta estimación debe hacerse a través de una escala numérica de entre 5 y 7 puntos (Haynes *et al.,* 1995) o con cualquier otro procedimiento que permita cuantificar la valoración de los jueces, aunque luego los datos a tener en cuenta sean meramente descriptivos, usándose si acaso el acuerdo inter-jueces para eliminar los ítems más problemáticos. Si después de esta fase se opta por modificar algunos ítems o escribir nuevos elementos, el proceso de evaluación debe repetirse. Finalizada la valoración de los ítems por parte de los jueces, el autor/adaptador debe informar con claridad qué ítems han sido eliminados y por qué, a la vez que debe especificarse cuál es finalmente la batería de ítems conservada.

Análisis estadístico de los ítems (D)

Tras el análisis cualitativo de los ítems, y para seleccionar los mejores del total de los disponibles, deben llevarse a cabo distintos estudios dirigidos a analizar métricamente las propiedades de dichos ítems, análisis que está basado en una serie de índices que van a permitir valorar a cada uno de ellos desde un punto de vista estadístico. El primer análisis de la batería de ítems suele basarse en la administración de éstos a una muestra de participantes con unas características semejantes a la de la población objetivo y que según Osterlind (1989) bastaría con que estuviese compuesta por entre 50 y 100 participantes. Esta administración debe hacerse tal y como si el autor tuviera la prueba definitiva desarrollada, y la intención es detectar los ítems más problemáticos, dificultades para comprender las instrucciones, errores en el formato del instrumento, erratas, etc. En el caso de que el número de ítems sea demasiado elevado se recomienda que éstos sean divididos y pasados a muestras diferentes. Con los resultados de este primer estudio, y con los ítems seleccionados, debe repetirse el proceso con la intención de obtener más garantías sobre éstos, pero ahora con una muestra de mayor tamaño, mínimo 300 participantes o entre 5 y 10 por ítem (Martínez-Arias, 1995) y también de características similares a la población objetivo. Es aconsejable que este proceso se repita (validación cruzada), dadas las fluctuaciones que los estadísticos derivados de las puntuaciones de los ítems presentan en función de la muestra con la que esté trabajándose.

¿Qué cálculos estadísticos deben presentarse o tenerse en cuenta en los análisis métricos de los ítems? En tests como los que aquí están siendo tratados (tests referidos a la norma), la selección de los ítems debe estar basada en que éstos tengan la capacidad de poner de manifiesto las diferencias existentes entre los individuos. Debido a esto, el objetivo es conseguir un grupo de ítems que maximice la varianza del test, seleccionando para ello a aquellos con un elevado poder de discriminación, alta desviación típica, y con puntuaciones medias de respuesta situadas entorno al punto medio de la escala (Nunnally y Bernstein, 1995). No obstante, la decisión de eliminar o conservar un ítem debe estar basada en una valoración conjunta de todos los índices estadísticos, junto con una consideración de los aspectos conceptuales que motivaron la creación de éste. La razón por la que presentar la media y desviación típica de cada ítem está en las propiedades de la curva normal. Así, son considerados ítems adecuados aquellos con una desviación típica superior a 1 y con una media situada alrededor del punto medio de la escala (simetría próxima a 0). Nótese que para este criterio de decisión, uno debe asegurar que en la muestra de estudio están representados todos los valores del constructo. Suele ocurrir que en los primeros trabajos sobre escalas clínicas sean usadas muestras de universitarios. Al analizar las puntuaciones medias de algunos ítems de dichas escalas, la media suele ser baja si puntuaciones altas indican cierta problemática (depresión, ansiedad, etc.) y la desviación típica escasa. Por ello, debe tenerse en cuenta la muestra con la que se trabaja, y en el caso de ser muestras no representativas, ser cuidadoso a la hora de tomar las decisiones, ya que un ítem puede resultar "problemático" para una muestra determinada pero muy adecuado para otra. Para calcular la discriminación de un ítem normalmente se recurre al coeficiente de correlación corregido entre la puntuación en el ítem y la total obtenida en la dimensión a la que éste pertenezca (aunque claro está, este cálculo no agota las posibilidades). Este procedimiento busca aumentar la consistencia interna de la dimensión. Se consideran adecuados valores mayores o iguales a 0,25-0,30 (Nunnally y Bernstein, 1995). En este sentido hay que apuntar que cuanto más elevadas sean estas correlaciones para todos los ítems de una faceta, mayor será la fiabilidad de este componente calculada a través de la consistencia de las respuestas a través de los ítems. Esto hace que cuando el análisis de la discriminación de los ítems sea efectuado siguiendo el procedimiento indicado, suela incluirse el cálculo de la fiabilidad de dicho componente a través del índice de consistencia interna. Así, aparece cuál sería el valor de este índice si un ítem determinado es eliminado. La idea es que si la eliminación de un ítem aumenta la fiabilidad, éste debe ser descartado (se recuerda de nuevo que estas decisiones siempre deben tener en cuenta criterios teóricos). No obstante, sobre esta regla de decisión, la cual los autores de este trabajo rechazan si se aplica indiscriminadamente, se discutirá en los siguientes párrafos, al igual que dentro del apartado de fiabilidad. Quiere resaltarse que si un constructo está configurado por distintas facetas o componentes, los cálculos de discriminación tienen que hacerse por faceta, y no considerando el total de la escala (véase un ejemplo de esta forma de proceder en Whiteside y Lynam, 2001). La idea es que cada componente del constructo debe ser una categoría homogénea de contenido y "aislada" en la medida de lo posible del resto de componentes, ya que de lo contrario no puede sostenerse su separación como categorías distintas de un mismo constructo.

La puntuación individual que se obtenga para cada componente debe tener elementos comunes con las otras facetas delimitadas, ya que han sido propuestas como integrantes de un mismo constructo. Sin embargo, estos elementos comunes no deben superar un límite, ya que de lo contrario no podría sostenerse que son componentes distintos. En esta dirección es donde resulta aconsejable hacer mención a nuevos análisis a incluir en el estudio estadístico de los ítems.

Junto al coeficiente de correlación ítem-total corregido deben efectuarse los análisis de correlación entre la puntuación de los ítems que configuran un componente y la puntuación total de los componentes que no sean el de pertenencia teórica. Algunos autores plantean que debe existir una diferencia positiva a favor del primer análisis de al menos dos décimas (Jackson, 1970). Junto con el cálculo anterior, es aconsejable incluir la correlación media inter-item. Este aspecto necesita de cierta reflexión. Cuando está elaborándose un instrumento con la intención de verificar una propuesta conceptual sobre un constructo determinado, un criterio normalmente tenido en cuenta es trabajar para que los componentes del constructo sean homogéneos. Para lograr esta homogeneidad, tradicionalmente ha sido usado el índice de fiabilidad de consistencia interna, intentándose que éste fuera lo mayor posible como indicativo de una faceta homogénea. Así, cuando es calculado el índice de discriminación de los ítems, se opta por eliminar los que provocan que la consistencia interna del componente se incremente (tal y como anteriormente ha sido explicado). Sin embargo, resulta necesario distinguir consistencia interna de homogeneidad. Tal y como Cortina (1993) especifica, la consistencia interna es el grado en el que los ítems de un componente o faceta están intercorrelacionados, mientras que la homogeneidad se refiere a si los ítems de ese componente evalúan fundamentalmente sólo a éste. Esto significa que la consistencia interna es algo necesario pero no suficiente para conseguir una faceta homogénea, o dicho de otro modo, puede tenerse un grupo de ítems altamente inter-correlacionados y que aún así no puedan ser considerados como representativos de un único componente (Clark y Watson, 2003). Debido a esto es recomendable llevar a cabo el cálculo de la correlación media entre los ítems.

La forma de proceder a la hora de llevar a cabo la correlación media inter-item, consiste en calcular ésta para los ítems de cada uno de los componentes por separado, para posteriormente calcularla teniendo en cuenta los posibles cruces entre componentes. La lógica que debe subyacer para interpretar los datos es que la correlación media entre los ítems de componentes distintos tiene que ser positiva para poder concluir que forman parte de un mismo constructo, pero inferior a la aparecida para los ítems de un mismo componente (una diferencia de al menos dos décimas según Clark y Watson, 2003). Llegado a este momento es donde algunos autores recomiendan el uso del análisis factorial como procedimiento inicial incluido dentro del estudio de las propiedades de los ítems (Floyd y Widaman, 1995). De hecho, las conclusiones que pueden derivarse de los cálculos de correlación media inter-item son fácilmente obtenidas a través de la "lectura" del patrón de saturaciones factoriales. Así, la técnica de análisis factorial podría ser usada en esta fase, no todavía como procedimiento de validez interna, sino como herramienta para la selección de ítems homogéneos (véase el siguiente apartado para un comentario más detallado sobre el análisis factorial).

Aunque en otro apartado será presentada la información acerca del cálculo de la fiabilidad de la prueba definitiva, el lector ha podido observar como ya se ha hecho referencia a ésta. Así, es recomendable hacer notar que es justo en esta fase de la construcción/adaptación cuando el autor tiene una primera estimación de la fiabilidad a partir de un grupo de ítems que todavía no son los definitivos. Este hecho es importante porque en función de sus intereses teóricos, con esta estimación podrá hacerse balance sobre la necesidad de elaborar nuevos ítems para poder alcanzar una fiabilidad concreta o bien reducir el número planificado de ítems por componente, ya que con ese menor número se alcanzan los objetivos deseados en cuanto a la fiabilidad, reduciéndose así el tiempo de evaluación.

En ocasiones el objetivo no es tanto "aislar" los componentes homogéneos de un constructo, sino conseguir a través de las puntuaciones en éstos, poder predecir un criterio externo. Esta forma de proceder está más enfocada a la validez de criterio que a la homogeneidad, aunque en la actualidad se recomienda un equilibrio entre ambos objetivos a la hora de seleccionar los ítems (Smith *et al.,* 2003). El interés por la predicción de un criterio suele estar muy presente en los tests dirigidos a evaluar una variable clínica (depresión, ansiedad, etc.), donde aparte de que los componentes del constructo consigan diferenciarse, interesa que éstos se vean muy relacionados con indicadores considerados como una señal distintiva de lo que esté evaluándose. Aquí los cálculos a efectuar son los mismos que en el caso anterior, pero además debe incluirse la correlación entre el ítem y la puntuación total obtenida en una variable externa. Aquí el análisis debe centrarse en la comparación de las correlaciones obtenidas entre los ítems y variables criterio, y entre éstos y variables teóricamente no relacionadas con los ítems. Nótese que para llevar a cabo estos cálculos es necesario que junto a la batería de ítems que está analizándose también se haga uso de otros cuestionarios sobre constructos con los que éstos deben y no deben relacionarse. Así, en la bibliografía especializada (Edwards, 2001; Paunonen y Ashton, 2001; Smith *et al.,* 2003) se insiste en que ésta sea la forma de proceder habitual, intentándose proporcionar información acerca tanto de la homogeneidad como de la validez de criterio de los ítems e insistiéndose en que los cálculos sean efectuados para cada faceta del constructo y no considerando la prueba en su totalidad.

Quiere resaltarse que los cálculos anteriormente especificados no agotan todas las posibilidades (véase Muñiz, 1998). Sería adecuado, por ejemplo, presentar información que asegure que todas las alternativas de respuesta de los ítems son elegidas, llevar a cabo el denominado análisis diferencial de los ítems (DIF), etc. Sea como fuere, para este apartado se ha optado por estos cálculos por considerarse imprescindibles para el desarrollo o adaptación de escalas como las que vienen analizándose (Haynes *et al.,* 1995) y por ser los más tratados en los artículos especializados.

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

Una vez que los ítems seleccionados han pasado filtros, tanto teóricos como estadísticos, el objetivo es ver si éstos empíricamente se "agrupan" tal y como teóricamente había sido predicho. Ahora la meta es explorar la estructura interna de la escala, su dimensionalidad. El estudio de la dimensionalidad de una prueba, el cual estaría for-

mando parte de los trabajos destinados a la obtención de evidencias de validez interna, persigue evaluar "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). Quizá sea uno de los aspectos más tratados de la psicometría y el análisis factorial la técnica que tradicionalmente sirve para distinguir este campo de estudio. En esta fase debe utilizarse una estrategia que permita contrastar estadísticamente la hipótesis del investigador basada en cómo van a agruparse los ítems. Es aquí donde tiene que hacerse uso de las denominadas ecuaciones estructurales y proceder con la puesta en marcha del análisis factorial confirmatorio. A pesar de lo comentado, puede recomendarse que antes de proceder con la aplicación de las ecuaciones estructurales sea utilizado un procedimiento exploratorio de análisis factorial como método de validación cruzada de todos los análisis de ítems previos, y como forma de llevar a cabo una primera "exploración" de la estructura interna del cuestionario (Floyd y Widaman, 1995). Posteriormente, y como elemento esencial, debería usarse el análisis factorial confirmatorio.

No es este el lugar para profundizar sobre el uso (y abuso) del análisis factorial. Por lo que aquí respecta, el asunto esencial es aclarar los elementos a tener en cuenta a la hora de su aplicación y, por lo tanto, información a considerar en un trabajo donde esta técnica sea usada. De este modo viene a asumirse que el investigador responsable es conocedor de los problemas asociados a la aplicación del análisis factorial exploratorio sobre las puntuaciones de los ítems de un test (véase apartado "cómo engañarse a uno mismo con el análisis factorial" en Nunnally y Bernstein, 1995, pp. 599-601) y que a partir de la reflexión sobre estos problemas finalmente se ha decidido aplicar este procedimiento de cálculo. Sobre el tamaño muestral necesario para poder aplicar cualquier procedimiento factorial habría que decir que la respuesta no es única. Stevens (1992) aconseja que al menos se cuente con 5 participantes por cada variable (ítem). Como regla general hay que decir que distintos estudios ponen de manifiesto que, considerando el número de ítems que normalmente es utilizado en los artículos instrumentales, con 300 participantes se obtienen soluciones fiables (Snook y Gorsuch, 1989). Como es sabido, el análisis factorial exploratorio proporciona agrupamientos de variables (o de ítems en el caso de aplicarse sobre una sola escala) en función de criterios matemáticos basados en la correspondencia entre éstos para que el responsable del análisis los "interprete". La primera estructura factorial proporcionada, o solución de primer orden, suele ser difícil de juzgar, recurriéndose a una rotación o combinación lineal de los factores iniciales. Suele ser esta solución factorial de segundo orden la que debe ser considerada a la hora de discutir los resultados.

Existen distintos "tipos" de análisis factoriales exploratorios a poder usar, al igual que de rotaciones. ¿Qué procedimientos deben aplicarse? Con un número de ítems superior a 20, la corroboración de que existe una adecuada inter-correlación entre ellos, y con muestras de participantes de al menos 300, las diferencias entre las soluciones factoriales proporcionadas por distintos métodos son despreciables (Snook y Gorsuch, 1989). Por ello, y teniendo en cuenta la mayor facilidad de aplicación e interpretación, se recomienda el uso del análisis de componentes principales (ACP) (Cortina, 1993) y la rotación ortogonal Varimax (véase Comrey, 1988 o Floyd y Widaman, 1995 para una

mayor información). Este método de rotación es aplicado partiendo del supuesto de la independencia entre los componentes del constructo, o por los intereses teóricos del investigador de separar lo máximo posible los factores resultantes. En el caso de que se tengan evidencias de una alta relación entre los componentes (alrededor de 0,40 según Nunnally y Bernstein, 1995), el método debe ser oblicuo, entre los que destacan la rotación Promax y la Oblimin directa. No obstante, téngase en cuenta que esto no deja de ser más que una regla genérica y que el uso de éstos u otros procedimientos debe justificarse.

Un requisito indispensable para la aplicación del análisis factorial exploratorio es que las variables (ítems) se encuentren relacionadas entre sí; es decir, la matriz de correlaciones debe ser tal que puedan "localizarse" agrupamientos relevantes entre variables. Por ello es necesario presentar antes de la aplicación del análisis los estimadores que aseguren que la matriz de correlaciones es apropiada, siendo las pruebas de elección la de esfericidad de Bartlett y el índice de Kaiser-Meyer-Olkin (KMO), recomendándose el cálculo de ambas (Cortina, 1993). Quiere resaltarse que la adecuación de la matriz de correlaciones no es el único criterio que debe analizarse antes del uso del análisis factorial exploratorio. En la bibliografía especializada son tratados otros aspectos, aunque su presentación desbordaría los objetivos de este trabajo. Sólo ha sido apuntada la relevancia de la matriz de correlaciones por ser el factor de mayor influencia en los resultados. El investigador debe analizar la solución rotada y, en concreto, la información a presentar debe incluir una tabla donde queden claros el número de factores resultantes, las saturaciones de los ítems en dichos factores, la cantidad de varianza explicada por cada factor y la proporción de varianza del ítem que es explicada por los componentes principales (comunalidad o h^2). Siguiendo las recomendaciones de Stevens (1992), deben señalarse las saturaciones que son al menos iguales a 0,40, aunque otros autores proponen un criterio menos restrictivo (0,25-0,30) para cuando las muestras están formadas por más de 300 participantes (Floyd y Widaman, 1995). En el caso de que un mismo ítem presente valores de saturación por encima del límite en más de un factor, deberán aparecer las dos saturaciones.

Un aspecto esencial del análisis factorial exploratorio es la interpretación de los resultados y, en concreto, decidir cuántos y cuáles parecen ser los componentes relevantes para explicar la dimensionalidad del test. Aunque las alternativas son variadas, quizá el criterio más potente es el de la replicación, es decir, que usando los mismos ítems en muestras diversas pueda concluirse que las soluciones factoriales son congruentes o similares, y para lo que deberán usarse índices pertinentes de valoración y no simplemente la "inspección visual". No obstante, los autores de este trabajo quieren resaltar algunos aspectos estimados como importantes para el caso de la "interpretación" de los factores en función de una sola administración.

El análisis factorial exploratorio no entiende de Psicología. Esto significa que el análisis sólo "agrupa" correlaciones similares, pero que esta agrupación puede ser debida a más elementos que los propiamente conceptuales. Así, ha sido puesto de manifiesto que una batería de ítems con formato similar aunque conceptualmente heterogéneos, redactados la mitad en sentido positivo y la otra mitad en sentido negativo van a agruparse en dos claros factores. Uno recoge a todos los ítems con sentido positivo y

el otro a los negativos. Así, el formato de los ítems puede pesar más que la significación conceptual y sin una exploración detallada de los resultados lo empírico, pero irrelevante, puede prevalecer sobre lo psicológicamente sustantivo. Así, cualquier criterio usado debe estar caracterizado por la flexibilidad en su aplicación. Los autores de este artículo quieren mostrar su total rechazo hacia aquellos trabajos donde a partir de un agrupamiento de ítems inadecuadamente derivados, y según la estructura factorial resultante, se señala el "descubrimiento" de los aspectos subyacentes de una realidad psicológica. 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, no recomendándose su uso en estas condiciones (Nunnally y Bernstein, 1995). Dado lo apuntado, viene a recordarse que los procedimientos exploratorios sirven para "indagar" y que, por lo tanto, esta indagación debe ser posteriormente sometida a confirmación. Así, aunque el autor deberá informar sobre los criterios tomados en cuenta para concluir sobre qué factores y cuántos son vistos como determinantes (véase una revisión en Ferrando, 1996 o Martínez-Arias, 1995), estos criterios deberán verse relacionados en la discusión con referentes teóricos, a la vez que se declara la momentaneidad de las conclusiones hasta que la replicación sea suficiente y la confirmación de la hipótesis llevada a cabo.

Cuando el objetivo es confirmar si la estructura empírica de la escala se corresponde con la teórica, la técnica de análisis no debe ser exploratoria. Aunque en los últimos años está observándose un incremento en el uso de los procedimiento confirmatorios en la publicaciones referentes a la creación/adaptación de tests, su uso aún está poco generalizado, siendo lo común la aplicación de procedimientos exploratorios (Batista-Foguet, Coenders y Alonso, 2004). No es este el espacio para explicar el uso de estrategias confirmatorias a través de los modelos de ecuaciones estructurales (se le recomienda al lector el trabajo de Batista-Foguet y Coenders, 2000); sin embargo, sí se resaltan las etapas que el encargado de su aplicación debe considerar (Batista-Foguet *et al*., 2004) y los datos que deben hacerse explícitos al publicar los resultados. El autor debe especificar claramente cuál es el modelo (forma en la que los ítems se agrupan) que pretende someterse a prueba, recomendándose el uso simultáneo de otros modelos alternativos para analizar el ajuste comparativo, y facilitando la información que asegure que los modelos pueden ser contrastados en función de los requisitos de las ecuaciones estructurales. El tamaño de la muestra debe ser adecuado para este tipo de análisis. Así, no debe aplicarse este método de análisis con muestras inferiores a 200 participantes, aunque dependerá del número de ítems, componentes propuestos, etc. (Batista-Foguet *et al*., 2004). Una vez especificado el o los modelos, comprobado que la aplicación de la técnica es posible y que la muestra es adecuada, debe seleccionarse el método de estimación a usar para concluir si lo teóricamente propuesto se ajusta a los datos empíricos. Cuando se usan ítems con una escala de respuesta tipo Likert, la recomendación es tratar a las puntuaciones como datos categoriales no continuos, ya que son en realidad las propiedades de dichas puntuaciones (Jöreskog y Sörbom, 1993).

El método recomendado es la estimación robusta de máxima verosimilitud (ML) o, en el caso de no normalidad de las puntuaciones, la estimación robusta de mínimos cuadrados no ponderados (ULS) o la asintóticamente libre de distribución (WLS) de Browne (1984), aunque aquí la muestra debería ser superior a 1000 participantes para que los resultados pueden considerarse estables. Una vez aplicado algunos de los métodos de estimación debe evaluarse la adecuación de los modelos sometidos a prueba, recordándose que "la etapa de diagnóstico nunca será capaz de demostrar que un modelo es correcto, sino, a lo sumo, incapaz de demostrar que es incorrecto" (Batista-Foguet *et al.,* 2004). Para esta valoración existe gran variedad de índices, recomendándose usar a la vez algunos de éstos (Tanaka, 1993). La decisión tradicional basada en el valor de la chi-cuadrado se desaconseja por ser muy susceptible a variaciones en función del tamaño de la muestra. Los más usados (Browne y Cudeck, 1993; Jöreskog y Sörbom, 1993), y que no se ven afectados por los grados de libertad y el tamaño muestral, son el índice de bondad de ajuste (*Goodness of Fit Index, GFI*), índice ajustado de bondad de ajuste (*Adjusted Goodness of Fit Index, AGFI*), error cuadrático medio de aproximación (*Root Mean Square Error of Approximation*; *RMSA*), residuo estandarizado cuadrático medio (*Standard Residual Mean Root*; *SRMR*) y el índice de ajuste no normado (*Non-Normed Fit Index*; *NNFI*). Todos estos índices se juzgan globalmente a partir de que se alcance o no unos valores establecidos como "correctos". El autor debe informar de esos valores y cuándo el ajuste global del modelo no va a ser considerado incorrecto. Tras la valoración global del modelo, normalmente puede observarse una determinada falta de ajuste de los datos a las predicciones y la salida del programa proporciona ciertas soluciones que conducirían a mejorar los resultados (eliminando ítems o asumiendo un nuevo parámetro, como la correlación antes no tenida en cuenta entre dos factores). Sea como fuere, de esta forma de proceder acaba concluyéndose sobre si la forma hipotetizada acerca de la distribución de los ítems puede mantenerse y cuáles van ser definitivamente los ítems a tener en cuenta. No obstante, esta conclusión debe ser tomada como parcial y momentánea, reclamándose que estos análisis se repitan con otra u otras muestras.

Estimación de la fiabilidad (F)

Al principio de este trabajo se indicó que iban a presentarse distintos apartados, y que éstos iban a aparecer según el orden en el que deben ser tenidos en cuenta a la hora de crear/adaptar un test. En función de esto algunos lectores podrán plantearse por qué es justo en este momento cuando aparece el estudio de la fiabilidad del test. La decisión de presentar justo en este momento el apartado referido a la fiabilidad tiene una intención clara para los autores. Así, el investigador que haya procedido para la creación/adaptación de un test, tal y como en este informe ha sido explicado, no es justo hasta este momento cuando debe tener el agrupamiento "definitivo" de ítems por componente para llevar a cabo los estudios de obtención de evidencias externas de validez. Por ello, es justo en este momento donde la estimación de la fiabilidad alcanza su relevancia, dado que ésta va a ser sobre la escala finalmente delimitada y no sobre formas experimentales previas. Lo comentado no significa que no deban hacerse estimaciones de la fiabilidad hasta que se llegue a esta fase. Ya ha sido indicado que ciertas

decisiones para la elección/descarte de ítems van a sustentarse en dicha fiabilidad y que su cálculo ocurre paralelamente a la construcción de la escala desde el momento en el que se emprenden los análisis estadísticos de los elementos. De esta forma, lo que quiere hacerse es llamar la atención sobre el hecho de que para considerar que tenemos estudiada la fiabilidad de un test primeramente tenemos que saber cuál es el test, en cuanto a qué componentes lo integran y cuáles son sus ítems, de lo contrario los análisis no dejan de ser aproximaciones previas que no pueden llevar a concluir sobre la fiabilidad del instrumento.

Para la aplicación de cualquiera de los métodos disponibles para la estimación de la fiabilidad, y antes de entrar a comentarlos, debe asegurarse que el tamaño de la muestra, la situación de evaluación y las características de los participantes son adecuadas. En cuanto al tamaño de la muestra las recomendaciones son las mismas que las apuntadas para el caso del análisis de la estructura interna, y es que ésta se sitúe entre los 200 y 300 participantes (Clark y Watson, 2003), aunque dependiendo de nuevo de la estrategia de cálculo. Los participantes deben tener características semejantes a las de la población objetivo del test y las condiciones de evaluación tienen que ser similares a aquellas que han sido usadas para los estudios previos y para las que la escala se diseñó. Pueden establecerse tres métodos para obtener estimaciones del coeficiente de fiabilidad (Traub, 1994): a) método de formas paralelas; b) método basado en el test-retest; c) método centrado en la aplicación única de la prueba. Los dos primeros son los que tienen una mayor relación con la conceptuación original de fiabilidad de Spearman, en cuanto a la semejanza que tendrían las puntuaciones obtenidas con un test si éste es pasado a una misma persona en momentos distintos, es decir, la correlación entre las puntuaciones de un test a través de medidas repetidas. Para el cálculo de la fiabilidad siguiendo estos procedimientos existen varios problemas a los que el investigador debe enfrentarse (Muñiz, 1998), entre los que destacan el hecho de contar realmente con formas paralelas de un test, el efecto de la experiencia o práctica debida a la primera evaluación sobre la segunda, los cambios "reales" que se producen en la variable medida o saber cuál es el intervalo de tiempo aconsejable para llevar a cabo una nueva administración del mismo test o de una forma paralela de éste. Frecuentemente, los constructores/adaptadores de un test tienen que conformarse, por razones prácticas más que de idoneidad, con estimar la fiabilidad a partir de una única administración del instrumento, y en estos casos los procedimientos son los del cálculo de la consistencia interna; éste consiste en la correlación entre las puntuaciones de partes distintas de un mismo test (generalmente entre dos mitades de un test puntuadas separadamente y que deben considerarse formas equivalentes) o en la covariación existente entre todos los ítems. En el caso de ítems con una escala tipo Likert, el índice de consistencia interna por excelencia es el *alpha* de Cronbach. Sin embargo, éste es un indicador imperfecto de la consistencia interna de una faceta, y esto a pesar de su uso extendido. En concreto, este índice está muy influido por el número de ítems, llegándose a señalar que para escalas o componentes con un número de ítems situado entre 30 y 40, los valores van a ser anormalmente altos, por lo que no es recomendable su uso (Cortina, 1993). En estos casos el cálculo de la correlación media inter-item resulta más adecuado (o recurrir a construir dos mitades de la prueba), por no verse influido por el número de

ítems. Se recomienda que en general el valor de la correlación media inter-item esté situado entre 0,15 y 0,50, aunque habría que matizar los objetivos de la escala (véase Briggs y Cheek, 1986).

En las pruebas que normalmente se construyen/adaptan en Psicología Clínica y de la Salud, lo frecuente es delimitar constructos multi-componente, es decir, definidos por varias facetas que se postulan como elementos a considerar aisladamente. En estos casos resulta totalmente inadecuado calcular el alpha de Cronbach para el total de la escala, ya que éste debería ser estimado para cada faceta del constructo. Cuando en un caso como éste, además del cálculo para el total de la escala, éste se efectúa por componente, lo que ocurre es que el índice de consistencia de la escala total es superior al de cada componente aisladamente, aunque como ya ha sido especificado, si las facetas son distintas, al agruparlas no debería aumentar la consistencia de las respuestas, sino disminuir. Esto pondría de relieve varias cosas. Sin ánimo de ser exhaustivos, la primera tiene que ver con el efecto del número de ítems sobre el *alpha* de Cronbach; si tenemos una escala con 6 componentes y cada uno de ellos tiene 7 ítems, al considerar todos a la vez, el cálculo se realiza sobre 42 ítems, y en estos casos el índice resultante es artificialmente alto (Cortina, 1993). La segunda tiene que ver con un factor que puede motivar que las respuestas de los sujetos sean consistentes a través de ítems referidos a facetas distintas; este factor se refiere al propio formato de los ítems. Veamos el siguiente ejemplo. Imagine el lector una prueba con dos facetas independientes, una referida al cálculo aritmético y otra a ortografía, y que quiere evaluarse a una muestra de universitarios. Ejemplo de un ítem del factor aritmética sería "¿Cuánto es el resultado de multiplicar 4 x 4?; del componente ortografía "A continuación puede ver una palabra escrita de tres formas distintas, señale la que es adecuada ortográficamente (suponga que la palabra presentada es hombre)". Estos ejemplos poco reales de ítems sirven para poner de manifiesto un efecto que se daría al calcular el índice de consistencia interna en estos casos, y es que aunque los ítems estén referidos a facetas distintas, los universitarios se mostrarían muy consistentes en sus respuestas dada la facilidad de éstos y no por lo que pretenden medir en sí mismo los ítems (ortografía y aritmética). El ejemplo anterior puede extrapolarse a algunos auto-informes diseñados en Psicología Clínica y de la Salud, en donde sus ítems, por la forma de preguntar o afirmar tan semejante y genérica, y por las opciones de respuesta facilitadas, van a provocar que la respuesta de los participantes sean "similares" a través de éstos, y que por lo tanto lo que mida la prueba sea más un factor denominado "formato del instrumento" y con una elevada consistencia interna (téngase en cuenta este comentario para reflexionar igualmente sobre los resultados del análisis factorial). Así, de nuevo se insiste en la necesidad de interpretar los valores con cierto distanciamiento y sin olvidar los aspectos teóricos que deben tenerse siempre en cuenta a la hora de discutir los resultados, y esto dando por hecho que los pasos iniciales de construcción/adaptación del test (en cuanto a elaboración teórica y desarrollo de ítems) han sido adecuados.

Otro aspecto problemático que puede ser comúnmente observado en las publicaciones que presentan la información referente al estudio de la fiabilidad de una prueba tiene que ver con una premisa que parece intentar cumplirse por todos los medios, y que es que la consistencia interna de la escala cuanto mayor mejor. Así, en las primeras

fases de desarrollo de un instrumento, cuando es observado que la eliminación de un ítem provocaría el aumento de la consistencia de un componente, la decisión suele ser descartar éste de inmediato. Sin embargo, desde aquí viene a rechazarse esta forma de proceder y se llama la atención sobre la paradoja de la atenuación (Loevinger, 1957). La paradoja de la atenuación vendría a poner de manifiesto que aumentar la consistencia interna más allá de cierto punto va a tener un efecto sobre la disminución de la validez de constructo. Dado que el valor de consistencia interna depende de la inter-correlación entre los ítems, una forma de aumentar ésta es haciendo que los ítems estén estrechamente inter-correlacionados. Sin embargo, los ítems altamente inter-correlacionados son ítems que están referidos a un mismo aspecto, por lo tanto redundantes y sin la capacidad para representar los elementos variados de un constructo. Por ejemplo, los ítems como "Estoy contento", "Estoy alegre" y "Estoy animado" podrían formar parte de un mismo componente evaluado (afectividad positiva); dado su contenido, uno podría casi asegurar que los participantes van a responder de la misma forma a cada uno de éstos y por lo tanto la inter-correlación será muy elevada, y la consistencia interna seguramente superior a 0,90. Sin embargo, el constructo afectividad positiva no está siendo bien recogido, el elemento capturado es muy concreto y, por lo tanto, faltaría información relevante que debido al uso de ítems similares no está siendo considerada. Los ítems que se agrupen dentro de un mismo componente deben estar relacionados, pero a su vez debe asegurarse que cada uno de éstos esté dedicado a representar aspectos diferentes de dicho componente. Esto no significa que para aumentar la validez de un instrumento tenga que "sacrificarse" la consistencia, lo que quiere decirse es que 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. Así, una vez alcanzados índices situados entre 0,70 y 0,80 en el *alpha* de Cronbach (Cortina, 1993), el objetivo debe ser representar adecuadamente el constructo medido (aunque la correlación entre algunos ítems sea moderada), de lo contrario tendremos un instrumento con mucha fiabilidad para la evaluación de nada. Cuando el objetivo es de diagnóstico o clasificación, la fiabilidad mínima calculada a través de la consistencia interna debe ser de 0,80. Sin embargo, cuando los intereses son de investigación y su aplicación no va a tener consecuencias directas sobre los participantes, la fiabilidad puede considerarse adecuada si está entorno a 0,70 (Nunnally y Bernstein, 1995).

Teniendo en cuenta la información presentada en este apartado, de nuevo se insiste en una aplicación reflexionada de las opciones disponibles. El autor debe justificar sus decisiones, éstas no deben estar motivadas por ser lo "normal", la aplicación del procedimiento seleccionado debe ser cuidadosa y la interpretación de los valores tiene que ser de nuevo reflejo del dominio teórico del constructo.

Obtención de evidencias externas de validez (G)

Aunque un constructo haya podido ser definido cuidadosamente en las primeras etapas teóricas y aunque esa definición se haya visto finalmente respaldada empíricamente a través de unos ítems concretos, en modo alguno puede entenderse que las puntuaciones que se puedan obtener con esa escala son indicativas de dicho constructo

o que puedan usarse para el objetivo inicialmente planteado. Para llegar a esta conclusión es necesario obtener las pertinentes evidencias externas de validez. Las evidencias de validez externa deben basarse en el estudio de las relaciones entre el test y a) un criterio que se espera prediga éste (validez de criterio), b) otros tests que supuestamente miden lo mismo o con otros constructos con los que tendría que mostrar relación (validez convergente); y c) otras variables teóricamente relevantes y de las que debería diferenciarse (validez discriminante) (AERA *et al*., 1999). Como puede deducirse, cuando se habla de evidencias externas de validez se produce una vuelta a la elaboración teórica inicial. De hecho, se trataría de establecer si aparecen las relaciones teóricamente predichas entre las puntuaciones obtenidas con el instrumento de evaluación y otras variables externas delimitadas como importantes para el constructo evaluado. Así, el análisis de la validez externa de las puntuaciones de un test no es ni más ni menos que el intento por "ubicar" al constructo en un entramado teórico significativo, dándole "coherencia psicológica". De esta forma, estos estudios supondrían el soporte a partir del cual interpretar las puntuaciones de la herramienta de evaluación y, por lo tanto, el modo de otorgar significado psicológico a un dato numérico (Paz, 1996). Los autores de este artículo quieren advertir algo que con frecuencia suele ser pasado por alto. Llegados a este momento del proceso de construcción/adaptación de un test, lo que los estudios siguientes deben aportar no es algo "exclusivo" de la tarea que supone esta construcción o adaptación. Lo que quiere decirse es que en cualquier campo de investigación, al estudiar una dimensión, se busca integrar ésta en un esquema general que le dé sentido y donde pueda dársele utilidad y significado. De esta forma, el estudio de la faceta psicológica que sea, tiene que enmarcarse en una tradición empírica y teórica previa, para posteriormente proceder a analizar si la propuesta resulta adecuada (validación). Con esto viene a sostenerse que el proceso de validación asentado en estos objetivos no es algo que se observe sólo dentro de esta fase de construcción/adaptación de un test, sino que debe tratarse más como una finalidad y quehacer común dentro de una disciplina científica. Así, esto tiene su reflejo en que no hay una "metodología de estudio" particular para esta tarea de validación externa de un instrumento, sino que la clave son las relaciones teóricamente propuestas como significativas, aplicándose la metodología y diseño más conveniente (experimental, cuasi-experimental o no experimental), en función de los intereses teóricos.

Para la elaboración o revisión de un trabajo dirigido a la obtención de evidencias externas de validez de un test, los autores deben justificar éste a partir de las teorías de referencia y resultados de investigación previos (lo cual debería estar concretado en la definición sintáctica de la variable realizada en las primeras fases de construcción/adaptación), y su puesta en marcha debe seguir los criterios consensuados para cualquier investigación (Bobenrieth, 2002), además de tener en cuenta los particulares de la metodología concreta que haya decidido usarse, como por ejemplo la experimental (Ramos-Álvarez y Catena, 2004). Este es el motivo por el que para esta fase en cuestión no podrá observarse al finalizar el informe una tabla resumen dedicada a los aspectos más sustantivos de ésta, remitiéndose al lector a los trabajos ya citados (Bobenrieth, 2002; Ramos-Álvarez y Catena, 2004) o a cualquier otro encargado de tratar las normas para la publicación de artículos en Ciencias del Comportamiento.

A pesar de lo dicho en el último párrafo, resulta conveniente apuntar algunas recomendaciones a tener en cuenta para este tipo de estudio. Así, para la obtención de evidencias de validez convergente y discriminante, el uso de la matriz multirrasgo-multimétodo (MRMM) resulta un procedimiento a considerar; éste, propuesto por Campbell y Fiske (1959), se basa en la evaluación de un mismo constructo a través de distintos métodos y de distintos constructos con igual método. Con los datos obtenidos se lleva a cabo una correlación cruzada que acaba facilitando una matriz de correlaciones donde deben tenerse en cuenta las aparecidas entre el test con otras medidas del mismo constructo pero con distinto método y las del test con otros constructos con el mismo método. Si las primeras correlaciones son altas se concluye a favor de una adecuada validez convergente y si además éstas se diferencian de las segundas, se concluye a favor de una adecuada validez discriminante (véase Paz, 1996 para un ejemplo). Cuando se recurre al uso de correlaciones entre las puntuaciones del test y otras medidas, ya sea a través del procedimiento MRMM o del uso tradicional de éstas, tiene que recordarse el efecto que la fiabilidad baja o moderada de los instrumentos tiene sobre estas correlaciones. Así, y dado el interés teórico que debe preceder a la puesta en marcha de estos estudios, en esos casos debe llevarse a cabo una corrección por atenuación de éstas, para así poder explorar la relación entre las variables en caso de que el efecto de esa fiabilidad se viera controlado.

La aportación que realizan las ecuaciones estructurales al estudio de la validez externa no puede dejar de ser comentada. Así, ya fue apuntado su uso para la confirmación de la dimensionalidad del instrumento. Sin embargo, las ecuaciones estructurales igualmente permiten poner a prueba modelos basados en las relaciones esperadas entre constructos distintos, tanto si éstas son tratadas como correlaciones simples, como si se propone a uno de éstos como "causa" del otro u otros. En el caso de decidirse por esta estrategia de trabajo, los comentarios a tener en cuenta son los ya efectuados para el caso del análisis factorial confirmatorio, aunque ahora el núcleo de interés no es en qué medida los ítems son predichos por una dimensión determinada, sino hasta qué punto, tal y como se tenía predicho, ciertos constructos se relacionan entre sí.

Según la información que hasta este momento ha sido presentada en este apartado, se deduce fácilmente que la validez externa de un test es analizada fundamentalmente a través del coeficiente de correlación de Pearson entre éste y otras medidas. Cuando el interés es explorar la validez de criterio, es decir, en qué medida una variable es predicha por la puntuación o puntuaciones en el test, debe explorarse la proporción de varianza del criterio que puede predecirse a partir del test, recurriéndose al coeficiente de determinación, que no es otra cosa que elevar al cuadrado el coeficiente de correlación de Pearson. En estos casos suele hacerse uso de los distintos tipos de análisis de regresión lineal múltiple (según los intereses), ya que generalmente no sólo se relacionan dos variables, predictora y predicha, sino que suele usarse un conjunto de tests para así analizar tanto la aportación conjunta de éstos para predecir el criterio, como la individual al considerar cada variable aisladamente mientras que se controla el efecto de las demás (correlación parcial). El análisis de regresión lineal múltiple en sus distintas posibilidades es una herramienta muy útil a tener en cuenta a la hora de llevar a cabo los estudios dirigidos a explorar la aportación diferencial de distintos constructos

sobre una variable de interés. Sin embargo, no está exento de problemas, y el autor deberá justificar su uso y reflexionar sobre sus condiciones de aplicación y verdaderas aportaciones a la hora discutir los resultados (véase Martínez-Arias, 1995 o Nunnally y Bernstein, 1995 para un análisis de los factores a tener en cuenta a la hora de usar el análisis de regresión lineal simple y múltiple). A pesar de haberse destacado el uso del análisis de regresión lineal, no debería perderse de vista que esta técnica de análisis no agota las posibilidades. Así, y aunque dependiendo del número de variables y de sus particularidades métricas, el hecho es que lo habitual sería poder acudir a casi la totalidad de técnicas de análisis multivariado, por lo que el investigador debería ser conocedor de las características principales de cada una de éstas, para así aplicarlas en las circunstancias adecuadas (véase Muñiz, 1998 para profundizar en el uso de estas técnicas dentro de los estudios de validez externa).

Para finalizar los comentarios referidos a la obtención de evidencias externas de validez habría que hacer una reflexión sobre un aspecto muy tratado en los últimos estándares para la creación de tests psicológicos y educativos (AERA *et al.*,1999). Se trata del estudio dirigido a explorar si las evidencias de validez obtenidas para determinadas muestras y en contextos concretos pueden generalizarse sin necesidad de nuevos estudios de validez. Aquí el problema residiría en que los coeficientes de validez obtenidos se ven afectados por la variabilidad de las muestras, los distintos instrumentos usados, criterios considerados en cada estudio, etc. Así, el objetivo es determinar si la variación en dichos coeficientes es simplemente producto de estas influencias "inevitables", o bien existen otras variables no tenidas en cuenta y que deberían ser pues incluidas dentro de los estudio de validez del instrumento en cuestión. Para este objetivo la propuesta se ha centrado en hacer uso de variaciones del meta-análisis tradicional y que viene a suponer más que una "puesta en común" y equiparación de los resultados logrados en diversos estudios. Como puede deducirse, las conclusiones son tomadas cuando éstas se ven precedidas por un amplio bagaje de resultados y, por lo tanto, por un conjunto de estudios numeroso. Esto de nuevo pone de relieve que hasta que un test cuenta con todas las evidencias necesarias de validez interna y externa, junto con las encargadas de reflejar la fiabilidad, las fases a seguir son varias, su puesta en marcha debe ser repetida y el tiempo que se necesita es igualmente elevado. Así, no hay test fiable y válido sin un trabajo estructurado, sistemático y prolongado detrás.

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.

Barbero, M.I. (2003). *Psicometría*. Madrid: Universidad Nacional de Educación a Distancia.

Batista-Foguet, J.M. y Coenders, G. (2000). *Modelos de ecuaciones estructurales*. Madrid: La Muralla.

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.

Bobenrieth, M. (2002). Normas para la revisión de artículos originales en Ciencias de la Salud. *Revista Internacional de Psicología Clínica y de la Salud/Internacional Journal of Clinical and Health Psychology, 2,* 509-523.

Briggs, S.R. y Cheek, J.M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54,* 106-148.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62-83.

Browne, M.W. y Cudeck, R. (1993). Alternative ways of assessing model fit. En K.A. Bollen y J.S. Long (Eds.), *Testing Structural Equation Models* (pp. 136-162). Thousand Oaks: Sage.

Buela-Casal, G., Carretero-Dios, H. y De los Santos-Roig, M. (2002). Estudio comparativo de las revistas de Psicología en castellano con factor de impacto. *Psicothema, 14,* 837-852.

Buela-Casal, G. y Sierra, J.C. (2002). Normas para la redacción de casos clínicos. *Revista Internacional de Psicología Clínica y de la Salud/Internacional Journal of Clinical and Health Psychology, 2,* 525-532.

Buela-Casal, G., Sierra, J.C., Carretero-Dios, H. y De los Santos-Roig, M. (2002). Situación actual de la evaluación psicológica en lengua castellana. *Papeles del Psicólogo, 83,* 27-33.

Campbell, D.T. y Fiske, D.W. (1959). Convergent and discriminant validation by Multitrait-Multimethod Matrix. *Psychological Bulletin, 56,* 81-105.

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

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: APA.

Comrey, A.L. (1988). Factor-analytic methods of scale development in personality and clinical psychology. *Journal of Consulting and Clinical Psychology, 56,* 754-761.

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

Edwards, J.R. (2001). Multidimensional constructs in organizational behavior research: An integrative analytical framework. *Organizational Research Methods, 4,* 144-192.

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

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.

Gable, R.K. y Wolf, J.W. (1993). *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings*. Boston: Kluwer Academic.

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

Grant, J.S. y Davis, L.L. (1997). Selection and use of content experts for instrument development. *Research and Nursing & Health, 20,* 269-274.

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.

Jackson, D.N. (1970). A sequential system for personality scale development. En C.D. Spielberger (Ed.), *Current topics in clinical and community psychology* (vol. 2) (pp. 61-96). Nueva York: Academic Press.

Jöreskog, K.G. y Sörbom, D. (1993). *LISREL 8. User´s referente guide*. Chicago, IL: Scientific Software.

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.

Lynn, M. (1986). Determination and quantification of content validity. *Nursing Research, 35,* 382-385.

Martínez-Arias, R. (1995). *Psicometría: teoría de los test psicológicos y educativos*. Madrid: Síntesis.

Montero, I. y León, O. (2002). Clasificación y descripción de las metodologías de investigación en Psicología. *Revista Internacional de Psicología Clínica y de la Salud/Internacional Journal of Clinical and Health Psychology, 2,* 503-508.

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

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

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

Muñiz, J. y Hambleton, R.K. (2000). Adaptación de los tests de unas culturas a otras. *Metodología de las Ciencias del Comportamiento, 2,* 129-149.

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

Navas, M.J. (2001). *Métodos, diseños y técnicas de investigación psicológica*. Madrid: Universidad Nacional de Educación a Distancia.

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

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

Paunonen, S.V. y Ashton, M.C. (2001). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81,* 524-539.

Paz, M.D. (1996). Validez. En J. Muñiz (Ed.), *Psicometría* (pp. 499-103). Madrid: Universitas.

Pelechano, V. (1997). Prólogo. En G. Buela-Casal y J.C. Sierra (dirs.), *Manual de evaluación psicológica. Fundamentos, técnicas y aplicaciones* (pp. 31-35). Madrid: Siglo XXI.

Pelechano, V. (2002). Valoración de la actividad científica en psicología? Pseudoproblema, sociologismo o ideologismo? *Analisis y Modificacion de Conducta, 28*, 323-362.

Prieto, G. y Delgado, A.R. (1996). Construcción de los ítems. En J. Muñiz (Ed.), *Psicometría* (pp. 139-170). Madrid: Universitas.

Ramos-Álvarez, M.M. y Catena, A. (2004). Normas para la elaboración y revisión de artículos originales experimentales en Ciencias del Comportamiento. *International Journal of Clinical and Health Psychology, 4,* 173-189.

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.

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

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

Smith, G.T., y McCarthy, D.N. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment, 7,* 300-308.

Snook, S.C. y Gorsuch, R.L. (1989). Component analysis versus common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106,* 148-154.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Tanaka, J.S. (1993). Multifaceted conceptions of fit in structural models. En K.A. Bollen y J.S. Long (Eds.), *Testing Structural Equation Models* (pp. 10-39). Thousand Oaks: Sage.

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.

Whiteside, S.P. y Lynam, D. R. (2001). The five factor model and impulsivity: Using a structural model of personality to understand impulsivity. *Personality and Individual Differences, 30,* 669-689.

ANEXO 1. Normas para la elaboración y revisión de investigaciones instrumentales.

JUSTIFICACIÓN DEL ESTUDIO (A)

		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.			

DELIMITACIÓN CONCEPTUAL DEL CONSTRUCTO A EVALUAR (B)

		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.			

CONSTRUCCIÓN Y EVALUACIÓN CUALITATIVA DE ÍTEMS (C)

		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.			

ANÁLISIS ESTADÍSTICO DE LOS ÍTEMS (D)

		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 las 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).			
C.7.	Se especifican con claridad los criterios a considerar para la selección-eliminación de los ítems.			
C.8.	Los cálculos estadísticos efectuados resultan pertinentes.			
C.9.	Los resultados (cualitativos y cuantitativos) se discuten con claridad.			
C.10.	Las decisiones sobre los ítems tienen en cuenta cuestiones teóricas.			
C.11.	Se especifica claramente que ítems son eliminados y por qué.			
C.12.	Los ítems seleccionados quedan claramente delimitados.			

ESTUDIO DE LA DIMENSIONALIDAD DEL INSTRUMENTO (ESTRUCTURA INTERNA) (E)

		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.	La información sobre la muestra es completa y pertinente.			

E.6.	La muestra de estudio tiene características similares a las 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 (teórica 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.			

ESTIMACIÓN DE LA FIABILIDAD (F)

		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.			