1
VOL. 31 - 2010
ENERO - ABRIL

REVISTA DEL CONSEJO GENERAL DE COLEGIOS OFICIALES DE PSICÓLOGOS
ISSN 0214 - 7823

PAPELES DEL PSICÓLOGO

METODOLOGÍA AL SERVICIO DEL PSICÓLOGO

ANÁLISIS DE DATOS - METODOLOGÍAS CUALITATIVAS
TEORÍAS, PROPIEDADES Y NUEVOS TIPOS DE TESTS

Papeles del Psicólogo, 2009. Vol. 31(1), pp. 2-6
http://www.cop.es/papeles

Presentación

METODOLOGÍA AL SERVICIO DEL PSICÓLOGO
METHODOLOGY FOR PSYCHOLOGISTS

Vicente Ponsoda
Departamento de Psicología Social y Metodología. Facultad de Psicología
Universidad Autónoma de Madrid

La publicación de las últimas secciones monográficas de *Papeles del Psicólogo* coincidió con las discusiones en mi universidad sobre la relevancia de la metodología para la formación del futuro psicólogo (en las reuniones preparatorias del nuevo plan de estudios) y con la concesión a nuestro grupo de investigación *Modelos y Aplicaciones Psicométricos* de una cátedra de patrocinio por parte del Instituto de Ingeniería del Conocimiento. Los tres acontecimientos están detrás de la propuesta inicial de la presente sección monográfica.

Cada vez somos más los convencidos de las ventajas recíprocas que se derivan del acercamiento entre la metodología y la profesión. La universidad genera conocimiento que no es fácilmente accesible al profesional. El objetivo principal de los nuevos másteres que las universidades ofrecen es precisamente acercar ambos mundos. Por otra parte, los profesionales se enfrentan a problemas, en ocasiones de no fácil solución, que pueden resultar muy fructíferos para los investigadores. Pondré dos ejemplos de mi campo de investigación.

Al poco de empezar a aplicarse los tests adaptativos informatizados, y cuando tanto los investigadores como los profesionales estaban maravillados con su eficacia (consiguen reducir a la mitad el número de ítems o el tiempo necesario para la aplicación del test), los profesionales se dieron cuenta de algunos de sus puntos débiles. Uno de ellos es que una parte importante del banco de ítems, en ocasiones hasta un 80% de los ítems disponibles (Hornke, 2000), no se administraba nunca. Sí, han leído bien, un 80%. Para hacer un buen test, se elabora un banco de por ejemplo 500 ítems, muy cuidado, se estudian todos ellos minuciosamente, se eliminan los ítem defectuosos... y, después, el nuevo test es tan eficaz (al presentar solo los ítems buenísimos) que ¡400 de los 500 ítems del banco no se administran nunca! Resultó evidente la necesidad de incorporar al test procedimientos de control de la exposición que hicieran posible la administración de muchos más ítems del banco y que redujeran la tasa de exposición de los que se administraban en todos o en casi todos los tests. Este objetivo se debía conseguir sin que el test perdiese precisión (Revuelta y Ponsoda, 1998). En la última década se ha investigado mucho sobre los métodos de control de la exposición (revisados en Georgiadou, Triantafillou y Economides, 2007). Un problema advertido por los profesionales ha dado lugar a una considerable investigación que ha generado nuevas soluciones.

Un segundo ejemplo. Es bien conocido que las medidas de personalidad mejoran moderadamente la predicción del desempeño laboral (Salgado y Moscoso, 2008) y que, en los procesos de selección de personal, los candidatos pueden falsear sus respuestas al cuestionario de personalidad y responder como creen que lo haría el candidato ideal (Salgado, 2005). El falseamiento de las respuestas puede ser incluso más importante en los procesos selectivos a ciertos puestos de la Administración, en los que en algún caso más de un 90% de los que aprueban la oposición han pasado por academias en las que son entrenados (por cierto, por psicólogos) a dar las respuestas que les faciliten la consecución de la plaza (Garrido, Ponsoda, Olea y Abad, 2009). La Academia de Policía Local de la Comunidad de Madrid requirió nuestra colaboración para que buscásemos soluciones que permitiesen seguir aplicando medidas de personalidad en los procesos selectivos, pero ganando en seguridad de que tales medidas informaban realmente de las características del candidato. El problema de la detección del falseamiento en los tests de personalidad aplicados en los

Correspondencia: Vicente Ponsoda. *Departamento de Psicología Social y Metodología. Facultad de Psicología. Universidad Autónoma de Madrid. c/ Iván Pavlov, 6. 28049 Madrid. España. E-mail: Vicente.ponsoda@uam.es*

procesos de selección de personal tiene un considerable interés y la investigación lo está abordando desde vías muy distintas (véase Salgado, 2005). Una vía reciente consiste en intentar obtener el nivel de los candidatos en los rasgos de personalidad tras eliminar la contaminación que la Deseabilidad Social pudiese haber introducido, mediante modelos factoriales (Ferrando y Anguiano-Carrasco, 2009) o modelos de Teoría de la Respuesta al Ítem de las medidas ipsativas[1] (Leenen, Ponsoda y Romero, 2009). De nuevo, un problema que han detectado y preocupa a los profesionales (el falseamiento de respuestas a los tests de personalidad) ha promovido y sigue promoviendo investigación.

En los dos ejemplos anteriores se ve natural y hasta inevitable la colaboración entre el metodólogo y el psicólogo; sin embargo, pese a que siempre ha habido preocupación en el campo metodológico por llegar al mundo psicológico, la sensación, a veces, es de cierto desaliento. Borsboom (2006) pretende encontrar una explicación de por qué los avances en metodología, incluso los muy interesantes, no llegan más fácilmente a los psicólogos investigadores no metodólogos. En un interesante y provocador artículo, cuyo título es "El ataque de los psicómetras", propone diversas explicaciones: a) de índole teórica (como el desconocimiento de otras teorías de los tests que no sean la Teoría Clásica), b) pragmática (como la escasa preparación cuantitativa de los estudiantes de psicología) y c) sustantiva (como la escasez de teorías psicológicas lo suficientemente precisas como para poder ser expresadas mediante modelos formales). En cuanto a las estrategias a seguir para resolver la falta de comunicación, propone escribir buenos libros de metodología, desarrollar programas informáticos que faciliten la aplicación de los últimos desarrollos metodológicos, y participar activamente con los psicólogos en el estudio sustantivo, en vez de ser exclusivamente los responsables del diseño y análisis de los resultados. La presente sección monográfica pretende colaborar en la tarea de establecer un puente entre la metodología y la psicología, facilitando al psicólogo el contacto con contenidos metodológicos que pueden serle de utilidad.

El repaso de otras secciones monográficas de la revista *Papeles del Psicólogo* me sugirió la conveniencia de hacer algo similar en metodología. El psicólogo clínico, de las organizaciones… cuando lee lo que se publica hoy sobre el tratamiento de la depresión o sobre los mejores predictores del rendimiento laboral se encuentra técnicas, conceptos y metodologías que van a ser tratados en los artículos de este monográfico. Los metodólogos hemos hecho monográficos sobre los avances en nuestros campos de investigación, dirigidos preferentemente a los demás metodólogos y publicados en "nuestras" revistas, como *Methodology, Metodología de las ciencias del comportamiento*, *Psicologica* , *Psicothema*… pero no, que yo sepa, en revistas no metodológicas. En concreto, ninguno de los monográficos de la revista *Papeles del Psicólogo* ha sido metodológico; aunque sí ha publicado trabajos sobre metodología, mayormente sobre los tests. Por cierto, la mayor parte lo han sido por autores de la presente sección monográfica. La revista *Papeles del Psicólogo* es además el vehículo ideal para tal objetivo por su calidad, su enorme difusión, la doble versión inglés y español, y por ser de libre acceso.

Una segunda razón que impulsó el monográfico fue la discusión que teníamos en 2008 en muchos centros universitarios sobre el papel de la metodología en la formación de los psicólogos, como ocurre siempre que se está cambiando el plan de estudios. Los mismos argumentos que exponíamos a nuestros colegas sobre la conveniencia de que los futuros psicólogos conozcan lo básico de las herramientas metodológicas modernas, nos debieran impulsar hacer eso mismo con los psicólogos en activo que muy probablemente no estudiaron esos desarrollos en su día. Una tercera razón es la cátedra de patrocinio IIC-UAM[2]. Uno de sus objetivos es precisamente acercar al profesional de la psicología los desarrollos metodológicos que puedan resultarles de interés, mediante seminarios y otras actividades. Una de ellas pensamos que podría ser el presente monográfico.

Convencido de que tenía sentido intentarlo, el primer paso fue plantear la propuesta al Director de *Papeles del Psicólogo* (el profesor Serafín Lemos). Su respuesta al proyecto fue muy positiva. Quiero agradecer públicamente no solo su respuesta, sino su colaboración y apoyo durante todo el proceso. El siguiente paso fue ver si mis colegas estaban de acuerdo en participar.

En cuanto a la elección de contenidos, pensé en un conjunto de temas que fuesen novedosos, poco conocidos por los profesionales y que pudieran resultarles útiles. Los temas finalmente propuestos fueron el escalamiento multidimensional, el sesgo de los tests, los tests adaptivos informatizados, el análisis factorial confirmatorio, los modelos estructurales, el meta-análisis, la idea actual sobre la validación de las puntuaciones de los tests, las nuevas teorías de los tests, los tests de desempeño, la metodología observacional y las metodologías cualitativas. Por falta de espacio, quedaron fuera algunos, también interesantes, como el análisis multinivel, la adaptación de tests, diseños y metodologías concretas para la investigación aplicada, la evaluación de programas, etc. En cuanto al estilo, los artículos van dirigidos a los psicólogos profesionales por lo que deberían evitar formalismos innecesarios que pudiesen dificultar su comprensión. Me consta el esfuerzo que han hecho los autores para exponer rigurosamente técnicas y conceptos complicados en un lenguaje asequible.

Elegidos los temas, me dirigí a los expertos. Todos lo son sobradamente sobre la temática de su artículo. La psicología española ha progresado mucho en las últimas décadas. Algo similar ha ocurrido en metodología. Cuando empezamos la mayoría de los firmantes a publicar, apenas había contribuciones de metodólogos españoles en las revistas internacionales. Hoy, afortunadamente, no es así. Como verán en las referencias, todos tienen investigación reconocida en el campo sobre el que han escrito el artículo. Estoy profundamente agradecido a todos ellos por haber aceptado participar y por el entusiasmo e interés que han puesto en la tarea.

[1] Uno de los objetivos del proyecto del Ministerio de Ciencia y Tecnología Psi2008-01685/Psic "Estudio psicométrico de las medidas ipsativas" es la consecución de tales modelos. Varios artículos del monográfico tratan los modelos factoriales y TRI citados y las medidas ipsativas.

[2] http://www.iic.uam.es/CatedraMYAP/

CONTENIDOS

Originalmente la sección monográfica contenía once artículos. Cinco tienen que ver con lo que venimos llamando teorías de los tests, cuatro sobre análisis de datos y dos sobre metodologías específicas. Mientras estábamos preparando los artículos se ultimó el análisis de las respuestas a una encuesta sobre el uso de los tests. Este último artículo no es uno de los temas previamente seleccionados, ni sigue el formato de los demás, pero pareció adecuado incluirlo por su contenido e interés. Son doce, por tanto, el número total.

El artículo de Sánchez-Meca y Botella es un buen ejemplo del sentido del monográfico. Los profesionales de la psicología en su quehacer diario han de realizar diagnósticos, aplicar tratamientos, decidir qué variables evaluar en un candidato… En estas situaciones no resulta fácil saber cuál es la mejor prueba a aplicar, el tratamiento más adecuado, las variables más apropiadas… Como respuesta a esta necesidad ha surgido la Psicología Basada en la Evidencia, que pretende fomentar que el profesional tome las anteriores decisiones basándose en lo que la investigación ha encontrado sobre el asunto que le interesa. Resulta imprescindible, entonces, aplicar procedimientos que resuman adecuadamente los resultados de las investigaciones. El trabajo de Sánchez-Meca y Botella muestra las bases de uno de esos acercamientos: el meta-análisis o conjunto de técnicas que permiten resumir, por ejemplo, la evidencia acumulada en las distintas investigaciones sobre la eficacia de un tratamiento concreto. El artículo detalla los pasos de un meta-análisis y los muestra en un ejemplo.

El análisis factorial es una técnica multivariante de reducción de dimensionalidad. Se ha aplicado a muchos campos y en algunos, como el de la personalidad y las aptitudes, su colaboración ha excedido la de un mero instrumento de análisis, hasta el punto de que se habla de teorías factoriales de la personalidad y de la inteligencia. La técnica acepta las respuestas de un conjunto de personas en un conjunto amplio de variables iniciales (ítems, tareas, pruebas…) y devuelve un número reducido y el significado de los factores, dimensiones o variables latentes que dan cuenta de las relaciones observadas entre las variables iniciales. Se distingue entre el análisis factorial exploratorio y el confirmatorio. El artículo de Ferrando y Anguiano-Carrasco presenta una minuciosa descripción de ambos, que no solo va a resultar informativa a los que apenas conozcan las técnicas. Dan sus recomendaciones sobre asuntos controvertidos y comentan las muchas posibilidades del programa de ordenador FACTOR, de libre distribución, desarrollado por ellos (Lorenzo-Seva y Ferrando, 2005).

Los modelos de ecuaciones estructurales se aplican cada vez más, en muy distintos campos, y no son demasiado conocidos por muchos profesionales. El análisis factorial confirmatorio es un caso particular de estos modelos. Con frecuencia se utilizan para determinar la relación que existe entre dos o más variables latentes, medidas cada una con sus correspondientes variables empíricas (ítems, tests…). El artículo de Ruiz, Pardo y San Martín expone lo fundamental de estos modelos, cómo se construyen sus diagramas, su estructura y los pasos a dar en su elaboración, reparando en la evaluación del ajuste y en los tipos de relaciones que pueden establecerse entre las variables latentes. La exposición se complementa con un ejemplo muy ilustrativo en el que se determina la relación existente entre las variables latentes Estrés, Cansancio Emocional y Síntomas Psicosomáticos. La técnica ayuda a determinar qué variable afecta (y cómo) a las demás.

El escalamiento multidimensional, objeto del artículo de Arce, de Francisco y Arce, es una técnica multivariante que está a caballo entre el análisis de datos y la psicometría. La psicometría se ocupa tanto de los procedimientos para la evaluación de las personas (teorías de los tests), como de las características psicológicas de los objetos (escalamiento psicológico). Hay procedimientos de escalamiento unidimensional y multidimensional. El escalamiento multidimensional descubre la dimensión o dimensiones que están detrás de la similaridad que percibimos entre los objetos. Los procedimientos aceptan de entrada la información sobre las similaridades entre los objetos y proporcionan, principalmente, el número de dimensiones necesarias para dar cuenta de ellas y la posición de cada objeto en cada dimensión. Algunos procedimientos informan también de la importancia que cada evaluador asigna a cada dimensión. El artículo describe los principales procedimientos de escalamiento multidimensional, indica los pasos que requieren y los aplica a varios ejemplos (al escalamiento de prácticas deportivas y de marcas de coches, entre otros).

El primer artículo del bloque sobre los tests expone las ideas centrales de las dos principales teorías: la Teoría Clásica (TC) y la Teoría de Respuesta a los Ítems (TRI). Cuando hablamos de tests, nos referimos a los que uno encuentra en los catálogos y también a las escalas, cuestionarios, exámenes, etc. Nos vamos a referir de hecho en el monográfico a tipos de tests muy diferentes. De la importancia de la TC se pueden decir muchas cosas, aunque quizás basta con una: es una teoría que ha cumplido los 100 años y sigue aplicándose para la elaboración de tests, en todo el mundo, pese a sus deficiencias. La TRI resuelve una importante: con la TC las medidas que se obtienen dependen del test particular administrado y no resulta fácil la comparación entre las proporcionadas por tests distintos de la misma variable psicológica. La TRI, por el contrario, puede producir medidas que son comparables, pese a haberse obtenido con distintos tests. Otra ventaja importante de la TRI es que permite ubicar a los evaluados y a los ítems (sus dificultades) en la misma escala, lo que tiene interesantes aplicaciones: permite, por ejemplo, ubicar a los estudiantes y las tareas en el continuo que indica el dominio de una materia, lo que mostraría qué tareas probablemente sabría y no sabría hacer bien cada estudiante; o ubicar al paciente y los estímulos fóbicos en la misma escala indicadora del deterioro, lo que facilitaría la decisión sobre el orden de presentación de estímulos más apropiado. El artículo de Muñiz sobre las teorías de los tests resume muy bien las principales características de ambas teorías y responde a la pregunta que nos hacen los estudiantes cuando les explicamos estos contenidos por primera vez: “construyo el test, lo aplico, obtengo las puntuación de cada evaluado y ya está. ¿Para qué necesito las teorías de los tests?”

Dos propiedades psicométricas cruciales de los tests o, mejor, de las puntuaciones de los tests, son la fiabilidad y la validez. Las dos resultan imprescindibles para determinar la calidad de los instrumentos de medida. En ambos conceptos ha habido cambios

con los años, pero ciertamente el cambio ha sido mucho mayor en el concepto de validez. Al construir un instrumento de evaluación hemos de hacer las cosas de modo que asigne puntuaciones similares en sucesivas ocasiones, si no hay cambio en el nivel del evaluado en la variable que el test mide. Si es así, decimos que la fiabilidad o consistencia de las puntuaciones del test es adecuada. Los estudios de fiabilidad no informan de los usos razonables o justificables que podemos hacer con las puntuaciones. Hace dos o tres décadas, cuando muchos de nosotros estudiamos por primera vez estos conceptos, la validez era casi solo la capacidad del test para predecir un criterio externo y venía indicada por el coeficiente de validez. El artículo de Prieto y Delgado expone los principales indicadores de la fiabilidad, cómo obtenerlos y las recomendaciones sobre su aplicación. En cuanto a la validez, se expone el concepto actual, en el que desaparecen los "tipos" de validez y se habla más bien de estrategias de validación con las que se acumulan evidencias que justifiquen el uso que hacemos de las puntuaciones. Interesantes son también los comentarios sobre los errores más frecuentes en la interpretación de ambos conceptos, que también se dan entre los que han recibido formación en estos temas (Frisbie, 2005).

Fiabilidad y validez no son las únicas exigencias psicométricas del test. El test además ha de carecer de sesgo, ser justo. Queremos hacer un test que evalúe razonamiento numérico. Si un ítem midiera solo razonamiento, la probabilidad de acierto en el ítem de los que tienen un mismo nivel de razonamiento deberá ser la misma si el evaluado es nativo o extranjero, si hombre o mujer, si de alto o bajo status socioeconómico. Si así fuera, estaríamos ante un ítem sin funcionamiento diferencial (DIF). El problema es que incluso los ítems mejor hechos a veces miden más de una dimensión, pudiendo ocurrir que las personas de un subgrupo u otro difieran en la dimensión no principal. Por ejemplo, si el ítem tiene un enunciado con mucho texto, pudiera ocurrir que los extranjeros tuviesen dificultades en entenderlo y, a pesar de tener el mismo nivel de razonamiento, lo acertaran menos que los nativos. Sería un ítem con DIF. Si el test tiene varios ítems con DIF, el test en su conjunto podría dar una puntuación menor a la persona con bajo dominio del idioma. Es evidente que estamos ante un problema de validez. Sus puntuaciones no debieran usarse para tomar decisiones sobre el nivel de razonamiento numérico, pues reflejan además el nivel de dominio del idioma. Es práctica común, por tanto, realizar estudios de funcionamiento diferencial de los ítems y del test cuando se construye o adapta uno. Las profesoras Gómez, Hidalgo y Guilera han resumido en su artículo su mucha experiencia en la detección del funcionamiento diferencial. Describen los principales procedimientos, dan recomendaciones de uso y delimitan conceptos que no siempre resulta fácil distinguir, como impacto, sesgo y equidad.

En los últimos años está habiendo un considerable desarrollo de los tests de desempeño ("performance assessment"), que apenas ha llegado a nuestro país. En estos tests las tareas a realizar por los evaluados son básicamente las mismas que han o habrán de hacer en su día a día. Por ejemplo, las redacciones y ejercicios que hace el estudiante; realizar un diagnóstico a partir de los síntomas del enfermo, que hará el médico en el puesto de trabajo al que aspira; organizar el plan de acción de la unidad a partir de las demandas recibidas, que tendrá que hacer el futuro jefe de policía si consigue el ascenso. Las tareas son ciertamente distintas de las requeridas por los tests ordinarios, pero no lo son las exigencias de calidad psicométrica que han de cumplir. El artículo de Martínez Arias describe qué es y cómo se construye un test de desempeño y expone sus principales ventajas e inconvenientes. Una ventaja de estos tests es que, por ser las tareas más ricas y complejas, permiten evaluar características psicológicas no fácilmente medibles con los tests ordinarios. Dos de sus inconvenientes son que resultan más difíciles de puntuar y que suelen tener peores propiedades psicométricas, debido principalmente a su escaso número de ítems.

Además de los tests de desempeño, otros tipos de tests han tenido un enorme desarrollo en los últimos años. El artículo de Olea, Abad y Barrada describe cinco nuevos tipos de tests: los informatizados, los tests basados en modelos, los ipsativos, los conductuales y los situacionales. De los tests administrados por ordenador, los adaptativos informatizados son sin duda los que más atención investigadora y aplicada han captado. Se ha creado incluso una asociación internacional (IACAT) para su avance. Los tests basados en modelos requieren tener un modelo de cómo la persona responde al ítem, lo que permite predecir sus características psicométricas y evitar total o parciamente los costosos procesos de calibración. Los tests ipsativos están resurgiendo con fuerza como una posible vía de control de algunos sesgos de respuesta en los tests de personalidad, como la Deseabilidad Social. Algunas dificultades de los tests ordinarios de personalidad, por ejemplo, se deben a que se registra la conducta verbal, son autoinformes. En los tests conductuales se está interesado y se registra lo que el evaluado *hace* (en vez de lo que *dice que hace*) en tareas muy bien pensadas. Los tests situacionales se aplican preferentemente en selección de personal. Miden características psicológicas, como rasgos de personalidad, competencias… muy frecuentemente en un formato de opción múltiple y no en el formato de categorías ordenadas, que es el ordinario para la medida de este tipo de características psicológicas. El evaluado ha de elegir la acción de las que se le ofrecen que tomaría en la situación que el ítem describe. Del cambio de formato comentado se siguen interesantes propiedades. El artículo describe los cinco tipos de tests y los evalúa críticamente.

¿Cuál es la opinión de los psicólogos españoles sobre los tests? Hace exactamente 10 años se aplicó una primera encuesta sobre el uso de los tests en nuestro país (Muñiz y Fernández-Hermida, 2000). El monográfico incluye un artículo en el que los mismos autores exponen los resultados de una segunda encuesta, con similar objetivo, aplicada recientemente, y contestada por más de 3000 psicólogos colegiados, mayoritariamente mujeres (72%) y de especialidad clínica (70%). Hay varios resultados de interés. Sigue habiendo en todas las especialidades psicológicas interés por los tests, y la valoración, que era buena hace 10 años, ha mejorado un poco durante la última década. En los ítems "*los tests constituyen una excelente fuente de información si se combinan con otros datos psicológicos*" y "*Utilizados correctamente, los*

tests son de gran ayuda para el psicólogo" las medias están por encima de 4.4, en una escala de 1 (desacuerdo total) a 5 (totalmente de acuerdo), y superan ligeramente las obtenidas hace 10 años. Interesantes también las dudas y reticencias manifestadas ante la evaluación informatizada y la tele-evaluación mediante internet. Se desprende de la encuesta que los profesionales desearían más formación sobre los tests. El presente monográfico contribuye a atender esta demanda en lo relativo a la metodología necesaria para su correcta elaboración e interpretación. El artículo comenta el trabajo realizado por la comisión encargada de elaborar la norma ISO 10667, que pretende regular todo lo relativo a la evaluación de personas en contextos laborales, en la que participa el COP y el primer firmante del artículo.

El grupo de investigación dirigido por la profesora Anguera lleva muchos años en la investigación y aplicación de la metodología de la observación a muy distintos campos. Su artículo proporciona una visión global de dicha metodología, que tiene varios puntos fuertes. Uno, fundamental, al que la autora presta atención en el artículo, es que la observación permite captar la cotidianeidad, respetando el medio en el que se producen las conductas. Otro es que, a veces, no hay otras alternativas. A un adulto se le puede aplicar un cuestionario para que nos diga cómo se siente; pero no a un recién nacido, por ejemplo. El artículo describe las cuatro fases de un estudio que aplique esta metodología (la delimitación del problema, la recogida de datos, el análisis de los datos y la interpretación de los resultados) y repara en sus singularidades. La perspectiva cualitativa sería prioritaria en la fase de recogida de datos; en la fase de registro, los programas informáticos tienen hoy un papel decisivo; y en la última fase, de análisis, sería prioritaria la perspectiva cuantitativa. Resulta evidente, entonces, la complementariedad de las perspectivas cualitativa y cuantitativa. El artículo ilustra las posibilidades de aplicación de la observación a muy distintos campos y muestra varias situaciones en las que sería la metodología más apropiada (el estudio de la interacción entre niños y adultos, las discusiones en una pareja, la comunicación no verbal, entre otras).

En los últimos años están ganando importancia las metodologías cualitativas. De ellas trata el artículo de López, Blanco, Scandroglio y Rasskin. Comienza con una contraposición entre las prácticas cualitativas y cuantitativas, y repara en los criterios de calidad de unas y otras. El artículo responde a las críticas de subjetividad, de falta de sistematicidad y transparencia y de falta de generalidad de los resultados. La posición de los autores es que hay buenos y malos investigadores en los dos campos y que no es correcto creer que solo la metodología cuantitativa tiene controles de calidad. La parte central del trabajo describe 10 técnicas de recogida de información (análisis de material documental, observación, entrevista, historia de vida, técnicas grupales…) y 7 prácticas de análisis (análisis de contenido, descripción etnográfica, inducción analítica, análisis del discurso…). Se comenta brevemente cada una y se ilustra con varios ejemplos de investigaciones en que podrían ser adecuadas. Se indica, por último, en qué situaciones esta metodología es especialmente apropiada. Por ejemplo, cuando estamos ante una primera exploración de fenómenos desconocidos, cuando el interés principal es conocer los significados que dan las personas a sus acciones, o en situaciones de dinámicas interactivas de elevada complejidad.

Ningún artículo requiere de otro para su comprensión, por lo que el lector puede seguir el orden que desee. A medida que los vaya leyendo verá que varios hacen referencia a la complementariedad de enfoques. Se dice en uno de ellos, al hablar de la teoría clásica frente a la TRI, que lejos de ser contrapuestas son complementarias, como lo son el coche y el avión. Tampoco parece que sea necesariamente mejor el análisis factorial confirmatorio que el exploratorio. Los autores de los dos últimos artículos abogan por la complementariedad entre las metodologías cualitativa y cuantitativa. Dejo al lector frente a un amplio conjunto de opciones metodológicas con la tranquilidad de que sabrá dar a cada una su mérito.

REFERENCIAS

Borsboom, D (2006). The attack of the psychometricians. *Psychometrika, 71*, 525-440.

Ferrando, P.J. y Anguiano-Carrasco, C. (2009). Assessing the impact of faking on binary personality measures: An IRT-based multiple-group factor analytic procedure. *Multivariate Behavioral Research, 44*, 497-524.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24(3)*, 21-28.

Garrido, L.E., Ponsoda, V., Olea, J. y F. J. Abad (2009, septiembre). *Efectos de la deseabilidad social sobre los tests de personalidad en muestras con alto entrenamiento.* XI congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

Georgiadou, E., Triantafillou, E., y Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Descargado el 23-12-2009 de http://www.jtla.org.

Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicologica, 21*, 175-189.

Leenen, I., Romero, S. y Ponsoda. V. (2009, septiembre). Análisis bayesiano de datos ipsativos: una evaluación y comparación de modelos competitivos. XI congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga.

Lorenzo-Seva, U. y Ferrando, P.J. (2005). *Factor*. http://psico.fcep.urv.cat/utilitats/factor/

Muñiz, J., y Fernández-Hermida, J.R. (2000). La utilización de los tests en España. *Papeles del Psicólogo, 76*, 41-49.

Revuelta, J. y Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 4, 311-327.

Salgado, J. (2005). Personalidad y deseabilidad social en contextos organizacionales: implicaciones para la práctica de la psicología del trabajo y de las organizaciones. *Papeles del Psicólogo, 26*, 115-128.

Salgado, J. y Moscoso, S. (2008). Selección de personal en la empresa y las administraciones públicas: de la visión tradicional a la visión estratégica. *Papeles del Psicólogo, 29*, 16-24.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 7-17
http://www.cop.es/papeles

REVISIONES SISTEMÁTICAS Y META-ANÁLISIS: HERRAMIENTAS PARA LA PRÁCTICA PROFESIONAL*

SYSTEMATIC REVIEWS AND META-ANALYSES: TOOLS FOR PROFESSIONAL PRACTICE*

Julio Sánchez-Meca[1] y Juan Botella[2]
[1]*Universidad de Murcia.* [2]*Universidad Autónoma de Madrid*

La práctica profesional y la investigación psicológica han estado demasiado separadas. Sin embargo, con la llegada de la Psicología Basada en la Evidencia (PBE) esta herramienta metodológica se ha convertido en el mejor modo de unir las mejores pruebas con la práctica psicológica. La PBE preconiza que la práctica profesional esté basada en las mejores pruebas obtenidas desde la investigación psicológica. Las revisiones sistemáticas (RSs) y los meta-análisis (MAs) se consideran actualmente como las mejores herramientas para sintetizar las pruebas científicas respecto a qué tratamientos, intervenciones o programas de prevención deberían aplicarse para un determinado problema psicológico. Así pues, los psicólogos tienen que saber qué son las RSs y los MAs, cómo se hacen y, lo que es más importante, cómo podemos hacer valoraciones críticas de ellos. El propósito de este artículo es presentar las RSs y los MAs, así como alguna guía orientativa sobre cómo hacer una lectura crítica de ellos.
Palabras clave: *Psicología basada en la evidencia, Revisiones sistemáticas, Meta-análisis, Evaluación de intervenciones.*

Professional practice and psychological research have traditionally been separated. However, over the last two decades Evidence-Based Psychology (EBP) has become a very useful methodological tool for linking the best evidence to psychological practice. EBP proposes to base professional practice on the best evidence obtained from psychological research. Systematic reviews (SRs) and meta-analyses (MAs) are considered as the best tools to synthesize the scientific evidence as to which treatments, interventions, or prevention programs should be applied for a given psychological problem. Thus, psychologists need to know what SRs and MAs are, how they are done and, most importantly, how we can carry out critical appraisal of SRs and MAs. The purpose of this article is to present SRs and MAs, together with some guidelines to warrant a critical reading of them.
Key words: *Evidence-based psychology, Systematic reviews, Meta-analysis, Intervention evaluation.*

No cabe duda de que el ejercicio profesional de la psicología, sea cual sea su ámbito, pasa necesariamente porque éste se base en las mejores pruebas y evidencias científicas. Sin embargo, existe un desfase entre las técnicas que se aplican en la práctica profesional en un determinado momento temporal y los avances que la investigación ha alcanzado en ese momento. Este desfase se debe, básicamente, a dos causas. Por una parte, no existe mucha conexión entre el mundo de la práctica profesional y el de la investigación, que se hace fundamentalmente en las universidades. Por otra parte, hasta hace unas dos décadas, las ciencias del comportamiento se han caracterizado por sufrir una pobre acumulación del conocimiento científico, de forma que los avances científicos llegan muy lentamente a la práctica rutinaria. Estos factores hacen que el profesional de la psicología vea al mundo de la investigación como algo muy alejado de su práctica habitual y sin una utilidad que pueda materializarse en resultados aplicables de forma rápida y directa en su quehacer cotidiano.

Si esto es lo que ocurre con los propios psicólogos, peor aún es lo que ocurre con los profesionales de otras disciplinas o los políticos y gestores que tienen que adoptar decisiones y preguntan la experta opinión de los psicólogos. Como ya hemos señalado en otro sitio (Botella y Gambara, 2006a), el problema es que los datos procedentes de la psicología con frecuencia son confusos y contradictorios. Recientemente hemos asistido a debates sobre temas con fuerte carga ideológica, como por ejemplo la adopción por parte de parejas homosexuales, en los que los políticos que defendían cada posición iban acompañados de psicólogos, supuestamente expertos, que decían lo contrario que el otro, pero que eran presentadas como si fuera "la posición" que se desprende de la evidencia recogida por la psicología.

Este panorama, un tanto descorazonador, está cam-

Correspondencia: Julio Sánchez-Meca. *Dpto Psicología Básica y Metodología. Facultad de Psicología. Campus de Espinardo. 30100-Murcia. España. E-mail: jsmeca@um.es*
www.um.es/metaanalysis

(*) Este artículo ha sido financiado al primer firmante por el Fondo de Investigación Sanitaria, convocatoria de Evaluación de Tecnologías Sanitarias (Proyecto No: PI07/90384) y al segundo firmante por el Ministerio de Educación y Ciencia (Proyecto No: SEJ2006-12546/PSIC).

biando gracias a dos avances metodológicos que tratan de evitar ese desajuste entre práctica profesional e investigación: el enfoque de la Psicología Basada en la Evidencia (PBE),[1] por una parte, y las revisiones sistemáticas y el meta-análisis, por otra.

El enfoque de la PBE es una herramienta metodológica mediante la cual se pretende modificar el modo de trabajo del profesional de la psicología de forma que tome en consideración en sus decisiones cotidianas las mejores evidencias o pruebas científicas acerca de un determinado problema. El problema podría ser qué mejor técnica de tratamiento utilizar para tratar a un paciente con un determinado trastorno psicológico, o qué programa de intervención sería el más apropiado para prevenir ciertas conductas desadaptadas, o cuál es el mejor método de diagnóstico de un trastorno psicológico. Una vez formulado adecuadamente el problema, el enfoque de la PBE consiste en realizar una búsqueda de las evidencias o pruebas que pongan de manifiesto el mejor curso de acción. Para que esa búsqueda de información resulte operativa, se requiere del uso de las nuevas tecnologías de la información y la comunicación y, en especial, de los recursos de internet. Una vez localizadas las pruebas científicas, las cuales estarán publicadas en revistas especializadas del ámbito de la psicología, el siguiente paso que promueve la PBE es hacer un análisis crítico de dichas pruebas, el cual requiere poner en práctica los conocimientos que el profesional de la psicología tiene sobre métodos de investigación, diseños, análisis de datos e instrumentos de medida. Y como última fase, la PBE implica aplicar los hallazgos encontrados a la práctica profesional.

¿Y cuáles son los mejores hallazgos o evidencias científicas que pueden avalar la aplicación rutinaria de un tratamiento, un programa de prevención o una técnica de evaluación o de diagnóstico? Es un hecho aceptado por la comunidad científica en ciencias sociales y de la salud considerar que las evidencias científicas más confiables son las que proporcionan los estudios primarios basados en la realización de ensayos clínicos aleatorizados (ECAs), que implican asignación aleatoria de los participantes a las condiciones experimentales (Nezu y Nezu, 2008).

Sin embargo, es fácil que cuando intentemos seleccionar las evidencias sobre un determinado problema nos encontremos con numerosos estudios empíricos, todos los cuales han abordado esa misma pregunta. Esta acumulación de información puede bloquear la puesta en práctica del enfoque de la PBE, al hacer inviable la selección de los estudios relevantes y su lectura crítica en un tiempo lo suficientemente corto como para que los profesionales puedan afrontar esta tarea, teniendo en cuenta su elevada carga de trabajo. Y es aquí donde entran en juego las revisiones sistemáticas (RSs) y el meta-análisis (MA). Como un modo de superar la pobre acumulación del conocimiento en las ciencias sociales, las RSs y los MAs constituyen una metodología de investigación que tiene como objetivo acumular de forma sistemática y objetiva las evidencias obtenidas en los estudios empíricos sobre un mismo problema. De esta forma, la lectura de una RS o de un MA sobre el problema en cuestión permite ahorrar tiempo a los profesionales y les ofrece una visión conjunta de lo que las evidencias científicas dicen sobre ese problema. Además, tal y como se resume en la Tabla 1, la comunidad científica ha aceptado a los MAs como la metodología que permite ofrecer las mejores pruebas o evidencias sobre un problema, cuando los estudios empíricos acumulados son estudios experimentales (o ECAs).

TABLA 1
NIVELES DE CALIDAD DE LAS EVIDENCIAS O PRUEBAS CIENTÍFICAS

Nivel	Tipo de prueba
1	(a) Meta-análisis (homogéneo)[1] de ECAs (b) Un ECA (con I.C. estrecho) (c) Todos o ninguno[2]
2	(a) Meta-análisis (homogéneo)[1] de estudios de cohortes (b) Un estudio de cohorte (incluyendo un ECA de baja calidad) (c) Investigación "de resultados"[3]; estudios ecológicos
3	(a) Meta-análisis (homogéneo)[1] de estudios de casos y controles (b) Un estudio de casos y controles
4	(a) Series de casos (y estudios de cohortes y de casos y controles con pobre calidad)
5	(a) Opinión de expertos sin una valoración crítica explícita, o basada en la fisiología, en los "primeros principios"[4], o investigación basada en el criterio de "autoridades"

[1] Los resultados de los ECAs son homogéneos.
[2] Se cumple cuando todos los pacientes morían antes de que el tratamiento estuviera disponible, pero ahora algunos sobreviven a él; o cuando algunos pacientes morían antes de que el tratamiento estuviera disponible, pero ahora ninguno muere tras él.
[3] Estudia una cohorte de pacientes con el mismo diagnóstico y relaciona sus resultados clínicos con los cuidados que han recibido.
[4] Se trata de los principios patofisiológicos utilizados para determinar la práctica clínica.

[1] Muy acertadamente, Frías Navarro y Pascual Llobell (2003) aclaran que la traducción más correcta del nombre en inglés, Evidence-Based Psychology, sería Psicología Basada en Pruebas, más que Psicología Basada en la Evidencia. No obstante, dado que PBE es el término que más se utiliza, hemos preferido mantenerlo en este artículo.

Papeles del Psicólogo ya se ha hecho eco del enfoque de la PBE en los dos excelentes artículos de Frías Navarro y Pascual Llobell (2003) y de Pascual Llobell, Frías Navarro y Monterde (2004). Además, existen en castellano excelentes presentaciones del enfoque de la Práctica Basada en la Evidencia (Grupo de Atención Sanitaria Basada en la Evidencia, 2007; Navarro, Giribet y Aguinaga, 1999; Vázquez y Nieto, 2003). Es por ello, que este artículo se centra en el otro avance metodológico que, en nuestra opinión, se ha convertido en un elemento clave para ayudar a conectar la investigación con la práctica profesional: el meta-análisis. Entendemos que el profesional de la psicología debe conocer esta metodología, ya que la lectura crítica de estudios meta-analíticos puede serle de gran utilidad en su toma de decisiones cotidiana sobre qué tratamientos o técnicas de diagnóstico aplicar. Al mismo tiempo, la lectura de estudios meta-analíticos facilita la puesta en acción del enfoque de la PBE al ofrecer de forma integrada las mejores evidencias o pruebas científicas sobre un determinado problema, con el consiguiente ahorro de tiempo para el profesional. Por otra parte, la proliferación de estudios meta-analíticos en el ámbito de la psicología garantiza que, antes o después, todo profesional de la psicología tendrá que afrontar la tarea de leer críticamente algún estudio de este tipo y, en consecuencia, conocer esta metodología se está convirtiendo, hoy por hoy, en una necesidad.

En lo que sigue, abordamos qué es un MA y qué son las RSs, desarrollamos las etapas en que se lleva a cabo un MA, ilustramos esta metodología con un ejemplo real y presentamos una guía para ayudarnos a leer críticamente estudios meta-analíticos. Finalmente, terminamos con algunas reflexiones y lecturas sugeridas. Para una mayor profundización en las RSs y los MAs pueden consultarse diversas fuentes (Borenstein, Hedges, Higgins y Rothstein, 2009; Botella y Gambara, 2002; Cooper, 2010; Cooper, Hedges y Valentine, 2009; Littell, Corcoran y Pillai, 2008; Marín Martínez, Sánchez Meca, Huedo y Fernández, 2007; Marín Martínez, Sánchez Meca y López López, 2009; Petticrew y Roberts, 2006; Sánchez Meca, 1999, 2003, 2008; Sánchez Meca y Ato, 1989; Sánchez Meca y Marín Martínez, en prensa).

REVISIONES SISTEMÁTICAS Y META-ANÁLISIS

Una RS es una revisión de una pregunta formulada con claridad, que utiliza métodos sistemáticos y explícitos para identificar, seleccionar y valorar críticamente investigaciones relevantes a dicha pregunta, así como recoger y analizar datos de los estudios incluidos en la revisión (Martín, Tobías y Seoane, 2006). Las RSs surgen como un intento de salvar las limitaciones de las revisiones tradicionales, caracterizadas por ser cualitativas y carentes de una adecuada sistematización. Los que defendemos las ventajas de las RSs frente a las revisiones tradicionales nos basamos en la premisa de que al proceso de revisión de la literatura científica sobre cualquier tema se le deben exigir las mismas normas de rigor científico que cuando se hace una investigación empírica: objetividad, sistematización y replicabilidad de sus resultados. Es decir, el proceso de revisión de los estudios empíricos sobre un determinado problema es una tarea científica del mismo modo que lo es la realización de un estudio empírico.

Si en una RS somos capaces de cuantificar, mediante algún índice estadístico del tamaño del efecto, los resultados de cada estudio empírico integrado y de aplicar técnicas de análisis estadístico para extraer la esencia de dichos estudios, entonces una RS se convierte en un meta-análisis (MA). Un MA es, pues, una RS en la que se utilizan métodos estadísticos para analizar los resultados de los estudios integrados en ella (Littell et al., 2008). Esto implica que todo MA es una RS, pero no toda RS tiene por qué ser un MA. Existen RSs cualitativas en las que no se aplican métodos estadísticos para integrar los resultados de los estudios, sino valoraciones cualitativas de dichos resultados.

Debido a su mayor nivel de cuantificación y rigor, dentro de las RSs, son los MAs los tipos de revisión que nos ofrecen las evidencias más válidas sobre un problema (Cooper, 2010). Es por ello que este artículo se centra específicamente en cómo se hace y cómo se interpreta un MA. Además, de los diversos problemas que es posible estudiar mediante un MA, posiblemente los más útiles para el profesional de la psicología son aquéllos que tienen por objeto examinar la eficacia de diferentes tratamientos, programas de intervención o de prevención de trastornos psicológicos, psicosociales o de conducta. Por tanto, nuestro interés se dirige a los MAs de este tipo.

¿Qué puede ofrecernos un MA? Al aplicar técnicas estadísticas para integrar los resultados de un conjunto de estudios empíricos acerca de la eficacia de tratamientos o programas de intervención, un MA permite responder a preguntas tales como: (a) ¿cuál es la magnitud del efecto global de los diferentes tratamientos?; (b) ¿son homogéneos los resultados de eficacia alcanzados por los diferentes tratamientos?; (c) caso de que no sean homogéneos, ¿qué factores son los que pueden explicar esa

heterogeneidad en los resultados? y (d) ¿es posible formular un modelo explicativo que sea capaz de dar cuenta de dicha heterogeneidad en los resultados? Para dar respuesta a estas preguntas, un MA implica desarrollar los pasos típicos de una RS y aplicar técnicas estadísticas de integración.

¿Dónde podemos encontrar MAs? Prácticamente en cualquier revista de psicología es posible encontrar algún estudio meta-analítico sobre algún tema que pueda ser de interés para la práctica profesional. En concreto, revistas que publican con frecuencia este tipo de investigaciones son *Psychological Bulletin*, *Clinical Psychology Review*, *Journal of Applied Psychology* o *Journal of Consulting and Clinical Psychology*. En castellano se pueden encontrar en revistas como *Psicothema*, *International Journal of Clinical and Health Psychology*, etc. Además, a través del buscador Google Académico es fácil encontrar estudios meta-analíticos. Por último, merecen mención especial la Colaboración Cochrane y la Colaboración Campbell, que son dos organizaciones internacionales cuyo fin es promover la realización de estudios meta-analíticos de alta calidad acerca de la eficacia de las intervenciones en distintos ámbitos que tienen que ver con el desempeño profesional del psicólogo. Así, la Colaboración Cochrane contiene en su sitio web numerosas RSs y MAs en el ámbito de la psicología clínica, mientras que la Colaboración Campbell ofrece esto mismo en los campos de la Educación, los Servicios Sociales y la Criminología (Sánchez Meca, Boruch, Petrosino y Rosa Alcázar, 2002).[2]

FASES DE UN META-ANÁLISIS

La realización de un MA implica seguir las mismas etapas que en cualquier estudio empírico, si bien algunas de ellas tienen ciertas peculiaridades que es preciso clarificar. Básicamente, podemos plantear la realización de un MA en cinco etapas:

(1) Formulación del problema
(2) Selección de los estudios
(3) Codificación de los estudios
(4) Análisis estadístico e interpretación
(5) Publicación

(1) *Formulación del problema*. El primer paso consiste en formular de forma clara y objetiva la pregunta a la que se pretende responder. Esto implica definir de forma teórica y operativa los constructos psicológicos objeto de estudio. Por ejemplo, en un meta-análisis sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia (Sánchez Meca, Rosa Alcázar, Marín Martínez y Gómez Conesa, en prensa), se definieron conceptos clave tales como cuáles eran los tratamientos psicológicos objeto de estudio, qué es el trastorno de pánico con o sin agorafobia y qué medidas del resultado de la eficacia se iban a admitir en el MA.

(2) *Búsqueda de los estudios*. El siguiente paso consiste en definir los criterios de selección de los estudios. Debe tenerse en cuenta que la realización de un MA implica seleccionar estudios empíricos que tengan ciertas características similares en cuanto al diseño de la investigación (e.g., todos los estudios deben incluir al menos un grupo tratado y un grupo de control, ambos con medidas pretest y postest), para que sea posible aplicar a todos ellos un mismo índice del tamaño del efecto que permita su comparabilidad métrica. Por tanto, aunque los criterios de selección dependerán del MA en cuestión, no pueden faltar especificaciones relativas al tipo de diseños admisibles en los estudios, al modo en que se han medido las variables de resultado, a las características de los participantes y a las características de los tratamientos. Por ejemplo, en el MA antes citado, para ser incluidos en el mismo los estudios empíricos tenían que incluir, al menos, un grupo tratado y un grupo de control formado por personas adultas diagnosticadas con el trastorno de pánico con o sin agorafobia, ambos grupos con medidas pretest y postest y el tratamiento aplicado tenía que ser sólo psicológico, quedando excluidos los psicofármacos. Además, los estudios tenían que estar realizados entre 1980 y 2006.

Una vez fijados los criterios de selección de los estudios, se lleva a cabo el proceso de búsqueda de los mismos, para lo cual se precisa utilizar bases electrónicas (e.g., PsycInfo, MedLine, ERIC), consultar revistas especializadas y contactar con autores reconocidos en el tema para solicitarles estudios de difícil localización. La combinación de fuentes formales e informales en el proceso de búsqueda debe garantizar la máxima comprehensividad en dicho proceso, así como la localización de estudios publicados y no publicados, con objeto de poder examinar el sesgo de publicación. En el

[2] Pueden consultarse los sitios web respectivos de la Colaboración Cochrane (www.cochrane.org) y de la Colaboración Campbell (www.campbellcollaboration.org). Puede también consultarse el sitio web del Centro Cochrane Iberoamericano que la Colaboración Cochrane tiene en Barcelona (www.cochrane.es).

MA sobre el trastorno de pánico se consultaron las bases electrónicas PsycInfo, Medline, Psicodoc y la Cochrane Library. Además, se consultaron MAs previos, artículos, libros, capítulos de libro de revisión, así como revistas especializadas en psicología clínica, y también se contactó con autores reconocidos en dicho ámbito, todo ello con el objeto de localizar el mayor número posible de estudios empíricos que cumplieran con los criterios de selección.

(3) *Codificación de los estudios*. Una vez localizados y recuperados todos los estudios empíricos seleccionados, la siguiente etapa consiste en registrar las características de dichos estudios. Con este propósito, se elabora un Manual de Codificación de las características de los estudios que podrían actuar como moderadores de los resultados de eficacia de los tratamientos analizados. A partir del Manual de Codificación se elabora un Protocolo de Registro de las variables moderadoras. Aunque las características de los estudios que se deben codificar dependerán del propósito de cada MA, podemos clasificarlas en varios clusters o categorías. Así, se habla de variables de tratamiento, variables de los participantes, del contexto, metodológicas y extrínsecas.

Variables de tratamiento son aquéllas que tienen que ver con el tratamiento aplicado en la investigación. Son, pues, variables de tratamiento el tipo de tratamiento aplicado (e.g., terapia cognitiva, exposición in vivo, relajación profunda, etc.), la duración del tratamiento, su intensidad, el modo de aplicación (individual versus grupal), etc.

Las *variables de los participantes* tienen que ver con las características de éstos. Así, se consideran variables de los participantes la edad media de la muestra analizada, su composición por sexos, su extracción social, la gravedad del trastorno, etc.

Las *variables de contexto* hacen referencia al lugar en el que se ha realizado la intervención. Por ejemplo, el lugar puede ser un hospital, una clínica particular, un gabinete psicológico, en la escuela, en el propio hogar, etc. También puede catalogarse como variable contextual el hecho de que los pacientes reciban el tratamiento estando internados o bien en régimen ambulatorio.

Las *variables metodológicas* son aquéllas que tienen que ver con el diseño y la instrumentación del estudio empírico. Así, son variables metodológicas muy relevantes en un MA el tipo de diseño (experimental versus cuasi-experimental), el tamaño de las muestras, la mortalidad experimental, la inclusión de medidas pretest o sólo postest, la realización de análisis estadísticos por intención de tratar o sólo con los que completan el tratamiento, el uso de evaluadores ciegos, es decir, desconocedores de qué tratamiento está recibiendo el paciente que está siendo evaluado, o el criterio diagnóstico utilizado en el estudio para evaluar a los participantes. Todas estas características permiten valorar la calidad metodológica de los estudios y, en consecuencia, la posible existencia de sesgos en las estimaciones de los resultados.

Por último, también se suelen codificar *variables extrínsecas*, así llamadas porque son características de los estudios que, en principio, no deberían tener nada que ver con el proceso científico de una investigación, pero que sin embargo, en ocasiones pueden afectar a los resultados de los estudios. Entran dentro de esta categoría variables tales como la fuente de publicación (publicado versus no publicado), la formación de los autores del estudio (psicólogo, psiquiatra, etc.) o el año de realización del estudio.

El propósito de la fase de codificación de las características de los estudios no es otro que disponer de un conjunto de variables que puedan ser capaces de explicar la variabilidad de los resultados de eficacia de los diferentes estudios, una cuestión que es analizada estadísticamente en la siguiente fase. En el MA sobre el trastorno de pánico, la heterogeneidad en los resultados de eficacia exhibida por los diferentes estudios empíricos integrados podía deberse a que dichos estudios habían aplicado técnicas de tratamiento psicológico diferentes, con una duración diferente, sobre muestras de pacientes que podían variar en edad, composición por sexo y gravedad del trastorno, y con diseños y características metodológicas variables. La codificación de todas estas variables en los estudios tiene precisamente como objeto poder comprobar cuáles de ellas pueden estar relacionadas con los resultados de eficacia.

En esta fase es muy importante que se compruebe la fiabilidad del proceso de codificación de las características. Para ello, lo habitual es que dos, o más, investigadores codifiquen de forma independiente todos o algunos de los estudios empíricos y comprobar el grado de acuerdo entre ellos. Sólo de esa forma podremos saber si el MA ha aplicado unas normas objetivas y sistemáticas en el proceso de codificación.

(4) *Análisis estadístico e interpretación*. Además de las variables moderadoras de los estudios, la realización de un MA requiere del cálculo de un índice cuantitativo que permita poner en la misma métrica los resultados de los estudios. Esto se debe a que los estudios medirán los efectos de los tratamientos con diferentes tests y escalas

psicológicas, de forma que sus resultados no son directamente comparables al estar en unidades de medida diferentes. Esta homogeneización de los resultados se logra mediante la aplicación de algún índice del tamaño del efecto. El tamaño del efecto es, pues, un índice que refleja el grado en que difieren, en promedio, los resultados de los participantes en el grupo tratado respecto de los del grupo de control. Aunque son muy variados los índices del tamaño del efecto que podemos encontrar en los MAs, el más utilizado es la diferencia de medias tipificada, definido como la diferencia entre las medias de los dos grupos dividida por la desviación típica conjunta de ambos. Al dividir por la desviación típica logramos obtener un índice cuantitativo homogéneo y comparable independientemente de los tests o escalas que se hayan utilizado en los diferentes estudios, ya que se pueden interpretar como unidades típicas de separación entre las medias de los dos grupos. Aparte de la diferencia de medias tipificada, también es frecuente encontrar en los MAs índices del tamaño del efecto para variables de resultado dicotómicas, como son la diferencia de proporciones, el riesgo relativo o el odds ratio, pudiéndose transformar unos índices en otros (Sánchez Meca, Marín Martínez y Chacón Moscoso, 2003).

Una vez que tenemos registrados para cada estudio sus características (variables moderadoras) y su tamaño del efecto, la base de datos resultante puede someterse a análisis estadísticos que permitan responder a las preguntas clave a que se enfrenta un MA: (a) ¿Cuál es la magnitud del efecto media de todos los estudios?; (b) ¿son homogéneos los tamaños del efecto de los estudios?, (c) caso de no ser homogéneos, ¿qué características de los estudios pueden dar cuenta de esa heterogeneidad? y (d) ¿es posible formular un modelo explicativo de la heterogeneidad de los tamaños del efecto a partir de un subconjunto de las variables moderadoras codificadas?

Para dar respuesta a estas preguntas se aplican técnicas de análisis estadístico en las que el peso que ejerce cada estudio en los cómputos meta-analíticos está en función de la precisión exhibida por su tamaño del efecto, y la precisión está en función del tamaño de la muestra: a mayor tamaño muestral, mayor precisión y, en consecuencia, mayor peso en los análisis. De esta forma, se calcula una media ponderada de los tamaños del efecto junto con su intervalo de confianza, se evalúa el grado de heterogeneidad de los tamaños del efecto y, si los tamaños del efecto no son homogéneos, se analiza el influjo de variables moderadoras sobre los tamaños del efecto. Esta última fase de los análisis se lleva a cabo aplicando procedimientos ponderados basados en el análisis de la varianza (o análisis por subgrupos) y en los modelos de regresión (meta-regresión), de forma que la variable dependiente está formada por los tamaños del efecto obtenidos en los estudios, mientras que las variables independientes o predictoras son las características de los estudios.

(5) *Publicación*. La última fase de un MA, como cualquier otra investigación, consiste en diseminar sus resultados. La publicación de un MA se rige por las mismas normas que las de cualquier estudio empírico (Botella y Gambara, 2006b). Las secciones de un estudio meta-analítico suelen ser, pues, la introducción, el método, los resultados y la discusión y conclusiones.

En la introducción se revisa el tema objeto de estudio, se definen los constructos psicológicos implicados y se formulan los objetivos del MA. En la sección de 'Método' se incluyen varias subsecciones. En primer lugar, la subsección 'búsqueda de los estudios' tiene por objeto especificar los criterios de selección de los estudios y los procedimientos de búsqueda de la literatura utilizados. En segundo lugar, se presenta una subsección dirigida a especificar el proceso de codificación de los estudios, donde se explican las características de los estudios codificadas. La tercera y última subsección, que suele titularse 'análisis estadístico', sirve para definir el índice del tamaño del efecto utilizado en el MA, así como las técnicas estadísticas de integración aplicadas. La sección 'Método' tiene por objeto permitir la replicabilidad del MA por otros investigadores, por lo que debe hacer lo más explícitas posible todas las decisiones adoptadas durante la realización del MA.

En la sección de 'Resultados' se representan los resultados de los análisis estadísticos aplicados en el MA, tratando de dar respuesta a las cuatro preguntas básicas de un MA antes reseñadas. Y en la sección de 'Discusión' los resultados del MA se ponen en relación con la literatura previa sobre el tema, se discute su relevancia práctica, sus implicaciones para la práctica profesional y se apuntan líneas futuras de investigación.

UN EJEMPLO

Siguiendo con el ejemplo antes planteado sobre la eficacia de los tratamientos psicológicos del trastorno de pánico con o sin agorafobia (Sánchez Meca et al., en prensa), en dicho MA se logró seleccionar 65 estudios que cumplían con los criterios de selección y en cada uno de ellos se

obtuvo una diferencia de medias estandarizada (*d*) que comparaba los resultados medios alcanzados por los grupos tratado y de control en el postest.[3] Los resultados del MA están basados en una muestra total de más de 2.300 pacientes con dicho trastorno psicológico, lo que da una idea del grado de generalización que pueden ofrecernos los resultados del MA.

A la pregunta ¿cuál es el grado de eficacia global obtenido con todo el conjunto de estudios?, este MA ofreció un tamaño del efecto medio $d_+ = 1.015$, con un intervalo de confianza estadísticamente significativo, que tomó valores entre 0.855 y 1.175. El valor 1.015 puede interpretarse en unidades típicas y, siguiendo el criterio de Cohen (1988), puede considerarse que valores en torno a 0.2, 0.5 y 0.8 reflejan una significación práctica de magnitud baja, media y alta, respectivamente. Por tanto, el valor 1.015 implica una elevada eficacia de los tratamientos psicológicos, en general, sobre el trastorno de pánico. Otra interpretación práctica del efecto medio puede hacerse asumiendo que la reducción de los ataques de pánico en los grupos tratado y de control se distribuyen según una ley normal, y que el valor d = 1.015 representa en unidades típicas la separación entre los niveles medios de los dos grupos. Así, tomando como población de referencia el grupo de control, el efecto d = 1.015 indicaría que, en promedio, los pacientes que han recibido tratamiento psicológico se sitúan en el percentil 84.4% de la distribución de los controles o, lo que es lo mismo, que los tratamientos psicológicos han logrado reducir los ataques de pánico en un 34.4% respecto de los controles. La Figura 1 ilustra de forma gráfica esta interpretación práctica del tamaño del efecto medio.

La segunda pregunta, estrechamente relacionada con la anterior, a la que tiene que dar respuesta un MA es si los tamaños del efecto son homogéneos en torno a su media o si, por el contrario, muestran tanta heterogeneidad que la media no representa bien al conjunto de los estudios. Mediante pruebas estadísticas apropiadas, como son el estadístico *Q* y el índice I^2 (Borenstein et al., 2009), el MA permite responder a esta pregunta que, en el caso del MA que nos ocupa llevó a la conclusión de que los estudios estaban reflejando resultados de eficacia (cuantificados en sus tamaños del efecto) muy heterogéneos entre sí.

Como consecuencia de la heterogeneidad manifestada por los tamaños del efecto, se hace preciso responder a la tercera pregunta: ¿qué características de los estudios pueden estar afectando a la heterogeneidad? Es en esta fase donde se aplican técnicas de análisis de varianza y de regresión para encontrar las variables moderadoras de la eficacia. En un MA sobre tratamientos psicológicos, la variable moderadora más importante es el tipo de tratamiento aplicado en los estudios. Al clasificar a

FIGURA 1
REPRESENTACIÓN GRÁFICA DE LA INTERPRETACIÓN PRÁCTICA DEL TAMAÑO DEL EFECTO 'DIFERENCIA DE MEDIAS TIPIFICADA'

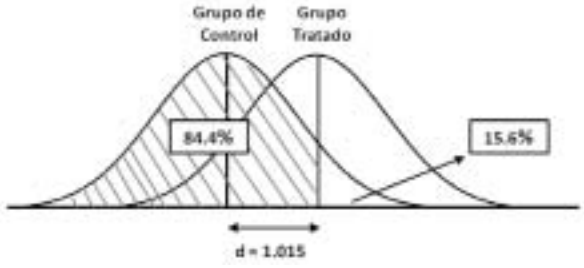
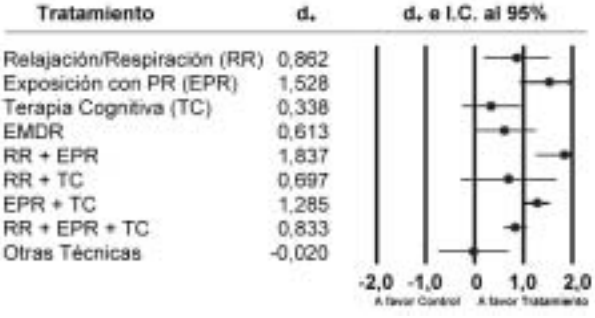


Figura 2
Representación gráfica tipo "forest plot" de los resultados de eficacia media alcanzados por diferentes tratamientos psicológicos del trastorno de pánico con o sin agorafobia. RR: Técnicas de relajación y/o respiración. EPR: Exposición con prevención de recaídas (básicamente, exposición *in vivo*). TC: terapia cognitiva (básicamente, técnicas de reestructuración cognitiva). EMDR ("Eye Movement Desensitization Reprocessing"): Técnicas de reprocesamiento por desensibilización de los movimientos oculares. d_+ es el tamaño del efecto medio obtenido por cada tratamiento. I.C.: Intervalo de confianza.



[3] Los resultados del MA que se presentan en este artículo son sólo una pequeña muestra de todas las evidencias que aportó. En este sentido, conviene aclarar que para cada estudio se calculó un tamaño del efecto para cada medida de resultado diferente (ataques de pánico, conductas agorafóbicas, nivel de ansiedad general, de depresión, de ajuste global, etc.). En este artículo hacemos referencia únicamente a los resultados obtenidos con las medidas de pánico, que son las más relevantes para este trastorno.

los estudios según la modalidad de tratamiento y calcular el tamaño del efecto medio alcanzado en cada una de ellas es posible comparar sus resultados de eficacia. Un modo muy útil de presentar la comparación entre tratamientos es mediante la construcción de un gráfico denominado 'forest plot', en el que se presenta de forma gráfica el efecto medio y el intervalo de confianza para cada categoría de tratamiento. En la Figura 2 se reproduce el 'forest plot' de este MA en el que se observan los tamaños del efecto medios obtenidos por las diferentes técnicas de tratamiento y las combinaciones entre ellas. Así, el gráfico permite observar cómo algunas técnicas de tratamiento han obtenido un efecto medio que no difiere significativamente del efecto nulo, al cruzar su intervalo de confianza al valor 0 (por ejemplo, la terapia cognitiva sola), y cómo otras técnicas sí han alcanzado un efecto medio estadísticamente significativo (por ejemplo, la exposición con prevención de recaídas, o ésta combinada con relajación/respiración).

GUÍA PARA LA LECTURA CRÍTICA DE META-ANÁLISIS

El examen de los resultados aportados por un MA sobre la eficacia de tratamientos, intervenciones o programas de prevención permite al lector valorar la eficacia diferencial de diferentes tratamientos y, en consecuencia, ayudarle en la toma de decisión sobre qué tratamiento aplicar en un caso particular. Sin embargo, la lectura crítica de los MAs pasa necesariamente por que el profesional tenga unos conocimientos apropiados de qué es un MA, cómo se hace y a qué sesgos pueden estar expuestos sus resultados. Conscientes de esta problemática, los expertos en MA han dedicado importantes esfuerzos a elaborar guías orientativas para la lectura crítica de estudios meta-analíticos, fruto de los cuales ha sido la publicación de diversas guías. En lugar de reproducir alguna de esas guías, aquí proponemos una que se basa fundamentalmente en las dos más recientemente propuestas en la literatura: la guía PRISMA ('Preferred Reporting Items for Systematic Reviews and Meta-Analyses'; Moher et al., 2009), que es una mejora de la guía QUOROM ('Quality Of Reporting Of Meta-analyses'; Moher et al., 1994) y la guía AMSTAR (Shea, Grimshaw, Wells et al., 2007; Shea, Hamel, Wells et al., 2009), que consta de 11 preguntas sobre el proceso de realización y publicación de una RS o MA.

El Protocolo que presentamos en la Tabla 2 pretende recoger las claves principales en que nos tenemos que fijar cuando estamos leyendo una RS o un MA, con objeto de poder valorar críticamente la calidad de los resultados que aporta y su relevancia para la práctica clínica. Las preguntas están orientadas a las diferentes secciones en que se divide la publicación de un MA: título, resumen, introducción, método, resultados y discusión. Básicamente, las preguntas están orientadas a comprobar si los meta-analistas han hecho explícitas todas las decisiones que han tenido que tomar durante la realización del MA, y ésta es una cuestión fundamental para poder valorar su calidad crítica y para garantizar que otros investigadores puedan replicar el MA.

REFLEXIONES FINALES

A pesar de la disociación existente entre la práctica profesional y la investigación, el enfoque de la PBE está posibilitando puntos de encuentro entre estos dos ámbitos que siempre deberían ir de la mano. Al mismo tiempo, las RSs y los MAs constituyen un modo rápido y seguro de conocer las últimas evidencias y pruebas científicas sobre cualquier tema relacionado con la práctica profesional. Es por ello que los psicólogos deben conocer esta metodología y saber hacer lectura crítica de RSs y MAs, así como de otros tipos de estudios que aportan evidencias.

En esta misma línea, sería conveniente que los planes de estudio de los Grados y Posgrados en Psicología contengan materias en las que se explique el enfoque de la PBE y la lectura crítica, no sólo de RSs y MAs, sino también de otros tipos de investigación, como los ensayos clínicos aleatorizados, los estudios de cohortes o los estudios observacionales y correlacionales. Sólo así lograremos que los profesionales vean la metodología como una disciplina cercana que tiene utilidad práctica en su desempeño profesional.

La lectura crítica de la investigación debería guiar las prácticas de los psicólogos, no sólo en el ejercicio directo con personas, sino en la toma de decisiones cuando se ocupan cargos directivos y de gestión en empresas y entidades que tengan que ver con la gestión de los servicios sociales, de salud y educativos. Y, en última instancia, debemos también tender a hacer Política Basada en la Evidencia, allá donde los psicólogos ocupen cargos públicos de responsabilidad o cuando otros nos pidan opiniones basadas en la evidencia acumulada por la psicología para adoptar sus decisiones.

Por último, para profundizar en la metodología del MA y de las RSs sugerimos algunas lecturas. En castellano pueden consultarse la monografía de Botella y Gambara (2002) o el capítulo de Sánchez Meca (2008). En inglés pueden consultarse los textos de Cooper (2010) y de Borenstein et al. (2009). Existen diversos programas de

TABLA 2
LISTA DE PREGUNTAS ORIENTADAS A LA LECTURA CRÍTICA DE RSS Y MAS

1. ¿Se identifica el estudio como un MA?		Sí
		No
		No disponible
2. ¿Incluye un **Resumen** con los objetivos, método, resultados y principales conclusiones? Debe proporcionarse un resumen estructurado que incluya: justificación; objetivos; fuente de los datos; criterios de selección de los estudios, participantes e intervenciones; valoración de la calidad de los estudios y métodos de síntesis; resultados; limitaciones del estudio; conclusiones e implicaciones de los principales resultados.		Sí
		No
		No disponible
3. ¿En la **Introducción** se describen de forma explícita las preguntas y los objetivos del MA? Debe proporcionarse una declaración explícita de las preguntas que se pretenden responder, con referencia a los participantes, las intervenciones, las comparaciones, las variables de resultado y el diseño de los estudios (PICOS: Participants, Interventions, Comparisons, Outcomes, and Study design).		Sí
		No
		No disponible
4. ¿En el **Método** se especifican los criterios de inclusión de los estudios? Deben especificarse las características de los estudios (e.g., PICOS, duración del período de seguimiento) y las características de los estudios utilizadas como criterios de elegibilidad, aportando su fundamentación (e.g., años considerados, idiomas, estatus de publicación).		Sí
		No
		No disponible
5. ¿En el **Método** se indican los procedimientos de búsqueda de los estudios? Deben describirse todas las fuentes de información (e.g., bases de datos con sus fechas de cobertura, contactos con autores de los estudios para identificar estudios adicionales) utilizadas en la búsqueda y fecha última de búsqueda. Debe presentarse la estrategia de búsqueda electrónica completa de al menos una base de datos, incluyendo los posibles límites impuestos, de forma que cualquiera pueda repetirla.		Sí
		No
		No disponible
6. ¿En el **Método** se especifican las variables de los estudios codificadas? Debe describirse el método de extracción de datos de los estudios primarios (e.g., protocolos de registro aplicados de forma independiente por dos o más codificadores), así como cualesquier procesos de obtención y confirmación de datos utilizados por los revisores. Debe incluirse una lista con todas las variables registradas en los estudios, así como su definición (e.g., PICOS, fuentes de financiación), así como cualesquier supuestos y simplificaciones adoptados en dicho proceso.		Sí
		No
		No disponible
7. ¿En el **Método** se hace alusión a la fiabilidad de la codificación? Un buen MA debe haber realizado un análisis de la fiabilidad de la codificación de las variables moderadoras de los estudios, y presentar los resultados de dicho análisis en términos de índices kappa y correlaciones intraclase.		Sí
		No
		No disponible
8. ¿En el **Método** se especifica/n el/los índice/s del tamaño del efecto? Debe especificarse cuál o cuáles fueron los índices del tamaño del efecto utilizados en el MA (e.g., diferencia de medias estandarizada, odds ratio, etc.).		Sí
		No
		No disponible
9. ¿En el **Método** se describen los métodos estadísticos utilizados en el MA? Deben describirse los métodos de tratamiento de los datos y cómo se combinaron los resultados de los estudios (e.g., modelo de efectos fijos, de efectos aleatorios o de efectos mixtos). Deben incluirse las medidas de consistencia utilizadas para analizar la heterogeneidad de los efectos (e.g., Q e I^2). Debe especificarse alguna valoración del riesgo de sesgo que pudiera afectar a la evidencia acumulativa (e.g., sesgo de publicación, reporte selectivo dentro de los estudios). Deben describirse los métodos de análisis adicionales (e.g., análisis de sensibilidad, análisis por subgrupos, meta-regresión).		Sí
		No
		No disponible
10. ¿En los **Resultados** se presentan las características de los estudios? Deben describirse las características de los estudios integrados, así como también debe proporcionarse una tabla con dichas características individuales, o bien ofrecer al lector la posibilidad de disponer de dicha tabla.		Sí
		No
		No disponible
11. ¿En los **Resultados** se analizan los estudios según su calidad? La calidad metodológica de los estudios debe haberse codificado y puesto en relación con los tamaños del efecto, con objeto de comprobar posibles sesgos debidos a una calidad pobre. Si se han incluido estudios aleatorizados y no aleatorizados deben compararse sus resultados.		Sí
		No
		No disponible

software diseñados para realizar los análisis estadísticos típicos de un MA. David B. Wilson ha desarrollado unos macros de MA para su uso en los paquetes estadísticos SPSS, SAS y STATA, que pueden obtenerse gratuitamente en el sitio web: http://mason.gmu.edu/~dwilsonb/ma.html. Además, la Colaboración Cochrane ha elaborado el programa *RevMan 5.0*, para hacer MAs, que puede obtenerse también gratuitamente en el sitio web de esta organización (www.cochrane.org). Y también cabe mencionar el programa comercial *Comprehensive Meta-analysis 2.0* elaborado por Borenstein, Hedges, Higgins y Rothstein (2005; www.meta-analysis.com).

REFERENCIAS

Borenstein, M.J., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2005). *Comprehensive Meta-analysis* (Vers.2). Englewood Cliffs, NJ: Biostat, Inc.

Borenstein, M.J., Hedges, L.V., Higgins, J.P.T. y Rothstein, H.R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Botella, J. y Gambara, H. (2002). *Qué es el meta-análisis*. Madrid: Biblioteca Nueva.

Botella, J. y Gambara, H. (2006a). El meta-análisis: una metodología de nuestro tiempo. *Infocop*, 29 mayo.

Botella, J. y Gambara, H. (2006b). Doing and reporting a meta-analysis. *International Journal of Clinical and*

TABLA 2
LISTA DE PREGUNTAS ORIENTADAS A LA LECTURA CRÍTICA DE RSS Y MAS (Continuación)

12. ¿En los **Resultados** se presentan los efectos medios y las medidas de consistencia? Deben presentarse los resultados de cada MA realizado, incluyendo los tamaños del efecto medios con sus intervalos confidenciales y las medidas de consistencia o heterogeneidad (e.g., *Q*, I^2). Opcionalmente, pueden presentarse los resultados de los estudios individuales y de cada MA mediante un 'forest plot'.		Sí
		No
		No disponible
13. Si ha habido heterogeneidad, ¿en los **Resultados** se presentan los análisis de moderadores? Caso de que exista heterogeneidad entre los tamaños del efecto, deben aplicarse modelos de efectos mixtos, tales como análisis por subgrupos (ANOVAs) y meta-regresión (análisis de regresión) para identificar características moderadoras de los resultados.		Sí
		No
		No disponible
14. ¿En los **Resultados** se hace algún análisis de sensibilidad? Si se diseñó realizar análisis de sensibilidad para comprobar la consistencia y robustez de los resultados del MA, deben describirse en esta sección.		Sí
		No
		No disponible
15. ¿En los **Resultados** se hace un análisis del sesgo de publicación? El MA debe haber realizado algún análisis de sesgo de publicación para comprobar si éste puede ser una amenaza contra la validez de sus resultados.		Sí
		No
		No disponible
16. ¿En la **Discusión** se resumen las evidencias? Deben resumirse los principales resultados, incluyéndose la fuerza de las evidencias logradas con cada variable de resultado principal; debe también considerarse su relevancia para los diferentes grupos implicados (e.g., profesionales de cuidados de salud, usuarios y políticos).		Sí
		No
		No disponible
17. ¿En la **Discusión** se plantean las limitaciones del MA? Deben discutirse las limitaciones tanto en el nivel de los estudios como en el de las variables de resultado (e.g., riesgos de sesgo) y en el nivel de la revisión (e.g., recuperación incompleta de investigaciones, sesgo de reporte).		Sí
		No
		No disponible
18. ¿En la **Discusión** se plantean las implicaciones para la práctica profesional? Deben discutirse las implicaciones que los principales resultados del MA tienen para el ejercicio profesional de los clínicos, los gestores y en la toma de decisiones política.		Sí
		No
		No disponible
19. ¿En la **Discusión** se plantean las implicaciones para la investigación futura? Debe aportarse una interpretación general de los resultados en el contexto de otras pruebas y evidencias, así como las implicaciones para la investigación futura.		Sí
		No
		No disponible
20. ¿Se especifica/n la/s fuente/s de **financiación**? Deben describirse las fuentes de financiación de la RS o del MA, así como otras ayudas recibidas (e.g., facilitación de datos) y el papel jugado por los financiadotes en la revisión sistemática, con objeto de valorar la posible existencia de conflicto de intereses		Sí
		No
		No disponible

Health Psychology, *6*, 425-440.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2ª ed.). Hillsdale, NJ: Erlbaum.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (3ª ed.). Thousand Oaks, CA: Sage.

Cooper, H., Hedges, L.V. y Valentine, J.C. (Eds.)(2009). *The handbook of research synthesis and meta-analysis* (2ª ed.). Nueva York: Russell Sage Foundation.

Frías Navarro, M.D. y Pascual Llobell, J. (2003). Psicología clínica basada e pruebas: Efecto del tratamiento. *Papeles del Psicólogo, 85*.

Grupo de Atención Sanitaria Basada en la Evidencia (2007). *Atención sanitaria basada en la evidencia: Su aplicación a la práctica clínica*. Murcia: Consejería de Sanidad de la Región de Murcia.

Littell, J.H., Corcoran, J. y Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford, UK: Oxford University Press.

Marín Martínez, F., Sánchez Meca, J., Huedo, T. y Fernández, I. (2007). Meta-análisis: Dónde estamos y hacia dónde vamos. En A. Borges y P. Prieto (Eds.), *Psicología y ciencias afines en los albores del siglo XXI* (pp. 87-102). La Laguna, Tenerife: Grupo Editorial Universitario.

Marín Martínez, F., Sánchez Meca, J. y López López, J.A. (2009). El meta-análisis en el ámbito de las Ciencias de la Salud: Una metodología imprescindible para la eficiente acumulación del conocimiento. *Fisioterapia, 31*, 107-114.

Martín, J.L.R., Tobías, A. y Seoane, T. (Coords.) (2006). *Revisiones sistemáticas en ciencias de la vida*. Toledo: FISCAM.

Moher, D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., et al. (1994) Improving the quality of reporting of meta-analysis of randomized controlled trials: The QUOROM statement. *Lancet, 354*, 1896–1900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., The PRISMA Group (2009) Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine, 6*(7): e1000097. doi:10.1371/journal.pmed.1000097.

Navarro, F., Giribet, C. y Aguinaga, E. (1999). Psiquiatría basada en la evidencia: Ventajas y limitaciones. *Psiquiatría Biológica, 6*, 77-85.

Nezu, A. M. & Nezu, C. M. (2008). *Evidence-Based Outcome Research: A Practical Guide to Conducting Randomized Controlled Trials*. Oxford University Press.

Pascual Llobell, J., Fías Navarro, M.D. y Monterde, H. (2004). Tratamientos psicológicos con apoyo empírico y práctica clínica basada en la evidencia. *Papeles del Psicólogo, 87*.

Petticrew, M. y Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA: Blackwell.

Sánchez Meca, J. (1999). Meta-análisis para la investigación científica. En F.J. Sarabia-Sánchez (Coord.), *Metodología para la investigación en márketing y dirección de empresas* (pp. 173-201). Madrid. Pirámide.

Sánchez Meca (2003). La revisión del estado de la cuestión: el meta-análisis. En C. Camisón, M.J. Oltra y M.L. Flor (Eds.), *Enfoques, problemas y métodos de investigación en Economía y Dirección de Empresas. Tomo I* (pp. 101-110). Castellón: Universitat Jaume I.

Sánchez Meca, J. (2008). Meta-análisis de la investigación. En M.A. Verdugo, M. Crespo, M. Badía y B. Arias (Coords.), *Metodología en la investigación sobre discapacidad*. Salamanca: Publicaciones del INICO (Colección ACTAS, 5/2008).

Sánchez Meca, J. y Ato, M. (1989). Meta-análisis: una alternativa metodológica a las revisiones tradicionales de la investigación. En J. Arnau y H. Carpintero (Eds.), *Tratado de psicología general. I: Historia, teoría y método* (pp. 617-669). Madrid:Alambra.

Sánchez Meca, J., Boruch, R.F., Petrosino, A. y Rosa Alcázar, A.I. (2002). La Colaboración Campbell y la práctica basada en la evidencia. *Papeles del Psicólogo*, *83,* 44-48.

Sánchez Meca, J. y Marín Martínez, F. (en prensa). Meta-analysis. En B. McGaw, E. Baker y P.P. Peterson (Eds.), *International encyclopedia of education* (3ª ed.). Oxford: Elsevier.

Sánchez Meca, J., Marín Martínez, F. y Chacón Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*, 448-467.

Sánchez Meca, J., Rosa Alcázar, A.I., Marín Martínez, F. y Gómez Conesa, A. (en prensa). Psychological treatment of panic disorder with and without agoraphobia: A meta-analysis. *Clinical Psychology Review*.

Shea, B.J., Grimshaw, J.M., Wells, G.A., et al. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BioMed Central*, *7*(10). doi: 10.1186/1471-2288-7-10.

Shea, B.J., Hamel, C., Wells, G.A., et al. (2009). AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *Journal of Clinical Epidemiology, 62*, 1013-1020.

Vázquez, C. y Nieto, M. (2003). Psicología (clínica) basada en la evidencia (PBE): Una revisión conceptual y metodológica. En J.L. Romero (Ed.), *Psicópolis: Paradigmas actuales y alternativos en la psicología contemporánea*. Barcelona: Kairós.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 18-33
http://www.cop.es/papeles

EL ANÁLISIS FACTORIAL COMO TÉCNICA DE INVESTIGACIÓN EN PSICOLOGÍA

FACTOR ANALYSIS AS A RESEARCH TECHNIQUE IN PSYCHOLOGY

Pere Joan Ferrando y Cristina Anguiano-Carrasco
Centro de investigación para la evaluación y medida de la conducta. Universidad 'Rovira i Virgili'

El presente texto explica los principales aspectos del análisis factorial como instrumento de investigación psicológica. Se revisan en primer lugar los aspectos básicos a nivel conceptual, de modo que su lectura sea adecuada tanto para el lector principiante como para aquellos que quieran profundizar más en sus conocimientos de la técnica. Después, se discuten con cierto detalle las diferencias entre el análisis exploratorio y confirmatorio, y los procedimientos para estimar el modelo y obtener la solución transformada. Estos puntos se discuten siguiendo cada una de las etapas recomendadas en una investigación: desde el diseño y recogida de datos hasta la interpretación de la solución final. Finalmente, se explica el funcionamiento del programa Factor, un programa más completo que la mayor parte de los programas comerciales, y que además es de distribución libre. Adicionalmente, se propone al lector la realización de un ejercicio resuelto, a modo de práctica para la aplicación de lo expuesto en el texto.
Palabras clave: *Análisis factorial exploratorio, Análisis factorial confirmatorio, Componentes principales, Programa de ordenador FACTOR*

The present text explains the main aspects of factor analysis as a tool in psychological research. First, the basic issues are revised at a conceptual level, so that the review is appropriate for beginners as well as for those who want an in deep knowledge of the technique. Next, the differences between exploratory and confirmatory analysis as well as the procedures for fitting the model and transforming the initial solution are discussed in detail. These issues are discussed according to the recommended steps in a factor-analytic research: from the design and data collection to the interpretation of the final solution. Finally, the functioning of the "Factor" program is explained. Factor is more complete than most commercial programs, and is also freely distributed. Additionally, a solved exercise is proposed as a practice to apply the material discussed in the text.
Key words: *Exploratory factor analysis, Confirmatory factor analysis, Principal components, FACTOR computer program*

Nacido con el siglo XX, el análisis factorial (AF) se ha desarrollado considerablemente a lo largo de sus más de 100 años de existencia. El sencillo modelo inicial propuesto por Spearman (1904) para validar su teoría de la inteligencia ha dado lugar a una amplia familia de modelos que se utilizan no sólo en ciencias sociales, sino también en otros dominios como Biología o Economía. Dado que un tratamiento completo del AF excedería con mucho las posibilidades de este artículo, conviene delimitar primero qué temas se van a tratar.

Desde hace años el primer autor revisa trabajos empíricos en los que se emplea el AF en la investigación psicológica, y la experiencia adquirida servirá para establecer las primeras delimitaciones. En primer lugar, la mayor parte de los estudios factoriales en psicología utilizan el AF para evaluar (a) la estructura de un test a partir de las puntuaciones en sus ítems, o (b) hipótesis de tipo dimensional utilizando como medidas puntuaciones en diferentes tests. Parece razonable, por tanto, centrar la exposición en este tipo de medidas: puntuaciones en ítems o tests.

En segundo lugar, la experiencia indica que los problemas metodológicos en estos estudios son casi siempre los mismos. Un primer grupo de problemas surge en la etapa del diseño de la investigación (etapa generalmente descuidada en los estudios factoriales). Los problemas del segundo grupo se refieren a las decisiones que debe tomar el investigador en la etapa de estimación y ajuste del modelo y en la de rotación. En particular, la mayor parte de los problemas se deben al empleo injustificado del "pack" conocido como "Little Jiffy": *Componentes principales - valores propios mayores que uno - rotación Varimax*. Dedicaremos especial atención al diseño y a la estimación y ajuste del modelo.

Aún con estas delimitaciones, el tema sigue siendo demasiado amplio. En este artículo nos centraremos tan sólo en el modelo general más básico de AF: el modelo lineal, basado en correlaciones, y que analiza medidas obtenidas en un solo grupo de sujetos y en una sola ocasión. Esta limitación deja fuera temas de gran interés: los modelos extendidos de medias y covarianzas, los mode-

Correspondencia: Pere Joan Ferrando. *Universidad 'Rovira i Virgili'. Facultad de Psicologia. Carretera Valls s/n. 43007 Tarragona. España. E-mail: perejoan.ferrando@.urv.cat*

los no-lineales y sus relaciones con la teoría de respuesta al ítem, y los modelos para grupos múltiples y múltiples ocasiones. Tampoco podremos entrar en el tema general de las puntuaciones factoriales.

El enfoque del artículo es conceptual y aplicado, y se ha tratado de reducir el formalismo al máximo. Sólo se incluyen las ecuaciones básicas del modelo y se presentan en cuadros de texto aparte. Asimismo, se ha tratado de reducir al mínimo el número de referencias, y, en cambio, se ha propuesto un apartado de lecturas recomendadas. En este sentido, debemos advertir que algunos de los tópicos discutidos son controvertidos, y que las recomendaciones reflejan la posición de los autores. El apartado de lecturas puede servir para que el lector vea otras posiciones y evalúe críticamente lo que le decimos aquí.

LAS IDEAS BÁSICAS DEL ANÁLISIS FACTORIAL

El AF es un modelo estadístico que representa las relaciones entre un conjunto de variables. Plantea que estas relaciones pueden explicarse a partir de una serie de variables no observables (latentes) denominadas factores, siendo el número de factores substancialmente menor que el de variables. El modelo se obtiene directamente como extensión de algunas de las ideas básicas de los modelos de regresión lineal y de correlación parcial. Del primer modelo se derivan las ecuaciones fundamentales del AF. Del segundo se derivan las ideas clave para evaluar el ajuste del modelo a los datos.

En el modelo de regresión lineal, la puntuación en una variable criterio puede aproximarse, o explicarse en parte, mediante una combinación lineal (una suma de variables multiplicadas cada una de ellas por un peso o coeficiente) de una serie de variables predictoras o explicativas denominadas regresores. Se asume explícitamente que la combinación es tan sólo una aproximación, y que una parte de la puntuación del criterio no podrá ser predicha o explicada desde los regresores. Esta parte no explicada es el término de error (véase ec. 1 en el cuadro).

$$Y = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m + e \qquad (1)$$

$$X_j = \mu_j + \lambda_{j1}f_1 + \lambda_{j2}f_2 + \ldots + \lambda_{jm}f_m + e_j \qquad (2)$$

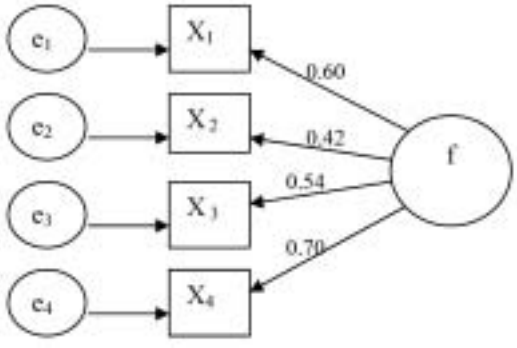
$$X_j = \mu_j + \lambda_j f + e_j \qquad (3)$$

En el AF se analiza un conjunto de variables observables (ítems, subtests o tests) cada una de las cuales puede considerarse como un criterio. Así entendido, el AF consiste en un sistema de ecuaciones de regresión como la descrita arriba (una ecuación para cada variable observable) en el que los regresores, denominados aquí factores, son comunes para un subconjunto (factores comunes) o todo el conjunto (factores generales) de variables (véase ec. 2 en el cuadro). Para cada una de estas ecuaciones la diferencia básica entre el AF y una regresión convencional es que los regresores, es decir los factores, no son observables. Esta diferencia es la que hace que el AF sea un modelo más complejo que el de regresión. Para empezar, al ser los factores no observables, carecen de una escala de medida o métrica determinada. Para resolver esta indeterminación, la práctica más simple, que seguiremos aquí, consiste en asumir que los factores están en escala típica: media cero y varianza uno. Si, además, las variables observables también están en escala típica, el modelo es más simple matemáticamente y más fácilmente interpretable.

Por analogía con el modelo de regresión, se sigue que el modelo AF más sencillo es aquel que plantea un solo factor general (ec. 3). Este modelo sería equivalente al de regresión simple, y fue el modelo AF inicial que planteó Spearman. Para empezar a fijar ideas vamos a estudiar una solución de Spearman basada en el AF de un conjunto de 4 ítems, junto con su representación en un diagrama de Wright. Para la elaboración de estos diagramas, el lector puede consultar el artículo de Ruíz, Pardo y San Martín en el presente volumen.

Ítem	f_i
X_1	0.60
X_2	0.42
X_3	0.54
X_4	0.70

Representado gráficamente sería:



La ecuación AF para el ítem 1 es:

$$X_{i1} = 0.6\, f_i + e_{i1}$$

Que se interpreta como sigue. La puntuación del individuo *i* en el ítem 1 viene en parte determinada por el efecto del factor común (el nivel de *i* en el factor común *f*) y en parte es error. En AF el término de error incluye todos aquellos efectos distintos al factor o factores comunes que influyen en la puntuación en la variable observada. De forma más sistemática, podemos distinguir tres grandes grupos de efectos o tipos de error: (a) el error de muestreo (error estadístico), (b) el error de medida (error psicométrico) y (c) el error de aproximación. Este último componente de error significa que el modelo especificado no se considera exactamente correcto ni siquiera en la población. En efecto, los modelos no son, ni pretenden serlo, exactos. Son, en el mejor de los casos, aproximaciones razonables a la realidad.

El valor 0.6 es el peso factorial y equivale a la pendiente en el modelo de regresión. Si las puntuaciones del ítem y del factor están en escala típica, este peso refleja la importancia que tiene el factor en la determinación de la puntuación en este ítem. Cuanto mayor el peso, mayor la importancia del factor, y, por tanto, menor la influencia del error. Además, en la escala típica asumida, este peso puede interpretarse como la correlación entre el factor y el ítem. Su cuadrado, que es el coeficiente de determinación, se interpreta como la proporción de varianza en las puntuaciones de este ítem que puede explicarse desde el factor. Así pues, de acuerdo con nuestra solución estimada, el ítem 1 correlacionaría 0.6 con el factor general; dicho factor explicaría el 36% de la varianza de las puntuaciones en este ítem (0.6^2=0.36) y, por tanto, el 64% restante sería varianza de error. En la terminología AF la proporción de varianza explicada recibe el nombre de "comunalidad".

Antes de seguir adelante, podría ser de interés como ejercicio que el lector interpretara la ecuación AF correspondiente a otro ítem, digamos el ítem 2. Asimismo, es muy importante tener en cuenta que las interpretaciones anteriores sólo son válidas si variables y factor están en escala típica. De no ser así, el peso no tiene una interpretación clara, ya que refleja entonces en mayor o menor grado las diferencias en la escala de medida de las variables. Al interpretar el output de un AF, es pues esencial asegurarse de que la solución que se interpreta es una solución tipificada.

Cuando se plantea más de un factor, tenemos el modelo AF múltiple, conocido también como modelo de Thurstone (1947), su principal impulsor. El modelo es el mismo para cualquier número de factores y, por simplicidad, lo explicaremos con dos. El aspecto clave aquí es la relación que se plantea entre los factores. Al igual que en regresión, el caso más simple y fácilmente interpretable es aquel en que los factores están incorrelados (conceptualmente, que son independientes entre sí). Una solución de este tipo se denomina "solución ortogonal". En una solución ortogonal los pesos factoriales siguen siendo interpretables como correlaciones variable-factor, sus cuadrados son proporciones de varianza explicada por el correspondiente factor y la suma de estos cuadrados es la comunalidad (véanse ecs. 4 y 5) o la proporción de varianza que explican conjuntamente los factores

.

$$\sigma_j^2 = \lambda_{j1}^2 + \lambda_{j2}^2 + \sigma_{ej}^2 = 1 \qquad (4)$$

$$1 = h_j^2 + \sigma_{ej}^2 \qquad (5)$$

$$r_{jk} = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} \qquad (6)$$

El caso de factores correlacionados, denominado 'solución oblicua' es el más complejo, pero posiblemente también el más realista en la práctica. El aspecto más importante a la hora de interpretar una solución de este tipo es que ahora los pesos y las correlaciones variable-factor son coeficientes distintos. Al igual que en teoría de la regresión lineal, los pesos factoriales son ahora coeficientes de regresión estandarizados y miden el efecto del factor sobre la variable de respuesta cuando los demás factores permanecen constantes. Estos pesos se presentan en la matriz denominada "patrón factorial". Por otra parte, las correlaciones variable-factor se denominan coeficientes estructurales, y se presentan en la matriz denominada "estructura factorial". Las ecuaciones correspondientes a los pesos y coeficientes estructurales se presentan en el cuadro que sigue abajo. Conceptualmente, los pesos indican hasta qué punto influye el factor en la variable, en tanto que los coeficientes estructurales indican hasta qué punto se parecen el factor y la variable. En el caso de soluciones oblicuas, en este capítulo nos centraremos sobre todo en los pesos, es decir la matriz patrón.

$$X_{ij} = \lambda_{j1} f_{i1} + \lambda_{j2} f_{i2} + \varepsilon_{ij} \qquad (7)$$

$$s_{j1} = \lambda_{j1} + \lambda_{j2}\phi_{12} \qquad (8)$$

$$\sigma_j^2 = \lambda_{j1}^2 + \lambda_{j2}^2 + 2\lambda_{j1}\lambda_{j2}\phi_{12} + \sigma_{ej}^2 \qquad (9)$$

$$r_{jk} = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + \phi_{12}\left(\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1}\right) \qquad (10)$$

En la ecuación (8) s_{j1} es el coeficiente estructural (correlación variable-factor) y ø es la correlación entre factores.

Veamos ahora un ejemplo de solución ortogonal múltiple con 6 ítems y dos factores:

Item	F_1	F_2
X_1	0.35	**0.82**
X_2	0.31	**0.85**
X_3	0.37	**0.80**
X_4	**0.79**	0.36
X_5	**0.77**	0.32
X_6	**0.79**	0.39

La representación gráfica es:

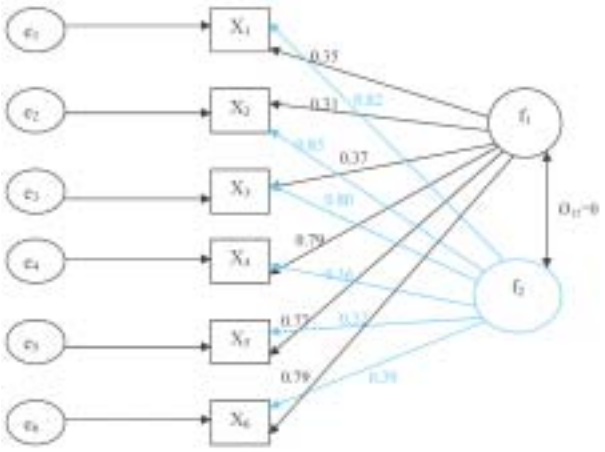


Diagrama de Wright para dos factores comunes incorrelados.

Si tomamos como ejemplo el ítem 2, obtenemos la siguiente ecuación:

$$X_{i2} = 0.31 f_1 + 0.85\, f_{i2} + e_{i2}$$

Esta ecuación nos indica que la puntuación del sujeto *i* en el ítem 2, además de por el error, viene determinada principalmente por el segundo factor (0.85), y, en menor medida, por el primero (0.31). En este caso para explicar la comunalidad (o varianza explicada) se sumarían los cuadrados de los pesos en ambos factores de modo que para el ítem 2,

$$h^2_2 = 0.31^2 + 0.85^2 = 0.82$$

obteniendo la parte de la varianza explicada por ambos factores. Restándola de uno obtenemos que la varianza de error es 0.18. En términos de proporciones, entre los dos factores explican un 82% de la varianza total (comunalidad) y el 18% restante sería error. Para el lector que lo desee sería interesante que realizara la interpretación de otro ítem, por ejemplo el 5.

A continuación mostramos una solución oblicua (patrón) obtenida con los mismos datos. La correlación estimada entre factores fue de 0.40. Con respecto a la solución ortogonal se aprecia que es más clara y simple (los pesos menores son ahora más cercanos a cero). Esto es lo que se observa generalmente al comparar ambos tipos de soluciones.

Ítem	F_1	f_2
X_1	0.23	**0.78**
X_2	0.18	**0.82**
X_3	0.25	**0.75**
X_4	**0.78**	0.18
X_5	**0.77**	0.14
X_6	**0.78**	0.21

El diagrama correspondiente sería ahora:

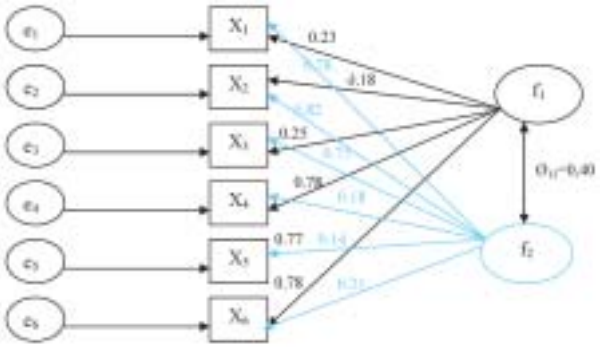


Diagrama de Wright para dos factores comunes correlados.

En este ejemplo vamos a trabajar con el ítem 4. Su ecuación básica tiene la misma forma que en el caso ortogonal:

$$X_{i4} = 0.78\, f_{i1} + 0.18\, f_{i2} + e_{i4}$$

Pero su interpretación es distinta, ya que los pesos (efecto del factor sobre la variable) y las correlaciones variable-factor son medidas distintas. Así el peso correspondiente al primer factor es 0.78. Sin embargo, la correlación entre la variable y este factor, es decir, el coeficiente estructural debe obtenerse como:

$$s_{jl} = 0.78 + 0.18 \cdot 0.40 = 0.85$$

El lector que desee practicar puede realizar los cálculos correspondientes al ítem 1.

Pasamos ahora a las aportaciones del modelo de correlación parcial. La ecuación básica se presenta en el cuadro siguiente para el modelo de Spearman y se basa en los resultados previos discutidos arriba.

$$r_{jk.i} = \frac{r_{jk} - r_{ij} r_{ik}}{\sqrt{1 - r_{ij}^2}\sqrt{1 - r_{ik}^2}} = \frac{\lambda_j \lambda_k - \lambda_i \lambda_j \lambda_i \lambda_k}{\sqrt{1 - \lambda_i^2 \lambda_j^2}\sqrt{1 - \lambda_i^2 \lambda_k^2}} = 0 \quad (11)$$

La ecuación (11) indica que, si el modelo es correcto, la correlación parcial entre cualquier par de variables después de eliminar de ambas la influencia del factor general es cero. El numerador de la correlación parcial es la diferencia entre la correlación observada entre las dos variables y la correlación reproducida desde el modelo y recibe el nombre de correlación residual. Así, si el modelo es correcto, la correlación observada y la reproducida son iguales y el residual cero. Conceptualmente este resultado se interpreta como que lo único que tienen en común las variables es el factor general que miden, por lo que, al eliminar esta causa común ya no hay nada que las relacione. El caso múltiple es más complejo pero la idea esencial es la misma. Si el modelo es correcto, y las variables tienen solo *m* factores en común, entonces la correlación parcial entre cualquier par de variables tras eliminar la influencia de estos factores comunes debe ser cero. Más que un resultado, esto es un principio de importancia básica. Sugiere que la vía más directa para evaluar si el modelo AF es apropiado deberá basarse en la evaluación de los residuales tras estimar el número propuesto de factores.

Globalmente, la mayor parte de las características del modelo expuesto hasta ahora vienen dictadas por el principio de parsimonia (Carroll, 1978). La parsimonia dicta que las ecuaciones del modelo sean lineales y por tanto lo más simples posible. Este principio recomienda también establecer una distinción clara entre varianza común (comunalidad) y varianza de error. Finalmente, el principio de parsimonia sugiere que el número de factores comunes debe ser considerablemente menor que el de variables. No se ganaría nada, en efecto, interpretando una solución con tantos factores como variables. La determinación del número de factores correcto (que sea lo bastante reducido como para ser claramente interpretable y lo bastante completo como para dar cuenta de las relaciones entre variables) es, posiblemente, la decisión mas importante del AF (Thurstone, 1947).

ANÁLISIS FACTORIAL EXPLORATORIO Y ANÁLISIS FACTORIAL CONFIRMATORIO

En la literatura (e.g. Mulaik, 1972) se distinguen de forma muy marcada dos tipos de análisis factorial: el análisis factorial exploratorio (AFE) y el análisis factorial confirmatorio (AFC). En nuestra opinión esta distinción no es tan clara como se presenta en los textos y, además, plantea una serie de problemas. En primer lugar, en la distinción AFE-AFC se mezclan dos conceptos: (a) la finalidad con la que se lleva a cabo el análisis y (b) el modelo que se pone a prueba. En segundo lugar, tanto para (a) como para (b) el AFE y el AFC no son dos categorías cualitativamente distintas sino que son, más bien, los dos polos de un continuo.

Tal y como se entiende tradicionalmente, en un análisis puramente exploratorio el investigador analizaría un conjunto de datos sin tener ninguna hipótesis previa acerca de su estructura, y dejaría que fuesen los resultados del análisis los que le proporcionasen información al respecto. Por otra parte, en un AFC el investigador habría planteado una serie de hipótesis bien especificadas que pondría a prueba evaluando el ajuste de un modelo. Estas hipótesis serían de tres tipos: (a) número de factores, (b) patrón de relaciones entre las variables y los factores, y (c) relaciones entre los factores.

En las primeras tentativas para evaluar un fenómeno nuevo, una postura puramente exploratoria como la que hemos descrito arriba sería aceptable. Sin embargo no parece la más adecuada cuando analizamos un test que hemos desarrollado o adaptado nosotros mismos. En este caso, es razonable suponer que tendremos una serie de hipótesis previas acerca del número de dimensiones que pretende medir el test, de cuales son los ítems que pretenden medir una u otra dimensión, y de si dichas dimensiones son o no independientes según la teoría. Posiblemente, sin embargo, estas hipótesis no sean aún lo suficientemente fuertes como para especificar un modelo AFC tal como se plantea habitualmente. Así pues, en cuanto a la finalidad, es útil considerar que la mayor parte de las aplicaciones psicométricas del AF se encontrarán en algún punto intermedio.

En cuanto al tipo de modelo que se pone a prueba, las distinciones se refieren aquí al grado de restricción en la solución propuesta. En un AFE las restricciones impues-

tas son las mínimas que se necesitan para obtener una solución inicial, solución que luego puede ser transformada. En un AFC las restricciones son mucho más fuertes y permiten poner a prueba una solución única que no es susceptible de posterior transformación. La finalidad con que se lleva a cabo el análisis y el tipo de modelo que se pone a prueba no son conceptos independientes. Cuanto mayor información previa se tenga y más fuertes sean las hipótesis, más especificada estará la solución puesta a prueba y mayor será el número de restricciones impuestas a dicha solución. Sin embargo, aún admitiendo esta clara relación, la distinción entre AF restricto y AF no restricto nos parecería más apropiada que la de AFE y AFC para referirnos al tipo de modelo que se pone a prueba.

En un AFC, tal como se usa habitualmente, las restricciones acerca del número de factores comunes así como las relaciones entre ellos son similares a las que se plantean en un AFE. Generalmente las correlaciones entre los factores se estiman libremente. Las diferencias principales se refieren a las restricciones que se imponen al patrón factorial. La solución que se propone casi siempre es una solución denominada "de conglomerados independientes" (McDonald, 1985) que sigue el principio de estructura simple. En dicha solución cada variable tiene una carga no nula en un solo factor común siendo cero las cargas en los restantes factores. Una solución de este tipo se presenta a continuación

Ítem	f_1	f_2
X_1	0.0	*
X_2	0.0	*
X_3	0.0	*
X_4	*	0.0
X_5	*	0.0
X_6	*	0.0

donde el asterisco indica que el peso correspondiente se estima como parámetro libre. En esta solución hipotética, los tres primeros ítems serian medidas puras del segundo factor y tendrían pesos nulos en el primero. Por otra parte los tres últimos ítems serían medidas puras del primer factor. Las soluciones de este tipo son teóricamente ideales. Tienen la máxima simplicidad estructural posible y permiten interpretar el contenido de cada factor sin ambigüedades.

Veamos ahora, en contraste, una solución exploratoria obtenida a partir del análisis de estos 6 ítems. Se obtuvo de un AFE en el que se propusieron dos factores comunes. La solución inicial arbitraria se transformó a una solución oblicua ya que, en teoría, los dos factores se consideraban relacionados. Es el patrón que habíamos presentado antes

Ítem	f_1	f_2
X_1	0.23	**0.78**
X_2	0.18	**0.82**
X_3	0.25	**0.75**
X_4	**0.78**	0.18
X_5	**0.77**	0.14
X_6	**0.78**	0.21

Es desde luego bastante clara, y el ajuste del modelo bifactorial fue bueno. Parece bastante evidente que los tres primeros ítems miden principalmente f_2 y los tres últimos f_1. Sin embargo, ¿son lo bastante "limpios" estos ítems como para ajustarse bien a la hipotética solución anterior?

Si evaluamos la adecuación de los datos a la solución hipotética presentada arriba, es decir, ajustamos un AFC convencional a estos ítems, lo que estamos planteando es que cada uno de ellos es una medida pura de un solo factor y, por tanto, que los pesos menores que aparecen en la solución AFE son debidos tan sólo a error de muestreo y son, por tanto, compatibles con valores exactamente de cero en la población.

El problema con este planteamiento es que, en el mundo real, la mayor parte de los ítems (y de los tests) no son medidas factorialmente puras. Con esfuerzo y tras un proceso de selección es posible llegar a obtener algunos ítems que son medidas (casi) puras. Estos ítems se denominan "marcadores" o "indicadores" en el lenguaje del AF. Sin embargo la pretensión que todos los ítems de un test sean marcadores nos parece una hipótesis poco realista.

Si aceptamos la idea de que en la mayor parte de los AF muchos de los ítems son factorialmente complejos, deberemos concluir que la hipótesis estructural más habitual en un AFC es falsa y que, por tanto, el modelo no ajustará bien. Más en detalle, si los pesos menores (estos que están en torno a 0.20 o por debajo) son menores pero no nulos, cada vez que fijamos uno de ellos a cero

cometemos un error de especificación del modelo. Si el modelo tiene pocos ítems, como en el ejemplo, quizás aún podamos llegar a un ajuste aceptable. Sin embargo, en modelos más grandes, la acumulación de los errores llevará necesariamente a ajustes inaceptables. Este razonamiento explica dos resultados que preocupan bastante en el campo aplicado (e.g. McCrae, et al. 1996). El primero es que estructuras factoriales obtenidas mediante AFE que son claras, interpretables y replicables a través de diferentes estudios, muestran ajustes inadmisibles cuando se evalúan mediante AFC. El segundo es que es más fácil obtener malos ajustes cuando se analizan cuestionarios de tamaño realista que cuando se analizan grupos muy reducidos de ítems. El primer resultado puede provocar en el investigador una desconfianza con respecto al AF. El segundo puede llevar a prácticas poco recomendables y a la eliminación innecesaria de ítems.

Nuestra posición puede resumirse así. En el análisis de ítems y tests, creemos que el AF debe venir guiado por la teoría previa. Esta teoría permitirá plantear hipótesis acerca del número de factores, del patrón (aproximado) que se espera encontrar, y de si los factores están o no relacionados. Sin embargo, generalmente, el conocimiento previo no será suficiente para especificar un modelo confirmatorio. Lo que proponemos es utilizar un modelo no restricto (exploratorio) pero con una finalidad confirmatoria hasta donde se pueda. Es decir, estimar una solución en la que se especifique el número de factores (o por lo menos un rango de valores) y también si estos factores son o no independientes. Además, debe tenerse una idea más o menos clara de cómo ha de ser el patrón transformado que se obtendrá. Por supuesto, si la investigación está lo bastante avanzada como para plantear una solución restricta, o todos los ítems son excepcionalmente simples, entonces el AFC es el modelo a usar.

EL DISEÑO DE UNA INVESTIGACIÓN BASADA EN EL ANÁLISIS FACTORIAL

Como en cualquier análisis estadístico, para que los resultados obtenidos mediante AF sean válidos, interpretables y generalizables, se requiere el cumplimiento de algunas condiciones básicas en el diseño de la investigación. Para ver la importancia de este punto en nuestro caso, es útil considerar el AF como un análisis a dos niveles. En el primer nivel se calculan las correlaciones entre una serie de medidas. En el segundo se analiza la estructura de dichas correlaciones. Si los resultados fallan ya en el primer nivel, nunca podrán ser correctos en el segundo. Por razones de claridad, discutiremos separadamente los dos aspectos básicos en el diseño: la muestra y las variables.

MUESTRA

En cualquier estudio factorial, y más aún en aquellos en que se desarrolla o adapta un test, debe tenerse una idea relativamente clara de la población de interés. Por tanto, el AF debería basarse una muestra representativa de esta población. Es muy habitual, sin embargo, utilizar muestras de conveniencia (generalmente estudiantes universitarios). Aparte de la no-representatividad, el problema estadístico más importante aquí es el de atenuación por restricción de rango. Si la muestra es muy homogénea en las variables a analizar (es decir, si las puntuaciones en los ítems/tests tienen poca variabilidad), las correlaciones obtenidas en el primer nivel del AF, estarán atenuadas. La matriz de correlaciones tendrá entonces mucho más "ruido" que "señal" y será difícil obtener una solución clara en el segundo nivel.

Posiblemente, el problema más discutido en AF en relación a la muestra es el de la estabilidad de la solución (¿Cuánta muestra se necesita para que una solución sea estable y generalizable?). Este es un problema complejo. La estabilidad de una solución factorial depende conjuntamente de tres factores: (a) el tamaño de muestra, (b) El grado de determinación de los factores y (c) la comunalidad de las variables. De forma que, si los factores están bien determinados y las variables tienen poco error de medida se podrán alcanzar soluciones estables con relativamente poca muestra. En este sentido queremos advertir que las "recetas" tradicionales tipo: 10 veces más sujetos que variables, etc. no tienen una base sólida.

Las medidas utilizadas habitualmente en psicología: tests y sobre todo ítems, contienen intrínsecamente mucho error de medida. Habrá que aceptar pues que las comunalidades serán generalmente bajas y, por tanto, se deberá actuar principalmente sobre los puntos (a) y (b). Con respecto al punto (b), que se discute con detalle más abajo, la idea de determinación de un factor refiere al número de variables que tienen pesos elevados en dicho factor. De otra forma, hasta qué punto el factor está bien definido y claramente medido por un buen número de indicadores. Con respecto al punto (a) es adecuado de nuevo pensar a "doble nivel". Los resultados del análisis al segundo nivel sólo podrán ser estables si lo son las correlaciones en que se basan. Y las correlaciones

tienen fluctuaciones muestrales altas. Cabe considerar entonces una muestra de 200 observaciones como un mínimo incluso en circunstancias ideales (altas comunalidades y factores bien determinados).

VARIABLES

El AF es un modelo para variables continuas e ilimitadas. Ni las puntuaciones de los ítems ni las de los test lo son. Por tanto, en la mayor parte de las aplicaciones psicológicas el AF deberá verse como un modelo aproximado cuya ventaja es la simplicidad. Es importante pues en primer lugar discutir en qué condiciones la aproximación será lo bastante buena para lo que se requiere en la práctica.

El AF funciona generalmente bien cuando se analizan puntuaciones en tests y subtests. En cuanto a los ítems, la aproximación suele ser también aceptable cuando se usan escalas de respuesta graduada (Likert) con 5 o más categorías. Finalmente, los ítems binarios y los ítems con 3 opciones y una categoría central son potencialmente los que pueden presentar más problemas. En principio recomendaríamos utilizar el formato de respuesta graduada siempre que sea posible.

Sea cual sea el tipo de respuesta, que el AF funcione bien o no depende sobre todo de la distribución de las puntuaciones. Las distribuciones simétricas no suelen dar problemas. Por otra parte los problemas más importantes suceden cuando (a) las distribuciones son marcadamente asimétricas y (b) las asimetrías van en ambas direcciones. Un ejemplo de esta situación sería el análisis de un test que contiene ítems muy fáciles e ítems muy difíciles. Las asimetrías de signo contrario dan lugar a relaciones no lineales y, por tanto, a la inadecuación del modelo AF lineal (Ferrando, 2002). Con relación a lo expuesto arriba, la magnitud del problema depende del tipo de variable a analizar. Con puntuaciones de tests es muy difícil que se produzcan relaciones no lineales. Con ítems de Likert es un problema a tener en cuenta. Finalmente es un problema muy frecuente en ítems binarios, conocido con el nombre de 'factores de dificultad' (McDonald y Alhawat, 1974).

Resultados obtenidos en simulación unidos a la propia experiencia, nos llevan a las siguientes recomendaciones. En el caso de tests y subtests el AF resulta casi siempre apropiado. En el caso de ítems de respuesta graduada, el AF se espera que funcione bien si los coeficientes de asimetría están todos en el intervalo entre -1 y +1. Finalmente, incluso los ítems binarios pueden ajustarse bien por el modelo lineal si los índices de dificultad se mueven entre 0.4 y 0.6. Cuando las variables tienen distribuciones más extremas, es necesario generalmente recurrir a enfoques no lineales que no podemos tratar aquí.

Aparte de la métrica y la distribución, hay otros factores a tener en cuenta en lo referente a las variables, sobre todo cuando se trata de ítems individuales. Como hemos dicho la fiabilidad de los ítems es intrínsecamente baja. Sin embargo, debería evitarse analizar ítems con fiabilidades excesivamente bajas, ya que dichos ítems sólo añadirían ruido a la solución factorial. Un estudio piloto convencional en el que se evalúen los índices de discriminación (correlaciones ítem-total) o las correlaciones test-retest ítem a ítem es muy recomendable. Permite eliminar aquellos ítems que sólo aportan ruido y empezar el AF desde un input más limpio.

En los cuestionarios de rendimiento típico (personalidad, motivación y actitudes), es relativamente frecuente incluir ítems redundantes: aquellos que son esencialmente la misma cuestión redactada quizás en forma ligeramente distinta. Estos ítems se utilizan para evaluar la consistencia de los sujetos o (solapadamente) para incrementar la consistencia interna del test. La presencia de ítems redundantes provoca siempre problemas en el AF. En efecto, los errores entre dos ítems redundantes no pueden ser independientes, ya que, aún después de eliminar los factores comunes, las respuestas siguen estando relacionadas debido a la semejanza de contenidos. La consecuencia es la necesidad de extraer factores adicionales definidos principalmente por parejas o tripletes de ítems redundantes. Estos factores pueden ser difíciles de identificar, sobre todo en soluciones rotadas. Un análisis de contenido previo puede eliminar redundancias y evitar estos problemas desde el principio.

Discutiremos por último el grado de determinación de los factores. Si es posible, una buena recomendación es la de utilizar marcadores o indicadores. Como hemos dicho antes los marcadores son, teóricamente, medidas puras de un factor. En forma más aplicada, Cattell (1988) las define como variables que, en estudios anteriores, han mostrado ser buenas medidas de los factores que se están evaluando. Su uso tiene principalmente dos funciones: (a) permiten identificar los factores aumentando su grado de determinación y (b) permiten relacionar los resultados del estudio con estudios anteriores. Cattell (1988) recomienda utilizar como mínimo dos marcadores por factor.

En cuanto a la relación entre el número de ítems y de factores, como sabemos, cuantos más ítems existan que midan con precisión un factor, más determinado estará dicho factor y más estable será la solución. Aunque existan recomendaciones divergentes (Cattell, 1988) nuestra opinión es que los mejores diseños en AF son aquellos en que se plantean pocos factores, se usan marcadores y se proponen un buen número de ítems para medir cada factor. Tanto si se usan marcadores como si no, para identificar claramente un factor se necesitan un mínimo de 4 variables con pesos substanciales en el mismo.

LAS ETAPAS DE UN ANÁLISIS FACTORIAL

Análisis preliminares: adecuación de los datos

De acuerdo con el planteamiento a doble nivel, parece lógico que antes de emprender un AF se utilicen indicadores para evaluar si las correlaciones obtenidas en el primer nivel son adecuadas para ser analizadas factorialmente en el segundo. Estos indicadores suelen denominarse "medidas de adecuación muestral" y su uso es muy importante como una etapa previa del AF: indicará si el AF es o no el modelo apropiado para los datos. Sin embargo, esta es la etapa que más se pasa por alto en investigación aplicada.

Para empezar, es conveniente inspeccionar los estadísticos descriptivos de las variables de acuerdo con la discusión en la sección anterior. A continuación debería contrastarse el test de esfericidad de Bartlett (1950). Dicho test pone a prueba la hipótesis nula de que la matriz de correlación poblacional es identidad, es decir, que las variables están incorreladas en la población. Si no puede rechazarse dicha hipótesis, habrá que aceptar que la matriz de correlación solo contiene "ruido". Es importante tener en cuenta que, aún así, si se factoriza dicha matriz se obtendrán factores. Sin embargo dichos factores serán totalmente espurios. En este sentido, es útil considerar el test de Bartlett como una prueba de seguridad y una condición necesaria. En la mayor parte de los AF se rechazará la hipótesis nula y, por tanto, se admitirá que existe alguna relación entre las variables. Sin embargo esto puede no ser suficiente. El modelo AF, como hemos visto, asume además que la relación es substancial. Si la relación es tan difusa que se necesitan prácticamente tantos factores como variables para explicarla, entonces no vale la pena llevar a cabo el análisis.

Supuesto que se cumpla la condición necesaria, en tercer lugar se evaluaría el grado de relación conjunta entre las variables. La medida más habitual es el KMO de Kaiser, (1970) que evalúa hasta que punto las puntuaciones en cada una de las variables son predecibles desde las demás. El rango de valores del KMO es de 0 a 1, y, cuanto más alto el valor, más substancialmente relacionadas entre ellas estarán las variables. Como valor de referencia, Kaiser (1970) sugiere que la matriz de correlación será apropiada para factorizar si el KMO es igual o superior a 0.80.

Estimación del modelo

Como hemos avanzado antes, esta es la etapa crucial del AF. En ella se estima una solución inicial y, sobre todo, se determina la dimensionalidad de los datos, es decir el número de factores más apropiado. La etapa de estimación debe guiarse por el principio de parsimonia. Se trata de determinar la solución más simple (es decir el menor número de factores) compatible con residuales suficientemente cercanos a cero.

El procedimiento de estimación implementado por defecto en los programas estadísticos suele ser el análisis en componentes principales (ACP). El ACP, sin embargo, no es un procedimiento para estimar el modelo factorial. Es un método para reducir el número de variables. En esencia, el AF es un modelo basado en el principio de que las variables tienen error de medida, distingue claramente entre varianza común (comunalidad) y varianza de error, y pretende reproducir tan sólo la varianza común, que es la que interviene en las correlaciones entre las variables. El ACP, en cambio, no hace esta distinción, sólo considera la varianza total y es esta varianza total la que pretende reproducir.

Los defensores del ACP argumentan que es más simple, que está mejor determinado y que produce virtualmente los mismos resultados que el AF (e.g. Velicer, 1990). Sin embargo esto último es una verdad a medias. Teóricamente, y desde el punto de vista del AF, el ACP podría considerarse como el caso extremo del modelo factorial en el que todas las variables a analizar están libres de error (es decir, varianza común y varianza total coinciden). En la práctica, el ACP y el AF llevan a resultados similares cuando: (a) el número de variables a analizar es grande (digamos más de 30) y (b) las variables tienen poco error y, por tanto, una elevada comunalidad (Mulaik, 1972). Un principio básico en Psicometría, sin embargo, es que las puntuaciones en los tests tienen error de medida (y las de los ítems mucho más). No parece, por tanto, muy razonable utilizar una técnica que no asume este principio.

El problema de usar ACP cuando el modelo correcto es el AF se ilustra con una pequeña simulación. Se generó

una matriz de correlación a partir de la siguiente solución factorial verdadera

0.50
0.50
0.50
0.50
0.50
0.50
0.50
0.50
0.50
0.50

A continuación la matriz de correlación se analizó mediante un método propiamente factorial y mediante ACP. La solución factorial directa (se especificaron dos factores) fue:

0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00
0.50	0.00

que recupera exactamente la solución verdadera. En cambio la solución ACP fue:

0.57	0	-0.82	0	0	0	0	0	0	0
0.57	0.03	0.09	-0.05	-0.78	0.22	0.03	-0.02	-0.01	-0.02
0.57	-0.04	0.09	-0.64	0.01	-0.47	-0.12	0	-0.1	-0.04
0.57	0.4	0.09	0.24	0.01	-0.34	0.44	0	0	0.34
0.57	0.15	0.09	0.18	0.12	-0.02	0.13	-0.15	-0.06	-0.74
0.57	0.06	0.09	0.18	0.12	0.13	-0.33	0.33	-0.59	0.1
0.57	0.09	0.09	0.24	0.01	-0.14	-0.55	0	0.51	0.05
0.57	0.03	0.09	-0.26	0.23	0.41	0.24	0.45	0.32	0
0.57	-0.73	0.09	0.24	0.01	-0.14	0.19	0.01	0.01	0.05
0.57	0	0.09	-0.13	0.23	0.35	-0.02	-0.63	-0.06	0.24

que muestra los dos problemas típicos del ACP: estimaciones sesgadas hacia arriba de los pesos en el factor de contenido y sobreestimación de la dimensionalidad. El primer componente es un estimador sesgado del único factor 'real' (cuyas cargas 'verdaderas' son todas de 0.50). Por otra parte, en los sucesivos componentes algunas de las variables tienen pesos por encima de 0.20-0.30. A este respecto es importante decir que, en la práctica, suele recomendarse interpretar tan sólo los pesos que estén por encima de estos valores mínimos (Catell, 1988, McDonald, 1985). McDonald (1985) propone un criterio heurístico más restrictivo en el que sólo se interpretarían aquellos factores que tuviesen al menos tres variables con pesos superiores a 0.30. Aún siguiendo este criterio más restrictivo, los resultados de la solución ACP llevarían a interpretar como factores 4 componentes que reflejan únicamente error. Si, además, esta solución hubiese sido posteriormente rotada la interpretación hubiera sido, posiblemente, totalmente errónea.

Existen varios métodos recomendables para estimar el modelo AF. Por limitaciones de espacio, discutiremos tan sólo los dos más utilizados: Mínimos Cuadrados Ordinarios (MCO) y Máxima Verosimilitud (MV). Sin embargo, hay otros métodos muy interesantes cuyo posterior estudio recomendamos al lector interesado. En particular vale la pena revisar el AF de rango mínimo (Shapiro y ten Berge, 2002).

El AF por MCO no es, propiamente, un método de estimación, sino un conjunto de métodos basados en un criterio general común. Para el número especificado de factores, los estimadores MCO son aquellos que hacen mínima la suma de cuadrados de las diferencias entre las correlaciones observadas y las reproducidas desde el modelo. Conceptualmente pues, los métodos MCO determinan la solución que hace que los residuales sean lo más cercanos a 0 posible. Esta es, como sabemos, la idea básica del ajuste en AF. Aunque el criterio es muy claro y directo, los métodos MCO son, en principio, puramente descriptivos. Como veremos, sin embargo, esto no es necesariamente una limitación.

Los principales métodos basados en el criterio MCO son: (a) AF de ejes principales, (b) MINRES de Harman (Harman y Jones, 1966), (c) ULS de Jöreskog (1977), y (d) Residual Mínimo de Comrey (1962). Para el mismo número de factores, las soluciones obtenidas con cualquiera de ellos son virtualmente idénticas. Sin embargo, recomendaríamos especialmente usar MINRES o ULS por dos razones: (a) no requieren la estimación inicial de las comunalidades y (b) son muy eficientes en términos de computación.

En contraste con los métodos MCO, el método MV (Lawley y Maxwell, 1971) es estadístico (inferencial). Su principal ventaja es que permite contrastar, de forma rigurosa, el ajuste del modelo a los datos mediante un índice referido a la distribución chi-cuadrado (χ^2). Esta ventaja, sin embargo, debe ser matizada. En primer lugar, la inferencia en AF MV se basa en el supuesto de que las variables que se analizan son continuas, métricas y distribuidas según la ley normal multivariante. En el caso de ítems y tests este supuesto nunca se cumple. En segundo lugar se asume que el modelo propuesto en

m factores ajusta perfectamente en la población y por tanto, que todo el error es error muestral (esta es la hipótesis nula del test de bondad de ajuste). Sin embargo, como hemos visto antes, los modelos se proponen tan sólo como razonables aproximaciones, y se acepta que parte del error será error de aproximación. Así pues, el método contrasta una hipótesis nula que ya sabemos desde el principio que es falsa y que se rechazará siempre en cuanto se tenga la potencia suficiente. Para agravar el tema, la potencia generalmente es muy alta, ya que en AF se trabaja habitualmente con muestras grandes. En suma, aún con distribuciones 'razonables' el uso del AF MV basado en el test de bondad de ajuste llevará casi siempre a la necesidad de estimar más factores de los que son substantivamente interpretables. Este fenómeno se denomina "sobrefactorización".

A pesar de los problemas discutidos arriba, existen razones para recomendar el uso del AF MV. En primer lugar, y aunque es poco conocido, la solución MV puede obtenerse también sin hacer supuestos inferenciales. Es la que hace mínimas las correlaciones parciales entre las variables tras eliminar de ellas la influencia de los factores (Mulaik, 1972). Esencialmente es el mismo criterio básico que el del AF por MCO. Los métodos MCO minimizan las correlaciones residuales. MV minimiza las correlaciones parciales (la correlación residual es el numerador de la correlación parcial). Por esta razón, en la práctica, las soluciones MCO y MV suelen ser muy similares. En segundo lugar, aunque el test χ^2 de bondad de ajuste evalúa una hipótesis falsa, existen indicadores de bondad de ajuste derivados de este test que evalúan el error de aproximación y el grado de ajuste del modelo. Como veremos en la siguiente sección estos indicadores son muy recomendables.

En una situación en que (a) las variables tienen distribuciones aceptables, (b) la solución esta bien determinada, y (c) el modelo propuesto es razonablemente correcto, las soluciones MCO y MV serán prácticamente idénticas. En este caso, el uso de MV tiene la ventaja de que permite obtener indicadores adicionales muy útiles en la evaluación del ajuste. En caso de distribuciones extremas, soluciones débiles o poco claras y notable error de aproximación la opción MV dará problemas. La convergencia del método es muy delicada y puede dar lugar a estimaciones inaceptables. Además, los indicadores adicionales no serán de fiar. En estos casos los métodos MCO son claramente superiores. De acuerdo con los estudios de simulación, son métodos muy robustos (convergen casi siempre) y, al no hacer distinciones entre las fuentes de error, suelen recuperar mejor la solución correcta que el método MV (MacCallum y Tucker 1991).

Evaluación del ajuste

Para decidir si un modelo con *m* factores resulta apropiado, debe evaluarse el grado de ajuste del modelo a los datos. Existen una variedad de criterios y procedimientos para llevar a cabo esta evaluación. En nuestra opinión, algunos son considerablemente mejores que otros.

Empezaremos con los criterios y procedimientos que no recomendamos. Posiblemente, el criterio más utilizado en Psicología, y el que suelen aplicar por defecto los programas comerciales es la regla de Kaiser: el número de factores relevantes es el número de valores propios mayores de 1 que tiene la matriz de correlación original. Este criterio presenta varios problemas, siendo el primero de ellos la falta de una justificación clara. Tiene varias (que no veremos aquí) pero ninguna convincente. El segundo problema es que se basa en la lógica del ACP no del AF. En efecto, los valores propios de la matriz sin reducir (con unos en la diagonal principal) equivalen a las proporciones de varianza total explicadas por los correspondientes componentes principales. Sin embargo, como hemos visto, la varianza que interesa realmente en el AF es la común, no la total. En tercer lugar, el número de factores determinado mediante esta regla está relacionado con el número de variables que se analizan. Más en detalle, si se analizan *n* variables, el criterio de Kaiser indicará un número de factores comprendido entre n/5 y n/3. Así, con 30 variables, el criterio indicará entre 6 y 10 factores. Sin embargo, si hemos diseñado una escala de 30 ítems para medir una sola dimensión, el número de factores esperado es 1, no de 6 a 10.

El test de sedimentación (Scree-test; Cattell, 1988) es un procedimiento gráfico ampliamente utilizado. En un gráfico bivariado, se representan puntos cuyas coordenadas son los valores propios de la matriz de correlación original (es decir, las proporciones de varianza total explicada) en el eje de ordenadas, y el número de componentes en el de abcisas. En una solución típica, el gráfico que une los puntos es una función decreciente, similar en forma a la ladera de una colina de residuos. A partir de cierto punto la función se hace prácticamente horizontal y es este punto el que, según Cattell, indica el número más adecuado de factores. La lógica es que, a partir de este número, los sucesivos factores son triviales y sólo explican varianza residual. Aunque la lógica es más convincente que la de la regla de Kaiser, el procedimiento tiene, en nuestra opinión, dos problemas. En primer lu-

gar la decisión se apoya en una inspección visual y, por tanto, tiene un fuerte componente de subjetividad. En segundo lugar, se basa en la lógica ACP y no distingue entre varianza común y de error. Ahora bien, si en lugar de los valores propios de la matriz sin reducir (proporciones de varianza total), se representasen los de la matriz reducida (es decir, las comunalidades) entonces el test sería útil como procedimiento auxiliar.

Existen dos criterios muy en boga actualmente, que son el MAP de Velicer (1976) y el análisis paralelo (AP; Horn, 1965). En nuestra opinión son de utilidad como criterios auxiliares, pero adolecen del mismo problema básico que tienen los criterios anteriores basados en la lógica ACP: no distinguen entre varianza común y varianza de error.

En el criterio MAP, se lleva a cabo un ACP en forma secuencial y en cada etapa se calcula la raíz media cuadrática de las correlaciones parciales que resultan si se elimina el componente correspondiente y los anteriores. Aunque las correlaciones residuales disminuyen siempre a medida que se estiman más componentes, las correlaciones parciales no lo hacen. De hecho, la función que relaciona la raíz media cuadrática de las parciales con el número de componentes suele tener forma de U. El mínimo de la función indicaría el número de componentes a retener.

El AP puede entenderse como una combinación del criterio de Kaiser y del scree test. En su forma más básica, consiste en generar una matriz de correlación aleatoria a partir de unos datos de la misma dimensión que los empíricos (sujetos y variables): teóricamente una matriz así debería tener todos los valores propios cercanos a 1. El método consiste en comparar los valores propios de la matriz empírica con los de la matriz generada al azar. Gráficamente la comparación sería como un doble scree test en el que se representan simultáneamente la curva correspondiente a los datos empíricos y la correspondiente a los datos al azar. La primera, como sabemos, se esperaría que mostrase un fuerte descenso seguido de estabilización. La segunda debería mostrar una tendencia mucho más plana (en una matriz muy grande sería una recta horizontal que cortaría a la ordenada en 1). El punto de intersección entre las dos curvas indicaría el número de factores a retener. Visto así el AP comparte muchas de las críticas del scree-test, pero tiene la ventaja de que el criterio para determinar el número de factores es mucho más objetivo.

Pasamos ahora a discutir los criterios y procedimientos que consideramos más recomendables. Empezaremos por aquellos de tipo general que pueden aplicarse cualquiera que sea el método de estimación. Después discutiremos aquellos específicamente relacionados con el AF MV.

Como bien sabemos ya, si el número de factores propuesto es apropiado, entonces las correlaciones residuales entre las variables tras eliminar la influencia de los factores deben ser todas prácticamente cero. De acuerdo con este principio, los criterios más claros para evaluar el ajuste de un modelo en *m* factores serán aquellos que más directamente se relacionan con la evaluación de las correlaciones residuales.

En problemas pequeños, la inspección visual de la matriz de residuales puede dar ya una idea importante del grado de ajuste. Sin embargo, habitualmente el AF trabaja con un número substancial de variables, y la inspección global de la matriz residual no resulta práctica. En este caso, deberá condensarse la información mediante estadísticos descriptivos.

Para empezar es conveniente inspeccionar la distribución de frecuencias de los residuales. Si el número de factores propuesto es adecuado, esta distribución será simétrica, aproximadamente normal y centrada en torno a una media de cero. Distribuciones asimétricas, descentradas o con las colas muy pesadas, indican que aún queda covariación sistemática por explicar y, por tanto, que es necesario estimar más factores.

La raíz media cuadrática residual (RMCR) es una medida descriptiva que indica la magnitud media de las correlaciones residuales. Si la media de éstas últimas es cero, entonces la RMCR coincide con la desviación típica de los residuales. Harman (1976), propone un valor de referencia de 0.05 o menor para considerar que el ajuste del modelo era aceptable. Este criterio es puramente empírico, pero funciona generalmente bien en la práctica. Mejor fundamentado es el criterio propuesto inicialmente por Kelley (1935). El error típico de un coeficiente de correlación de cero en la población es, aproximadamente, $1/\sqrt{N}$ donde *N* es el tamaño de muestra. Este sería, por tanto, el valor de referencia. Así, en una muestra de 300 sujetos, $1/\sqrt{N}=0.058$. Si la RMCR se mueve en torno a este valor, o es menor, cabe interpretar que los valores residuales observados no difieren significativamente de cero y, por tanto, que no quedan ya relaciones sistemáticas por explicar.

El índice gamma o GFI propuesto inicialmente por Tanaka y Huba (1985) es una medida de bondad de ajuste normada entre 0 y 1 que puede utilizarse con la mayoría de los procedimientos de estimación. Puede interpretarse como un coeficiente de determinación multivariado que indica la proporción de covariación entre las variables explicada por el modelo propuesto. De acuerdo con los criterios actuales (véase el artículo de Ruíz, Pardo y

San Martín en este volumen) para considerar un ajuste como bueno, el GFI debería estar por encima de 0.95.

Discutiremos por último dos indicadores que se utilizan en el caso de estimación por MV. Son, por tanto, inferenciales, pero aquí recomendamos su uso de acuerdo con una lógica descriptiva. El primero de ellos es el coeficiente TLI-NNFI, propuesto inicialmente por Tucker y Lewis (1973) precisamente para el modelo AF. Es un índice relativo y mide la mejora de ajuste que produce el modelo propuesto con respecto al modelo nulo en 0 factores, en relación a la mejora esperada por un modelo que ajustara bien. Sus valores se mueven entre 0 y 1 (aunque no esté estrictamente normado) y Tucker y Lewis recomendaron interpretarlo como un coeficiente de fiabilidad. En este sentido, y aunque los criterios actuales son más rigurosos (véase Ruíz, Pardo y San Martín en este volumen), nuestra opinión es que valores por encima de 0.85-0.90 empezarían a considerarse aceptables.

El índice RMSEA (Steiger y Lind, 1980), muy en boga actualmente, estima el error de aproximación del modelo propuesto. Más específicamente, estima la discrepancia que habría entre la matriz de correlación poblacional y la matriz reproducida por el modelo propuesto, también en la población. Conceptualmente, el RMSEA se basa en el enfoque, discutido antes, de que los modelos son sólo aproximaciones y estima hasta qué punto el modelo puesto a prueba es una aproximación razonable. El RMSEA es un índice relativo a los grados de libertad (complejidad) del modelo y, por tanto, puede penalizar a los modelos menos parsimoniosos. Como referencia, valores por debajo de 0.05 podrían considerarse como indicadores de buen ajuste, en tanto que valores entre 0.05-0.08 indicarían un ajuste admisible.

Cerraremos esta sección con tres observaciones. En primer lugar, no recomendamos tomar una decisión tan importante como la del número apropiado de factores basándose en un solo indicador o criterio. Es conveniente utilizar múltiples indicadores que nos proporcionen más información y, por tanto, elementos de juicio. En segundo lugar, en situaciones reales el proceso que lleva a la determinación del número de factores no es tan lineal como lo describimos aquí por razones didácticas. En efecto, una vez obtenida la solución transformada podemos encontrarnos con que uno o más factores sean muy débiles, estén pobremente identificados o reflejen contenidos triviales (por ejemplo, que no lleguen al mínimo de 3 variables con pesos superiores a 0.3 antes comentado). Este resultado podría llevar a la reconsideración del número de factores y a una nueva inspección de la solución. Por último, en muchos casos la información teórica previa no indica un número claro de factores sino que es más difusa. Puede indicar un rango plausible de factores o quizás varias soluciones alternativas plausibles. Es conveniente en este caso examinar las diferencias entre los valores de los indicadores de ajuste correspondientes a las diferentes soluciones evaluadas.

Obtención de la solución transformada (rotación)

En el análisis factorial no restricto, la solución inicial estándar que se obtiene mediante la aplicación de los métodos de MCO o MV se denomina “forma canónica”. Para el número de factores especificado, esta solución tiene la propiedad de que los factores sucesivos explican el máximo posible de la varianza común. En algunos casos la solución canónica es directamente interpretable y no requiere ninguna transformación posterior. El caso más claro es cuando se analiza un conjunto de ítems que pretenden medir una sola dimensión. Dado que el primer factor explica el máximo posible de la varianza común, si el conjunto es esencialmente unidimensional, entonces los factores que siguen al primero deberán ser residuales. Un ejemplo extremo es la solución directa en dos factores que hemos usado arriba para comparar el AF con ACP. Es una solución canónica en la que el primer factor explica ya toda la varianza común y al segundo no le queda nada por explicar. Este grupo de variables sería pues perfectamente unidimensional.

Cuando se espera encontrar una solución en múltiples factores, sin embargo, la solución canónica inicial es arbitraria. Debe ser entonces transformada o rotada hasta obtener una solución interpretable de acuerdo con la teoría previa.

La principal decisión que debe tomar el investigador en esta etapa es si utilizará una rotación oblicua o una ortogonal. Este es un tema controvertido. Los autores que defienden las soluciones ortogonales, consideran que son más simples y fáciles de interpretar. Además creen que son también más estables en estudios de replicación. La base estadística de este argumento es que, si los factores son independientes en la población, no lo serán exactamente en las muestras y, por tanto, si se utilizan soluciones oblicuas, las correlaciones entre factores reflejarán tan sólo error muestral Por otra parte, los autores que defienden las soluciones oblicuas consideran que la mayor parte de constructos psicológicos están relacionados y que exigir factores incorrelados es imponer artificialmente una solución que no es correcta tan sólo porque es más sencilla (e.g. Mulaik, 1972). En definitiva, que ha de ser la teoría la que guíe esta decisión.

Los autores nos situamos esencialmente en la segunda postura. Sin embargo, también creemos que debe tenerse en cuenta (como siempre en AF) el criterio de parsimonia. Si la teoría no permite hipótesis fuertes al respecto, parece razonable empezar con una solución oblicua. Si las correlaciones estimadas entre factores son substanciales, esta es la solución a mantener. Sin embargo, si las correlaciones entre factores son consistentemente bajas (digamos por debajo de 0.30 o 0.20), entonces podría hacerse un segundo análisis especificando una solución ortogonal. Si las dos soluciones fuesen similares, sería preferible aceptar provisionalmente la solución ortogonal.

Una vez decidido el tipo general de rotación, la importancia de la decisión respecto al método específico a utilizar depende de la solidez del diseño. Si las variables son 'limpias', se utilizan marcadores y los factores están bien determinados, entonces los diferentes métodos deben llevar esencialmente a la misma solución. No es una mala estrategia probar diferentes métodos y evaluar si convergen aproximadamente a una solución común. En soluciones complejas, ruidosas y pobremente determinadas el empleo de diferentes métodos de rotación puede llevar a soluciones muy dispares. Este no es un problema de los métodos, es un problema de diseño.

Los métodos analíticos de rotación ortogonal son generalmente 'cuárticos' en el sentido de que se basan en las potencias cuartas de las los pesos factoriales (conceptualmente, varianzas de los cuadrados de los pesos). Existen dos métodos generales de este tipo (véase e.g. Mulaik, 1972): Quartimax y Varimax. Dado el patrón inicial no rotado (variables ´ factores), la transformación Quartimax es la que maximiza la varianza de los cuadrados de los pesos por filas. En cambio Varimax maximiza la varianza por columnas. La solución Quartimax es pues compatible con una columna del patrón en el que la mayor parte de los pesos son elevados y tiende a dar soluciones con un factor general. En cambio Varimax tiende a dar soluciones múltiples en las que no hay un factor dominante. Existe un tercer método, Equamax que combina ambos criterios y lleva por tanto a soluciones intermedias. Los tres métodos funcionan bien, y tal vez la elección de uno de ellos debería venir guiada por lo esperado desde la teoría (existencia o no de un factor general). Nuestra consideración positiva de los métodos (Varimax en particular) no es incompatible con la crítica hecha al principio del capítulo. Nuestra crítica inicial se refiere al uso indiscriminado de Varimax por defecto aun cuando la teoría indique claramente que los factores están relacionados o cuando se espera encontrar un factor general.

La experiencia con los métodos anteriores basada en estudios de simulación sugiere que pueden llevar a resultados erróneos o inestables (sobre todo Varimax) cuando una elevada proporción de las variables a analizar son factorialmente complejas. Como hemos dicho arriba, este no es un problema del método de rotación sino de un mal diseño. Para minimizar el problema se han propuesto versiones ponderadas (weighted) de los tres métodos (Quartimax, Varimax y Equamax) en los que se asigna mayor peso a los ítems inicialmente evaluados como más simples.

Los métodos analíticos tradicionales de rotación oblicua son una extensión de los métodos cuárticos ortogonales, con el problema adicional de determinar el grado de relación (oblicuidad) entre los factores. En la práctica, el método más utilizado con diferencia es Oblimin (véase e.g. Mulaik, 1972) cuyo criterio puede ser considerado como una combinación de los criterios Quartimax y Varimax extendida al caso oblicuo. El criterio Oblimin incluye un parámetro (delta, que puede tomar valores entre 0 y 1) y que pondera la maximización de la simplicidad por filas o por columnas. Indirectamente el parámetro delta controla también el grado de oblicuidad permitido en la solución. Los programas factoriales utilizan por defecto el valor delta=0 siguiendo la recomendación de Jennrich (1979). Este valor suele llevar a una buena convergencia y a una solución simple e interpretable. En contrapartida, para conseguir esta simplicidad, los factores resultantes suelen estar muy relacionados entre ellos. Browne (comunicación personal) recomienda utilizar delta=0.5.

Una alternativa interesante a Oblimin son ciertos métodos analíticos que incorporan la idea de rotación sobre una matriz diana que se construye analíticamente. Esencialmente la idea es la siguiente. En primer lugar, a partir de una solución ortogonal se construye una matriz hipótesis o diana. Dicha matriz es una modificación de la solución ortogonal lo más cercana posible a la estructura simple; así, y principalmente, los pesos muy bajos en la solución ortogonal se hipotetizan como ceros en la matriz diana. En segundo lugar se determina la solución transformada oblicua que mejor se aproxima a la matriz diana. El método inicial que incorporaba estas ideas es Promax (Hendrickson y White, 1964). Lorenzo-Seva (1999) ha propuesto recientemente un método mucho más refinado denominado Promin. Con respecto a los métodos previos, Promin incorpora mejoras en todas las etapas: solución ortogonal de partida, determinación de la matriz diana y procedimientos y criterios de aproximación. Es, por tanto, el método más recomendable dentro de esta familia.

SOFTWARE: EL PROGRAMA FACTOR[1]

Si bien es cierto que el AFE es una técnica *clásica* de análisis de datos, hoy en día la investigación estadística sobre el propio análisis continúa siendo muy activa. Así, en los años recientes se han publicado en revistas especializadas diversos avances relacionados con los métodos utilizados en AFE. Sin embargo, los autores de los programas informáticos de análisis de datos más populares (no hace falta citar nombres) no parecen interesados en implementar estos nuevos avances. Por esta razón, los investigadores de las propias universidades se han preocupado por desarrollar programas específicos que recojan tanto los métodos clásicos como las nuevas aportaciones. Un ejemplo de este tipo de programas es Factor (Lorenzo-Seva y Ferrando, 2006). Se trata de un programa fácil de usar (basado en los típicos menús de Windows) que tiene como finalidad el cálculo de AFE.

Factor implementa procedimientos e índices clásicos, así como algunas de las aportaciones metodológicas más recientes. Esto incluye todos los discutidos en este capítulo y además otros de indudable interés y que debieran ser objeto de estudio futuro para los lectores interesados en el AF. Así por ejemplo Factor permite trabajar con correlaciones policóricas cuando se sospecha que el modelo lineal puede ser inadecuado. Buenos ejemplos de la metodología propuesta recientemente que se han implementado en Factor son: (1) Minimum Rank Factor Analysis, que es el único procedimiento de extracción de factores que permite estudiar la proporción de varianza explicada por cada factor; y (2) el método de rotación Simplimax, que ha demostrado ser el método de rotación más eficiente de todos los propuestos hasta el momento. Muchos de estos métodos no están disponibles en ningún programa comercial.

Finalmente, cabe destacar que Factor es un programa que se distribuye gratuitamente como software de libre disposición. Se puede obtener en Internet, en la página: http://psico.fcep.urv.es/utilitats/factor/. En la misma página se ofrece un breve manual de utilización, así como una versión de demostración para ejemplificar su utilización. Hasta la fecha, y desde que se puso a disposición de los investigadores en el año 2006, se ha utilizado por investigadores de las más diversas procedencias en 29 artículos internacionales aparecidos en revistas recogidas en el ISI.

[1] El autor de esta sección es Urbano Lorenzo-Seva

EJEMPLO ILUSTRATIVO

A continuación se propone al lector la realización de un ejercicio práctico para aplicar el material expuesto. Los datos se pueden encontrar en la siguiente dirección: http://psico.fcep.urv.cat/ejemplopapeles. Se trata de un test de 14 ítems que mide dos tipos de ansiedad. Se propone al lector que mediante el programa factor, realice el análisis factorial del test para determinar la estructura y propiedades del mismo. En la misma web se encontrará la resolución de este ejemplo, aunque es recomendable que el lector trabaje sobre el ejemplo antes de leer la resolución.

BIBLIOGRAFÍA RECOMENDADA

- **(Manuales de tipo general)** los de Harman (1976) y Mulaik (1972), ambos en la lista de referencias, tratan bastante a fondo los principales aspectos del AF, incluyendo los que no hemos podido desarrollar. Su nivel técnico es bastante más elevado que el que hemos utilizado aquí.

  El manual de McDonald (1985, en lista de referencias) es mucho más personal que los dos anteriores. Se muestra muy crítico con el enfoque tradicional del AF y refleja las fuertes posiciones del autor. Muy recomendable.

  Por último, el siguiente texto del autor:

  Ferrando, P.J. (1993) Introducción al análisis factorial. Barcelona: PPU.

  Puede usarse como una ampliación más técnica de los aspectos básicos desarrollados en las primeras secciones.
- **La problemática AFE vs AFC en el análisis de ítems se discute a fondo en**:

  Ferrando, P.J. y Lorenzo-Seva, U. (2000).

  "Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions", *Psicológica*. 21, 301-323.
- **La siguiente tesis proporciona una buena revisión de los principales criterios y métodos de rotación**

  Lorenzo-Seva, U. (1997). Elaboración y evaluación de métodos gráficos en análisis factorial. Universidad de Barcelona.
- **Por último, el desarrollo paso a paso de un AF mediante un programa informático se describe en:**

  Ferrando, P.J. y Lorenzo, U. (1998)

  "Análisis factorial".

  En: Renom, J. (Coord.) 'Tratamiento informatizado de datos'. Cap. 5 pp 101-126. Barcelona. Masson.

REFERENCIAS

Bartlett, M.S. (1950). Tests of significance in factor analysis. British Journal of Mathematical and Statistical Psychology, 3, 77-85.

Carroll, J.B. (1978). How shall we study individual differences in cognitive abilities?-Methodological and theoretical perspectives. Intelligence, 2, 87-115.

Cattell, R.B. (1988). The meaning and strategic use of factor analysis. In J.R. Nesselroade and R.B. Cattell (eds.) Handbook of multivariate experimental psychology (pp 131-203). New York: Plenum Press.

Comrey. A.L. (1962). The minimum residual method of factor analysis. Psychological Reports, 11, 15-18.

Ferrando, P.J. (1993) Introducción al análisis factorial. Barcelona: PPU.

Ferrando, P.J. y Lorenzo-Seva, U. (1998). Análisis factorial. En: Renom, J. (Coord.) Tratamiento informatizado de datos. Cap. 5 (pp 101-126). Barcelona: Masson.

Ferrando, P.J. y Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: some aspects of the problem and some suggestions. Psicológica, 21, 301-323.

Ferrando, P.J. (2002). Theoretical and empirical comparisons between two models for continuous item responses. Multivariate Behavioral Research, 37, 521-542.

Harman, H.H. y Jones, W.H. (1966). Factor analysis by minimizing residuals (minres). Psychometrika, 31, 351-368.

Harman, H.H. (1976). Modern factor analysis. Chicago: Univ. of Chicago press.

Hendrickson, A.E. y White, P.O. (1964). Promax: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology. 17, 65-70.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179-185.

Jöreskog, K.G. (1977). Factor analysis by least-squares and maximum-likelihood methods. En: K. Enslein, A. Ralston, & H.S. Wilf (Eds.): Statistical methods for digital computers. (pp 125-153). New York: Wiley.

Kaiser, H.F. (1970). A second generation little jiffy. Psychometrika, 35, 401-415.

Kelley, T.L. (1935). Essential Traits of Mental Life. Harvard Studies in Education, 26 (pp. 146). Cambridge: Harvard University press.

Lawley, D.N. y Maxwell, A.E. (1971). Factor analysis as a statistical method. London: Butterworths.

Lorenzo-Seva, U. (1997). Elaboración y Evaluación de Métodos Gráficos en Análisis Factorial. Tesis doctoral. Universidad de Barcelona.

Lorenzo-Seva, U. (1999). Promin: A Method for Oblique Factor Rotation. Multivariate Behavioral Research, 34, 347-365.

Lorenzo-Seva, U. y Ferrando P.J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. Behavior Reserch Methods, 38, 88-91.

MacCallum, R.C. y Tucker, L.R. (1991). Representing sources of error in the common factor model: implications for theory and practice. Psychological Bulletin, 109, 502 511.

McCrae, R.R., Zonderman, A.B., Costa, P.T., Bond, M.H. y Paunonen, S.V. (1996). Evaluating replicability of factors in the revised NEO personality inventory: confirmatory factor analysis versus Procustres rotation. Journal of Personality and Social Psychology, 70, 552-566.

McDonald, R.P. (1985). Factor analysis and related methods. Hillsdale: LEA

McDonald, R.P. and Ahlawat, K.S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

Mulaik, S.A. (1972). The foundations of factor analysis. New York: McGraw-Hill.

Shapiro, A. y ten Berge, J.M.F. (2002). Statistical inference of minimum rank factor análisis. Psychometrika, 67, 79-94.

Spearman, Ch. (1904). General intelligence; objectively determined and measured. American Journal of Psychology, 115, 201292.

Steiger, J.H. y Lind, J. (1980). Statistically based tests for the number of common factors. Comunicación presentada en el meeting anual de la Psychometric Society. Iowa City, Mayo de 1980.

Tanaka, J.S. y Huba, G.J. (1985). A fit index for covariance structure models under arbitrary GLS estimation. British Journal of Mathematical and Statistical Psychology, 38, 197-201.

Thurstone, L.L. (1947). Multiple factor analysis. Chicago: University of Chicago press.

Tucker, L.R. y Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. Psychometrika 41, 321–337.

Velicer, W.F. y Jackson, D.N. (1990). Component analysis versus common factor analysis: Some further observations. Multivariate Behavioral Research. 25, 97-114.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 34-45
http://www.cop.es/papeles

MODELOS DE ECUACIONES ESTRUCTURALES

STRUCTURAL EQUATION MODELS

Miguel A. Ruiz, Antonio Pardo y Rafael San Martín
Facultad de Psicología. Universidad Autónoma de Madrid

En este capítulo se presentan los modelos de ecuaciones estructurales, una técnica de análisis estadístico multivariante utilizada para contrastar modelos que proponen relaciones causales entre las variables. Tras la definición de este tipo de modelos y la presentación de un ejemplo típico, se discute el concepto de causalidad, para entender su utilización en este contexto. A continuación se discute la estructura general que tiene un modelo, los tipos de variables que se pueden utilizar en ellos y su representación mediante diagramas estructurales, acompañado de la discusión de un ejemplo. Posteriormente se presentan los pasos en la elaboración de un modelo y los tipos de relaciones posibles. También se comentan brevemente el concepto de ajuste y los problemas típicos de estos modelos. Por último se ofrecen algunos recursos adicionales.
Palabras clave: *Modelos de ecuaciones estructurales, Variables latentes, Variables observadas, Diagrama de rutas, Modelos de rutas, Análisis de estructuras de covarianza, Análisis factorial confirmatorio, Bondad de ajuste, Modelos causales.*

In this chapter, structural equation models (SEM) are presented. SEM is a multivariate statistical technique used to test models proposing causal relations between their variables. After defining this type of models and presenting a typical example, the concept of causation is discussed, in order to understand it's meaning in the present context. The general model structure, the types of variables used, and how to represent them in path diagrams are discussed, accompanied with an example. Steps needed to build a model are presented and the different types of relations are commented. Goodness of fit is also briefly commented and also typical problems found in these models. Some additional resources are also presented.
Key words: *Structural equation models, Latent variables, Observed variables, Path diagrams, Path analysis, Analysis of covariance structures, Confirmatory factor analysis, Goodness of fit, Causal models.*

Los modelos de ecuaciones estructurales son una familia de modelos estadísticos multivariantes que permiten estimar el efecto y las relaciones entre múltiples variables. Los modelos de ecuaciones estructurales nacieron de la necesidad de dotar de mayor flexibilidad a los modelos de regresión. Son menos restrictivos que los modelos de regresión por el hecho de permitir incluir errores de medida tanto en las variables criterio (dependientes) como en las variables predictoras (independientes). Podría pensarse en ellos como varios modelos de análisis factorial que permiten efectos directos e indirectos entre los factores.

Matemáticamente, estos modelos son más complejos de estimar que otros modelos multivariantes como los de Regresión o Análisis factorial exploratorio y por ello su uso no se extendió hasta 1973, momento en el que apareció el programa de análisis LISREL (*Linear Structural Relations*; Jöreskog, 1973). El LISREL fue perfeccionado, dando lugar al LISREL VI (Jöreskog y Sörbom, 1986), que ofrecía una mayor variedad de métodos de estimación. El EQS (Abreviatura de *Equations*: Bentler, 1985) es el otro paquete utilizado tradicionalmente para este tipo de análisis. En la actualidad, existen otros programas de estimación en entorno gráfico, como el AMOS (Analysis of Moment Structures; Arbuckle, 1997). Tal ha sido la influencia de los programas de estimación en la posibilidad de desarrollo de los modelos de ecuaciones estructurales, que no es infrecuente que se los denomine modelos LISREL. En la literatura internacional se los suele llamar modelos SEM, abreviatura de Structural Equation Models.

La gran ventaja de este tipo de modelos es que permiten proponer el tipo y dirección de las relaciones que se espera encontrar entre las diversas variables contenidas en él, para pasar posteriormente a estimar los parámetros que vienen especificados por las relaciones propuestas a nivel teórico. Por este motivo se denominan también *modelos confirmatorios*, ya que el interés fundamental es "confirmar" mediante el análisis de la muestra las relaciones propuestas a partir de la teoría explicativa que se haya decidido utilizar como referencia.

Como podemos apreciar en el siguiente ejemplo, la especificación teórica del modelo permite proponer estructuras causales entre las variables, de manera que unas variables causen un efecto sobre otras variables que, a su vez, pueden trasladar estos efectos a otras variables, creando concatenaciones de variables.

Correspondencia: Miguel Ángel Ruiz. *Departamento de Psicología Social y Metodología. Facultad de Psicología. Universidad Autónoma de Madrid. Calle Iván Pavlov 6. 28049 Madrid. España. Email: miguel.ruiz@uam.es*

La figura 1 muestra un modelo de ecuaciones estructurales perteneciente al campo de la salud (González y Landero, 2008). Este tipo de modelos en particular también se denominan modelos de análisis de rutas (path analysis) y en él todas las variables son observables, excepto los errores de predicción. La finalidad de este modelo concreto es predecir la magnitud de los síntomas psicosomáticos de una persona a partir de un conjunto de antecedentes personales. El modelo plantea la existencia de tres variables predictoras (autoestima, autoeficacia y apoyo social) que influyen en el nivel de estrés del individuo. A su vez el estrés influye de manera directa sobre la magnitud de los síntomas psicosomáticos y también de manera indirecta, modulado por el nivel de cansancio emocional. Como puede observarse, el modelo propuesto es algo más complejo que un modelo de regresión ya que algunas variables juegan el papel de variable predictora y de variable dependiente de manera simultánea.

Interpretando brevemente la magnitud y el signo de los parámetros estimados, los resultados constatan que las variables predictoras tienen un efecto negativo sobre el nivel de estrés, de manera que una menor autoeficacia percibida, una menor autoestima y un menor apoyo social generan un mayor nivel de estrés. Además, la autoeficacia percibida es el predictor con mayor efecto y todos los predictores se relacionan unos con otros. Con los predictores utilizados se puede explicar el 42% de la variabilidad del estrés. Además, el estrés influye directa y positivamente (0,16) sobre los síntomas psicosomáticos, pero el efecto indirecto a través del cansancio emocional es mayor (0,21=0,54*0,39). En total se explica el 24% de las diferencias encontradas en los síntomas psicosomáticos de los sujetos. El significado de estos y otros elementos de la figura se explicarán más adelante.

El nombre que reciben los modelos de ecuaciones estructurales es debido a que es necesario utilizar un conjunto de ecuaciones para representar las relaciones propuestas por la teoría. Para representar las relaciones del ejemplo anterior se están utilizando y estimado simultáneamente tres ecuaciones de regresión.

Existen muchos tipos de modelos con distinto nivel de complejidad y para distintos propósitos. Todos ellos son modelos de tipo estadístico. Esto quiere decir que contemplan la existencia de errores de medida en las observaciones obtenidas de la realidad. Habitualmente incluyen múltiples variables observables y múltiples variables no observables (latentes), aunque algunos como el del ejemplo sólo contemplan como variables latentes los errores de predicción.

Respecto a su estimación, los modelos de ecuaciones estructurales se basan en las correlaciones existentes entre las variables medidas en una muestra de sujetos de manera transversal. Por tanto, para poder realizar las estimaciones, basta con medir a un conjunto de sujetos en un momento dado. Este hecho hace especialmente atractivos estos modelos. Ahora bien, hay que tener en cuenta que las variables deben permitir el cálculo de las correlaciones y por ello deben ser variables cuantitativas, preferentemente continuas.

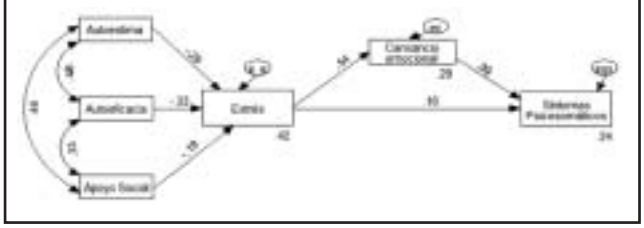
Los puntos fuertes de estos modelos son: haber desarrollado unas convenciones que permiten su representación gráfica, la posibilidad de hipotetizar efectos causales entre las variables, permitir la concatenación de efectos entre variables y permitir relaciones recíprocas entre variables.

Son muchos los tipos de modelos que se pueden definir con esta metodología. A continuación se enuncian los más populares de los mencionados en la literatura estadística: Regresión múltiple con multicolinealidad, Análisis factorial confirmatorio (ver Ferrando y Anguiano, 2010), Análisis factorial de 2º orden, Path analysis, Modelo causal completo con variables latentes, Modelo de curva latente (ver Bollen y Curran, 2006), Modelos multinivel (ver Skrondal y Rabe-Hesketh, 2004), Modelos multigrupo, Modelos basados en las medias (ANOVA, ANCOVA, MANOVA y MANCOVA; ver Bagozzi y Yi, 1994) y Análisis de mediación (ver Preacher y otros, 2007).

EL CONCEPTO DE CAUSALIDAD

Una potencialidad interesante de estos modelos es la posibilidad de representar el efecto causal entre sus variables. Aunque resulte muy atractivo el hecho de poder representar gráficamente la influencia causal de una variable sobre otra y aunque también seamos capaces de estimar el parámetro correspondiente a ese efecto, debemos tener claro que la estimación del parámetro no "demuestra" la existencia de causalidad. La existencia de una relación causal entre las variables debe venir sustentada por la articulación teórica del modelo y no por su estimación con datos de tipo transversal. Para demostrar

FIGURA 1
MODELO DE RUTAS DE EJEMPLO



científicamente la existencia de una relación causal deberemos recurrir al diseño de un experimento controlado con asignación aleatoria de los sujetos a las condiciones del estudio (ver Pardo, Ruiz y San Martín, 2009, págs. 356-359). No debemos olvidar que los modelos de ecuaciones estructurales se utilizan en estudios de tipo correlacional en los que tan solo se observa la magnitud de las variables y en los que nunca se manipulan éstas.

Los trabajos de Boudon (1965) y Duncan (1966) abrieron una nueva posibilidad de aproximación al problema de la causalidad, distinta de la manipulación experimental, proponiendo el análisis de dependencias o *análisis de rutas* (path analysis). En este tipo de análisis se estudia una teoría causal mediante la especificación de todas las variables importantes para dicha teoría. Posteriormente, se pueden derivar las relaciones entre los efectos causales a partir de la teoría causal para, en último término, estimar el tamaño de estos efectos. La generalización del modelo de análisis de rutas dio lugar a los modelos de ecuaciones estructurales para la comprobación de teorías o, lo que es lo mismo, de modelos causales. La lógica de estos modelos establece que, basándose en la teoría que fundamenta el modelo, será posible derivar las medidas de covariación esperadas entre las variables a partir de los efectos causales del modelo. Si la teoría es correcta, las medidas de covariación derivadas del modelo y las medidas de covariación obtenidas a partir de los datos deberán ser iguales.

ESTRUCTURA DE UN MODELO

Un modelo de ecuaciones estructurales completo consta de dos partes fundamentales: el modelo de medida y el modelo de relaciones estructurales.

El modelo de medida contiene la manera en que cada constructo latente está medido mediante sus indicadores observables, los errores que afectan a las mediciones y las relaciones que se espera encontrar entre los constructos cuando éstos están relacionados entre sí. En un modelo completo hay dos modelos de medida, uno para las variables predictoras y otro para las variables dependientes.

El modelo de relaciones estructurales es el que realmente se desea estimar. Contiene los efectos y relaciones entre los constructos, los cuales serán normalmente variables latentes. Es similar a un modelo de regresión, pero puede contener además efectos concatenados y bucles entre variables. Además, contiene los errores de predicción (que son distintos de los errores de medición).

Existen dos casos excepcionales en los que el modelo no contiene ambas partes y que se usan con relativa frecuencia. En primer lugar, los modelos de análisis factorial confirmatorio sólo contienen el modelo de medida y las relaciones entre las variables latentes sólo pueden ser de tipo correlacional. En segundo lugar, los modelos de análisis de rutas no contienen variables latentes; en su lugar, las variables observables son equiparadas con las variables latentes; consecuentemente, sólo existe el modelo de relaciones estructurales. Como contrapartida, los errores de medición y los errores de predicción se confunden en un único término común.

TIPOS DE VARIABLES

En un modelo estructural se distinguen distintos tipos de variables según sea su papel y según sea su medición.

- ✔ Variable observada o indicador. Variables que se mide a los sujetos. Por ejemplo, las preguntas de un cuestionario.
- ✔ Variable latente. Característica que se desearía medir pero que no se puede observar y que está libre de error de medición. Por ejemplo, una dimensión de un cuestionario o un factor en un análisis factorial exploratorio.
- ✔ Variable error. Representa tanto los errores asociados a la medición de una variable como el conjunto de variables que no han sido contempladas en el modelo y que pueden afectar a la medición de una variable observada. Se considera que son variables de tipo latente por no ser observables directamente. El error asociado a la variable dependiente representa el error de predicción.
- ✔ Variable de agrupación. Variable categóricas que representa la pertenencia a las distintas subpoblaciones que se desea comparar. Cada código representa una subpoblación.
- ✔ Variable exógena. Variable que afecta a otra variable y que no recibe efecto de ninguna variable. Las variables independientes de un modelo de regresión son exógenas.
- ✔ Variable endógena. Variable que recibe efecto de otra variable. La variable dependiente de un modelo de regresión es endógena. Toda variable endógena debe ir acompañada de un error.

LOS DIAGRAMAS ESTRUCTURALES: CONVENCIONES Y DEFINICIONES

Para representar un modelo causal y las relaciones que se desea incluir en él se acostumbra a utilizar diagramas similares a los diagramas de flujo. Estos diagramas se denominan *diagramas causales*, *gráfico de rutas* o *diagramas estructurales*. El diagrama estructural de un modelo es su representación gráfica y es de gran ayuda a

la hora de especificar el modelo y los parámetros contenidos en él. De hecho, los programas actuales permiten realizar la definición del modelo en su totalidad al representarlo en el interfaz gráfico. A partir del diagrama estructural el propio programa deriva las ecuaciones del modelo e informa de las restricciones necesarias para que esté completamente identificado. Los diagramas estructurales siguen unas convenciones particulares que es necesario conocer para poder derivar las ecuaciones correspondientes.

- ✔ Las variables observables se representan encerradas en rectángulos.
- ✔ Las variables no observables (latentes) se representan encerradas en óvalos o círculos.
- ✔ Los errores (sean de medición o de predicción) se representan sin rectángulos ni círculos (aunque algunos programas las dibujan como variables latentes).
- ✔ Las relaciones bidireccionales (correlaciones y covarianzas) se representan como vectores curvos con una flecha en cada extremo.
- ✔ Cualquier efecto estructural se representa como una flecha recta, cuyo origen es la variable predictora y cuyo final, donde se encuentra la punta de la flecha, es la variable dependiente.
- ✔ Los parámetros del modelo se representan sobre la flecha correspondiente.
- ✔ Cualquier variable que reciba efecto de otras variables del modelo deberá incluir también un término error.
- ✔ Aunque no es necesario que el usuario lo especifique, los programas suelen incluir, junto a cada variable, su varianza y, si se trata de una variable dependiente, su correspondiente proporción de varianza explicada.

Los diagramas estructurales también sirven para especificar adecuadamente el modelo de cara a la estimación con un programa estadístico. Las restricciones se hacen de manera gráfica o imponiendo valores sobre el propio gráfico. Además, los programas estadísticos permiten comprobar el modelo especificado a partir del gráfico que genera el programa. Esto ayuda a no olvidar parámetros fundamentales en la definición del modelo, evitando que el usuario tenga que escribir de forma explícita las ecuaciones del modelo y confiar en que las ecuaciones sean las correctas.

Revisemos el modelo planteado anteriormente como ejemplo pero, esta vez, definido con mayor complejidad. La figura 2 muestra una nueva versión del modelo que contiene seis variables latentes: autoestima, autoeficacia, apoyo social, estrés, cansancio emocional y síntomas psicosomáticos. Las tres primeras variables latentes son exógenas (porque no reciben efecto directo de otra variable), y las tres últimas variables latentes son endógenas, porque reciben efecto de otras variables. Las tres variables endógenas cuentan con un término que representa su error de predicción (e_e, ece y esx).

Cada variable latente endógena está medida mediante tres variables observables que se denominan indicadores. La variable latente síntomas psicosomáticos se mide a los sujetos mediante tres escalas llamadas INDI1SX, INDI2SX e INDI3SX. El modelo asume que una persona con muchos síntomas psicosomáticos puntuará alto en los tres indicadores y una persona con pocos síntomas psicosomáticos puntuará bajo. Los indicadores son ob-

FIGURA 2
MODELO ESTRUCTURAL COMPLETO DE PREDICCIÓN DE LOS SÍNTOMAS PSICOSOMÁTICOS

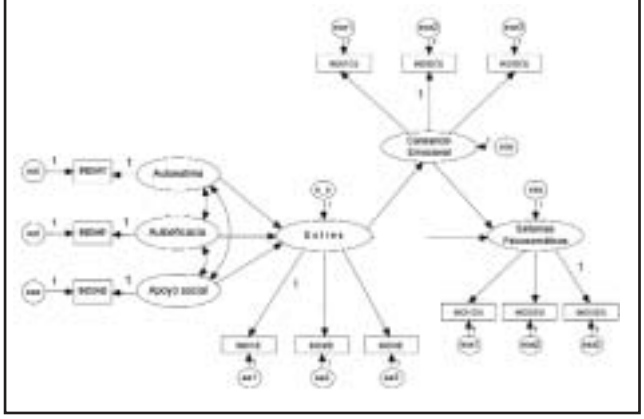
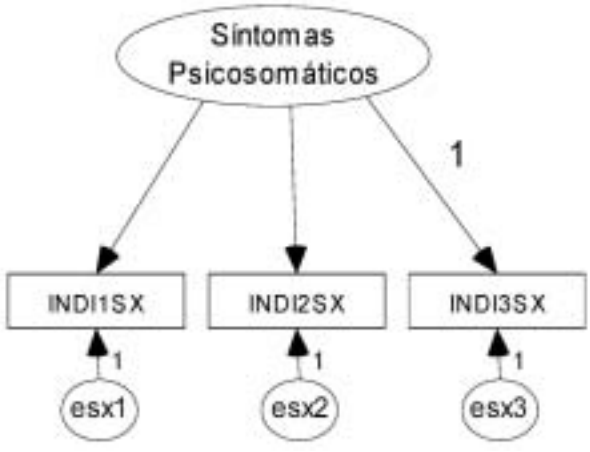


FIGURA 3
MODELO DE MEDIDA DE LA VARIABLE SÍNTOMAS PSICOSOMÁTICOS



servables pero no son medidas perfectas de su variable latente. Por ese motivo, cada indicador tiene asociado un error de medida. El error de medida del indicador INDI1SX es la variable no observable esx1. La figura 3 representa el modelo de medida de la variable latente síntomas psicosomáticos. En el caso de las variables latentes exógenas, cada constructo se encuentra medido por un solo indicador y por ese motivo se puede simplificar esa parte del modelo identificando la variable latente con su indicador, como se indica en el modelo final estimado de la figura 5.

La figura 4 representa el modelo de relaciones estructurales. Este modelo sólo contiene las variables latentes. En él es fácil apreciar que las variables exógenas pueden correlacionar entre sí (cosa que no sería posible en un modelo de regresión ordinario) y que cada variable endógena tiene asociado un error de predicción que explica parte de su variabilidad (este error no está asociado a los errores de medida, que están recogidos en el modelo de medida).

La figura 5 representa el modelo final estimado, una vez simplificado con respecto a las variables exógenas.

FIGURA 4
MODELO DE RELACIONES ESTRUCTURALES

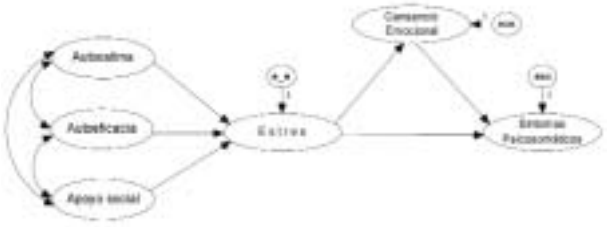
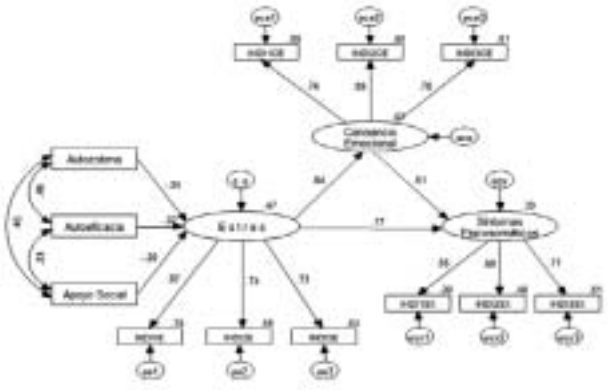


FIGURA 5
MODELO FINAL ESTIMADO



En la parte izquierda se encuentran las tres variables exógenas utilizadas para predecir el nivel de estrés. Las tres variables son observables y correlacionan entre sí (son multicolineales). El efecto negativo que tienen sobre el estrés indica que un menor nivel de autoestima, autoeficacia y apoyo social permite predecir un mayor nivel de estrés. (En el gráfico no se indica, pero todos los pesos de regresión son significativamente distintos de cero). La combinación de los tres predictores permite explicar el 47% de la varianza del estrés (libre de error de medición), lo que se indica numéricamente sobre la variable latente. La proporción de varianza del estrés explicada por sus predictores es inversamente proporcional a la varianza de su error de predicción y por eso no es necesario indicar su valor, pero sí se representa la variable de error correspondiente (e_e). Cada variable latente endógena se encuentra medida por tres indicadores. Cada flecha que parte de una variable latente hacia su indicador se interpreta igual que la saturación en un análisis factorial y (en la solución estandarizada) se corresponde con la correlación del indicador con la variable latente que intenta medir. El valor numérico representado junto al recuadro de una variable observada es la proporción de varianza compartida por el indicador y la variable latente (similar a la comunalidad) y que no es atribuible al error de medición. En la parte central del modelo se encuentran los efectos de unas variables latentes sobre las otras. Se aprecia que el estrés tiene mayor efecto directo sobre el cansancio emocional que sobre los síntomas psicosomáticos. A su vez, el efecto que reciben los síntomas psicosomáticos del estrés es menor que el que reciben del cansancio emocional. En la figura no se representa el efecto total del estrés sobre los síntomas psicosomáticos (0,50) que sería la suma del efecto directo (0,17) y el indirecto (0.64*0.51=0,33) a través del cansancio emocional.

Comparando el modelo completo con el modelo de rutas de la figura 1 podemos constatar que los efectos se han incrementado en algunos casos de manera sustancial y que, además, se ha incrementado la proporción de varianza explicada de las variables endógenas. También se aprecia que no todos los indicadores son igual de precisos. Por último, es esperable que este modelo equivalente, siendo estimable, obtenga peores valores de ajuste que el modelo de rutas por el mero hecho de contener un mayor número de variables (lo que afecta a los grados de libertad del modelo y a los estadísticos de bondad de ajuste).

Al igual que existe un conjunto convenciones para representar los modelos de manera gráfica, también exis-

ten convenciones para nombrar cada elemento de un modelo, ya sean variables o parámetros, en su notación matemática. No entraremos aquí a explicar esta notación, pero sí es bueno saber que se suelen utilizar letras griegas (ver Ruiz, 2000; Hayduk, 1987).

PASOS EN LA ELABORACIÓN DE UN MODELO

La estimación de un modelo comienza con la formulación de la teoría que lo sustenta. Dicha teoría debe estar formulada de manera que se pueda poner a prueba con datos reales. En concreto, debe contener las variables que se consideran importantes y que deben medirse a los sujetos. El modelo teórico debe especificar las relaciones que se espera encontrar entre las variables (correlaciones, efectos directos, efectos indirectos, bucles). Si una variable no es directamente observable, deben mencionarse los indicadores que permiten medirla. Lo normal es formular el modelo en formato gráfico; a partir de ahí es fácil identificar las ecuaciones y los parámetros.

Una vez formulado el modelo, cada parámetro debe estar correctamente identificado y ser derivable de la información contenida en la matriz de varianzas-covarianzas. Existen estrategias para conseguir que todos los parámetros estén identificados, como por ejemplo, utilizar al menos tres indicadores por variable latente e igualar la métrica de cada variable latente con uno de sus indicadores (esto se consigue fijando arbitrariamente al valor 1 el peso de uno de los indicadores). Aun así, puede suceder que el modelo no esté completamente identificado, lo que querrá decir que se está intentando estimar más parámetros que el número de piezas de información contenidas en la matriz de varianzas-covarianzas. En ese caso habrá que imponer más restricciones al modelo (fijando el valor de algún parámetro) y volver a formularlo.

Por otra parte, una vez seleccionadas las variables que formarán parte del modelo, hay que decidir cómo se medirán las variables observables. Estas mediciones (generalmente obtenidas mediante escalas o cuestionarios) permitirán obtener las varianzas y las covarianzas en las que se basa la estimación de los parámetros de un modelo correctamente formulado e identificado (asumimos que estamos trabajando con una muestra representativa de la población que se desea estudiar y de tamaño suficientemente grande).

Una vez estimados los parámetros del modelo se procede, en primer lugar, a valorar su ajuste. Si las estimaciones obtenidas no reproducen correctamente los datos observados, habrá que rechazar el modelo y con ello la teoría que lo soportaba, pudiendo pasar a corregir el modelo haciendo supuestos teóricos adicionales. En segundo lugar se procede a hacer una valoración técnica de los valores estimados para los parámetros. Su magnitud debe ser la adecuada, los efectos deben ser significativamente distintos de cero, no deben obtenerse estimaciones impropias (como varianzas negativas), etc. Puede ocurrir que alguna de las estimaciones tenga un valor próximo a cero; cuando ocurre esto es recomendable simplificar el modelo eliminando el correspondiente efecto. Por último, el modelo debe interpretarse en todas sus partes. Si el modelo ha sido aceptado como una buena explicación de los datos será interesante validarlo con otras muestras y, muy posiblemente, utilizarlo como explicación de teorías de mayor complejidad que se desee contrastar. El proceso expuesto se resume gráficamente en la figura 6.

TIPOS DE RELACIONES

En las técnicas multivariantes estamos acostumbrados a estudiar la relación simultánea de diversas variables entre sí. En estas técnicas las relaciones entre variables dependientes e independientes son todas del mismo nivel o del mismo tipo. En un modelo de ecuaciones estructurales podemos distinguir distintos tipos de relaciones. En-

FIGURA 6
PROCESO DE ELABORACIÓN Y ESTIMACIÓN DE UN MODELO

tender estos distintos tipos de relaciones puede ser de gran ayuda a la hora de formular los modelos a partir de las verbalizaciones en lenguaje común. A continuación vamos a discutir estos tipos de relaciones, siguiendo el esquema propuesto por Saris y Stronkhorst (1984).

COVARIACIÓN Vs CAUSALIDAD

Decimos que dos fenómenos *covarían*, o que están correlacionados, cuando al observar una mayor cantidad de uno de los fenómenos también se observa una mayor cantidad del otro (o menor si la relación es negativa). De igual forma, a niveles bajos del primer fenómeno se asocian niveles bajos del segundo. Así, por ejemplo, cuando decimos que la aptitud y el rendimiento correlacionan entre sí, esperamos que los sujetos con un mayor nivel de aptitud manifiesten un mejor rendimiento y viceversa. Sin embargo, ya hemos enfatizado que covariación y causalidad no son la misma cosa. Cuando se observa una alta relación (covariación) entre dos variables, no debemos interpretarla como una relación causal entre ambas. Pueden existir otras variables que no hemos observado y que potencien o atenúen esta relación. Por ejemplo, es posible que la motivación y el rendimiento estén relacionados y que esa relación esté condicionando la relación de la aptitud con el rendimiento (potenciándola o atenuándola). Un ejemplo tal vez más claro es el propuesto por Saris. Si recolectamos datos sobre el número de vehículos y el número de aparatos telefónicos en distintas poblaciones, es seguro que encontraremos una covariación entre ambas variables. Pero no por ello pensamos que un mayor número de vehículos es el causante de que haya un mayor número de aparatos telefónicos.

Otro nivel de análisis es la causalidad. Si recogemos información sobre el número de fumadores en una habitación y la cantidad de humo existente en la habitación, observaremos que existe una alta covariación entre ambas variables. Parece razonable dar un paso más en la interpretación de este resultado y argumentar, conceptualmente, que la cantidad de fumadores *causa* la cantidad de humo y que los cambios en la cantidad de fumadores causarán un cambio en la cantidad de humo.

El cambio de perspectiva desde la covariación observada a la causalidad atribuida a dos variables lo lleva a cabo el investigador, que es quien hipotetiza la causalidad. Es una buena costumbre que las verbalizaciones, o enunciados, sean explícitos respecto al tipo de relación que deseamos probar entre dos variables.

Los ejemplos que hemos expuesto en este apartado pueden representarse mediante los gráficos que hemos desarrollado hasta aquí.

Si estamos estudiando la correlación entre aptitud y rendimiento deberemos representarla como una flecha curva entre ambas variables.

Figura 7 Relación de covariación

Por el contrario, la relación causal entre el número de fumadores y la cantidad de humo la representaremos como un vector que apunte de la causa hacia el efecto.

Figura 8: Relación de tipo causal

RELACIÓN ESPURIA

En una relación causal básica o una relación de covariación hay involucradas dos variables. En una relación espuria la relación comprende al menos tres variables. Una relación espuria se refiere a la existencia de covariación entre dos variables que es debida, total o parcialmente, a la relación común de ambas variables con una tercera. Esta es la razón por la cual la covariación entre dos variables puede ser muy elevada y, sin embargo, ser nula su relación causal. Un ejemplo típico de relación espuria es la que se da entre estatura e inteligencia en preescolares. Si medimos ambas variables en niños de preescolar es muy posible que encontremos una alta relación entre ellas; sin embargo, a nadie se le ocurre pensar que la estatura causa la inteligencia. Existe una tercera variable, el desarrollo del niño (la edad), que es causa de ambas variables y que hace que se observe esa relación. Gráficamente se puede representar de la siguiente forma:

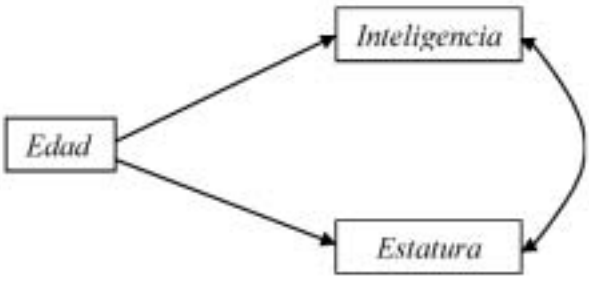


Figura 9: Relación espuria

Para estudiar la presencia de este fenómeno se utiliza el coeficiente de correlación parcial, que mide la relación entre dos variables tras eliminar el efecto de una tercera (también puede eliminarse el efecto de más de una variable). En nuestro ejemplo, la correlación entre las tres variables será alta y positiva, mientras que la correlación parcial entre la inteligencia y la estatura (eliminando el efecto de la edad) será prácticamente nula.

En general, podemos decir que la relación causal entre dos variables implica que ambas variables covarían, permaneciendo constantes el resto de las variables. Pero lo contrario no es cierto: la covariación entre dos variables no implica necesariamente que exista una relación causal entre ambas; la relación puede ser espuria, falsa, ficticia (ver Pardo, Ruiz y San Martín, 2009, págs. 356-357).

RELACIÓN CAUSAL DIRECTA E INDIRECTA

Hasta ahora sólo hemos mencionado relaciones causales directas. Una relación causal indirecta implica la presencia de tres variables. Existe una relación indirecta entre dos variables cuando una tercera variable modula o mediatiza el efecto entre ambas. Es decir, cuando el efecto entre la primera y la segunda pasa a través de la tercera. A las variables que median en una relación indirecta se las denomina también variables moduladoras.

Consideremos la relación entre la aptitud, el rendimiento y la motivación. Podemos pensar en el nivel de motivación como una variable que modula la relación entre la aptitud y el rendimiento. Esta relación puede representarse gráficamente como:

Figura 10: Relación causal indirecta

El modelo de la figura propone que existe un efecto directo de la aptitud sobre la motivación y de la motivación sobre el rendimiento. Además, existe un efecto indirecto entre la aptitud y el rendimiento. El efecto indirecto de la variable aptitud sobre el rendimiento puede ser potenciado (o atenuado) por la variable moduladora motivación.

La existencia de un efecto indirecto entre dos variables no anula la posibilidad de que también exista un efecto directo entre ellas. Así, las relaciones propuestas en la figura 10 pueden hacerse más complejas de la siguiente forma:

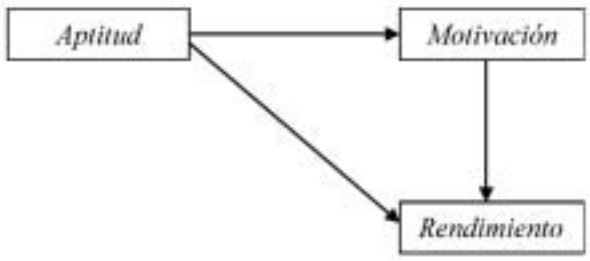


Figura 11: Relaciones directa e indirecta

Una vez más, es el investigador quién debe explicitar el tipo de relaciones que su teoría es capaz de justificar.

RELACIÓN CAUSAL RECÍPROCA

La relación causal entre dos variables puede ser recíproca o unidireccional. Cuando la relación es recíproca (bidireccional) la variable causa es a su vez efecto de la otra. Este tipo de relaciones se representa como dos flechas separadas orientadas en sentidos contrarios. Una relación recíproca es en definitiva un bucle de retroalimentación entre dos variables. La relación causal recíproca puede ser directa o indirecta, implicando a otras variables antes de cerrase el bucle.

La relación entre la Ansiedad y el Rendimiento puede representarse como un bucle recíproco: cuanto mayor es la ansiedad, peor es el rendimiento; y cuanto peor es el rendimiento, mayor es la ansiedad.

Figura 12: Relación causal recíproca

EFECTOS TOTALES

Hemos visto que cada tipo de relación causal se representa mediante un tipo de efecto. Existe un último tipo de efecto (o relación) que no hemos mencionado; se trata de los efectos *no analizados*. En la representación gráfica son las flechas que podrían estar representadas y que no lo están. Estas ausencias pueden obedecer a dos motivos. Por un lado, puede ocurrir que se hayan dejado fuera del modelo variables importantes para explicar la covariación presente en los datos (error de especificación). Por otro, puede ser debido a que se asume que el resto de las variables no consideradas en el modelo se compensan entre sí, incorporándose su efecto en los términos de error del modelo. A la suma de los efectos espurios más los efectos no analizados se les denomina *efectos no causales*. Una vez que el modelo está defini-

do, los efectos espurios aparecen cuando las variables endógenas están correlacionadas más allá de los efectos estimados (apareciendo covarianzas entre los errores de predicción). Los efectos no analizados aparecen cuando las variables observables están correlacionadas más allá de lo que el modelo predice (apareciendo covarianzas entre los errores de medición).

Como sea que una variable endógena puede recibir un efecto directo de otra variable y también un efecto indirecto de esa misma variable modulado por otras terceras variables, se acostumbra a sumar ambos tipos de efectos dando lugar al *efecto total*.

EL CONCEPTO DE "AJUSTE"

Para entender la fundamentación de los modelos de ecuaciones estructurales, es necesario reorientar nuestro conocimiento de lo que significa el concepto de *ajuste* de un modelo. En regresión lineal, cuando hablamos de las estimaciones de los parámetros, escogemos aquellas estimaciones que mejor ajustan el modelo a los datos, en el sentido de que minimizan los errores de predicción cometidos con el modelo para el conjunto de sujetos de la muestra (en el método de mínimos cuadrados). Por el contrario, en los modelos de ecuaciones estructurales, lo que se pretende ajustar son las covarianzas entre las variables, en vez de buscar el ajuste a los datos. En lugar de minimizar la diferencia entre los valores pronosticados y los observados a nivel individual, se minimiza la diferencia entre las covarianzas observadas en la muestra y las covarianzas pronosticadas por el modelo estructural. Este es el motivo por el que a estos modelos también se les llama de estructura de covarianza (*covariance structure models*; Long, 1983). Por tanto, los residuos del modelo son la diferencia entre las covarianzas observadas y las covarianzas reproducidas (pronosticadas) por el modelo estructural teórico.

El ajuste de un modelo se puede expresar en una hipótesis fundamental, que propone que, *si el modelo es correcto* y conociéramos los parámetros del modelo estructural, la matriz de covarianzas poblacional podría ser reproducida exactamente a partir de la combinación de los parámetros del modelo. Esta idea de ajuste se resume en la siguiente ecuación

$$H_0: \ \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) \qquad (1)$$

donde Σ es la matriz de varianzas-covarianzas poblacional entre las variables observables, θ es un vector que contiene los parámetros del modelo y $\Sigma(\theta)$ es la matriz de varianzas-covarianzas derivada como una función de los parámetros contenidos en el vector θ.

Veamos el significado y extensión de esta hipótesis con un ejemplo (Bollen, 1989). Consideremos el modelo que muestra la Figura 13.

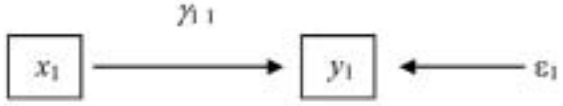


Figura 13: Modelo de regresión simple

La ecuación de regresión que lo define es la siguiente (se han eliminado los subíndices)

$$y = \gamma x + \varepsilon \qquad (2)$$

Donde γ es el coeficiente de regresión y ε la variable que representa el término error, que se asume que es independiente de x y cuyo valor esperado es cero. La matriz de varianzas-covarianzas entre las variables observadas x e y es

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathrm{VAR}(y) & \mathrm{COV}(x,y) \\ \mathrm{COV}(x,y) & \mathrm{VAR}(x) \end{pmatrix} \qquad (3)$$

Esta es la matriz que obtenemos directamente al analizar descriptivamente los datos y representa las relaciones existentes entre las variables en la muestra. Ahora bien, la variable dependiente y es función de las variables x y ε, y del parámetro γ. Podemos volver a escribir los elementos de la matriz Σ en función de la ecuación (2). Operando, es relativamente fácil demostrar que la varianza de la variable dependiente es función del parámetro γ y de la varianza de los errores:

$$\mathrm{VAR}(y) = \gamma^2 \ \mathrm{VAR}(x) + \ \mathrm{VAR}(\varepsilon) \qquad (4)$$

También es posible demostrar que la covarianza entre x e y es función del parámetro γ y de la varianza de la variable predictora:

$$\mathrm{COV}(x,y) = \ \gamma \ \mathrm{VAR}(x) \qquad (5)$$

Sustituyendo en la ecuación (3) las expresiones derivadas escritas en función de los parámetros del modelo llegamos a la matriz de varianzas- covarianzas poblacional reproducida:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \gamma^2 \mathrm{VAR}(x) + \mathrm{VAR}(\varepsilon) & \gamma \mathrm{VAR}(x) \\ \gamma \mathrm{VAR}(x) & \mathrm{VAR}(x) \end{pmatrix} \qquad (6)$$

A esta matriz también se le llama matriz de varianzas-covarianzas *implícita*. Podemos sustituir ahora en la ecuación (1) y volver a expresar la hipótesis básica com∂

$$H_0: \Sigma = \begin{pmatrix} VAR(y) & COV(x,y) \\ COV(x,y) & VAR(x) \end{pmatrix} = \begin{pmatrix} \gamma^2 VAR(x) + VAR(\varepsilon) & \gamma VAR(x) \\ \gamma VAR(x) & VAR(x) \end{pmatrix} = \Sigma(\theta) \quad (7)$$

En esta igualdad, los elementos de la parte derecha y los de la parte izquierda se corresponden uno a uno, dadas las especificaciones del modelo que hemos propuesto. Si el modelo es el correcto y conociéramos los valores de los parámetros de la parte derecha de la igualdad, no sería difícil comprobar la igualdad de los términos. El objetivo de la estimación es obtener los valores de los parámetros (en este caso el coeficiente de regresión y la varianza de los errores) que permiten mantener esta igualdad con los datos muestrales.

Para poder estimar los parámetros del modelo ha sido necesario esperar al desarrollo de programas informáticos. En esta breve aproximación a los modelos de ecuaciones estructurales basta con saber que las estimaciones se realizan intentando maximizar el ajuste del modelo. Para ello se utiliza alguna medida que resuma la magnitud de las diferencias entre las varianzas y covarianzas observadas (parte izquierda de la igualdad) y las reproducidas (parte derecha de la igualdad), y se intenta minimizar dichas diferencias.

LOS ESTADÍSTICOS DE BONDAD DE AJUSTE

Una vez que se ha estimado un modelo es necesario evaluar su calidad. Para ello se utilizan los estadísticos de bondad de ajuste. Existen tres tipos de estadísticos de bondad de ajuste: los de ajuste absoluto (valoran los residuos), los de ajuste relativo (comparan el ajuste respecto a otro modelo de peor ajuste) y los de ajuste parsimonioso (valoran el ajuste respecto al número de parámetros utilizado). Ninguno de ellos aporta toda la información necesaria para valorar el modelo y habitualmente se utiliza un conjunto de ellos del que se informa simultáneamente (ver Schreiber y otros, 2006).

En la siguiente tabla se enumeran los más utilizados, junto con su abreviatura y el valor de referencia que debe alcanzar cada uno para indicar un buen ajuste. El estadístico chi-cuadrado es conceptualmente el más atractivo; permite contrastar la hipótesis nula de que todos los errores del modelo son nulos, por lo que interesa mantener dicha hipótesis con la muestra utilizada. Sin embargo, es muy sensible al tamaño muestral: con muestras grandes (mayores de 100 ó 200 casos) es relativamente fácil rechazar la hipótesis nula cuando el modelo de hecho consigue un buen ajuste. Por este motivo, además de valorar su significación estadística, suele compararse con sus grados de libertad. Siempre se informa de este estadístico.

PROBLEMAS TÍPICOS

Es necesario mencionar varios problemas típicos que se suelen encontrar en los modelos publicados, algunas limitaciones que debemos tener en cuenta y las precauciones que debemos tomar al utilizarlos.

En la definición de un modelo no deben excluirse variables importantes desde el punto de vista teórico. En primer lugar, debe hacerse un esfuerzo por medir todas las variables pertinentes. En segundo lugar, deben cuestionarse los modelos en los que las variables conceptualmente centrales carezcan de efecto significativo.

El hecho de que un modelo obtenga buen ajuste con una muestra no excluye que puedan existir otros modelos tentativos que también puedan ajustarse bien a los datos. Siempre es interesante contrastar otros modelos que también puedan estar soportados por la teoría (o por teorías rivales).

En ocasiones se publican los modelos conteniendo tanto los efectos correspondientes a parámetros distintos de cero como efectos que tras la estimación se pueden considerar nulos. Aunque el espacio requerido para dar explicaciones sea mayor, debe informarse tanto del modelo teórico con todos los parámetros y variables propuestas como del modelo final que sólo contenga los parámetros distintos de cero y las variables con efecto estadístico.

TABLA 1
ESTADÍSTICOS DE BONDAD DE AJUSTE Y CRITERIOS DE REFERENCIA

Estadístico	Abreviatura	Criterio
Ajuste absoluto		
Chi-cuadrado	χ^2	Significación > 0,05
Razón Chi-cuadrado / grados de libertad	χ^2/gl	Menor que 3
Ajuste comparativo		
Índice de bondad de ajuste comparativo	CFI	≥ 0,95
Índice de Tucker-Lewis	TLI	≥ 0,95
Índice de ajuste normalizado	NFI	≥ 0,95
Ajuste parsimonioso	NFI	
Corregido por parsimonia	PNFI	Próximo a 1
Otros		
Índice de bondad de ajuste	GFI	≥ 0,95
Índice de bondad de ajuste corregido	AGFI	≥ 0,95
Raíz del residuo cuadrático promedio	RMR	Próximo a cero
Raíz del residuo cuadrático promedio de aproximación	RMSEA	< 0,08

Es sabido que los estadísticos de bondad de ajuste se deterioran rápidamente con el aumento del tamaño muestral y muchos investigadores informan de muestras pequeñas para no deteriorar los valores de ajuste. Por este motivo deben cuestionarse los modelos estimados con muestras pequeñas o poco representativas. Se acostumbra a exigir tamaños muestrales superiores a los 100 sujetos y los tamaños superiores a los 200 sujetos son una buena garantía.

Estos modelos admiten pocas variables (10-20). Cuanto mayor es el número de variables, más difícil resulta reproducir correctamente las covarianzas observadas. Además, cuanto mayor sea el número de variables mayor debe ser también el tamaño muestral (se recomienda una tasa superior a los 10 sujetos por variable observada).

Muchos estudios en los que se utilizan estos modelos abusan del ajuste y reajuste de las posibles relaciones teóricas, incluyendo y excluyendo efectos y variables de manera tentativa. Para ello se utilizan los valores de significación y los índices de modificación de los parámetros individuales (tanto los de los efectos analizados como los de los efectos excluidos) y que informan de los problemas de ajuste existentes en los datos. Estos modelos sobre-manipulados suelen ser muy inestables y pierden sus buenas propiedades de ajuste cuando se replican con otras muestras. Por desgracia, los estudios de replicación son escasos, por lo que es recomendable mantener un cierto escepticismo cuando en un estudio no se informe detalladamente de las manipulaciones que hayan podido sufrir los datos y el modelo.

No se deben utilizar variables categóricas ya que, idealmente, todas las variables deberían ser cuantitativas continuas para justificar el uso de los estadísticos varianza y covarianza. Como hemos visto, es fundamental que la estimación muestral de las varianzas y covarianzas entre las variables observadas sea precisa para que el proceso de estimación de los parámetros del modelo sea exitoso. Sin embargo, es muy frecuente que utilicemos preguntas en formato ordinal tipo Likert para medir a los sujetos, por la facilidad que supone responder en ellas. En esos casos será conveniente agrupar las preguntas individuales para formar escalas con una métrica más continua (ver Finney y DiStefano, 2006).

CONSIDERACIONES FINALES

A pesar de las limitaciones mencionadas, los modelos de ecuaciones estructurales son una herramienta muy potente para formalizar de manera explícita teorías relativamente complejas, permite contrastarlas y posibilita incluir relaciones complejas o jerárquicas entre múltiples variables.

También permiten extender algunos modelos tradicionales al incluir, por ejemplo, errores de medición en los modelos de análisis factorial, o al estimar directamente las saturaciones y las correlaciones entre los factores (sin recurrir a la rotación) o al incluir pruebas de significación individuales para las saturaciones estimadas.

Además, en ellos se pueden separar los errores de medida de los errores de predicción, atenuando el efecto de los errores de medición sobre la valoración de la capacidad predictiva del modelo.

Estos modelos, junto con los modelos de regresión canónica, son los únicos que permiten analizar problemas en los que se dispone de más de una variable dependiente y analizarlas de forma simultánea.

Aunque la estimación de estos modelos se ha simplificado mucho con los programas de estimación que cuentan con un interfaz gráfico es importante tener en cuenta que su uso es laborioso. Si bien es cierto que son una ayuda inestimable para afrontar el reto del desarrollo de teorías explicativas del comportamiento humano.

RECURSOS ADICIONALES

Aquellos que quieran profundizar más en estos modelos manteniéndose a un nivel básico pueden consultar los manuales de Byrne (1994, 1998, 2001, 2006) y los que quieran una introducción aún más elemental pueden consultar el libro de Saris y Stronkhorst (1984) y las breves monografías de Long (1983a, 1983b, 1990). Una buena exposición de cómo desarrollar e interpretar estos modelos son los tres capítulos finales del manual de Hair y otros (2006), es muy práctico, aunque apenas contiene formulación y carece de demostraciones. El manual de Bollen (1989) es excelente y muy completo, pero requiere un buen nivel de conocimientos previos en estadística.

También son muy recomendables los manuales de los programas de estimación más utilizados: el AMOS (Arbuckle, 1997), el LISREL (Jöreskog y Sörborn, 1986; SPSS, 1990, 1993), el EQS (Bentler, 1985) y el CALIS, perteneciente a SAS (Hatcher, 2003).

Se encuentran disponibles dos programas de estimación de modelos de uso gratuito HYBALL (http://web.psych.ualberta.ca/~rozeboom/) y TETRAD (http://www.phil.cmu.edu/projects/tetrad/).

REFERENCIAS

Arbuckle, J. L. (1997). *Amos Users' Guide. Version 3.6*. Chicago: SmallWaters Corporation.

Bagozzi, R. O. y Yi, Y. (1994). Advanced Topics in Structural Equation Models. In: R. P. Bagozzi (Ed.). *Advanced Methods of Marketing Research*. Cambridge: Blackell Publishers.

Bentler, P. M. (1985). *Theory and implementation of EQS: A structural equations program*. Los Angeles: BMDP Statistical Software.

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

Bollen, K. A., Curran, P. J. (2006). *Latent Curve Models. A Structural Equation Perspective*. Wiley.

Boudon, R. (1965). A method of linear causal analysis: Dependence analysis. *American Sociological Review*, 30: 365-373.

Byrne, B. M. (1994). *Structural Equation Modeling with EQS and EQS/WINDOWS: Basic Concepts, Applications, and Programming.* SAGE Publications.

Byrne, B. M. (2001). *Structural Equation Modeling with AMOS. Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates.

Byrne, B. M. (2006). *Structural Equation Modeling With Eqs: Basic Concepts, Applications, and Programming (Multivariate Applications)*. SAGE Publications.

Byrne, B. M. (1998). Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming. Mahwah: Lawrence Erlbaum Associates, Inc.

Duncan, O. D. (1966) Path analysis: Sociological examples. *American Journal of Sociology*, 72: 1-12.

Ferrando, P.J. y Anguiano, C. (2010). El análisis factorial como técnica de investigación en Psicología. *Papeles del Psicólogo*, 31(1), 18-33.

Finney, A.J. y DiStefano C. (2006). Non-Normal and Categorical Data in Structural Equation Modeling. In: G. R. Hancock y R. O. Mueller (Eds). Structural Equation Modeling: A Second Course, 269-314. Greenwich: Information Age Publishing

González, M.T. y Landero, R. (2008). Confirmación de un modelo explicativo del estrés y de los síntomas psicosomáticos mediante ecuaciones estructurales. *Revista Panamericana de Salud Pública, 23* (1), 7-18.

Hayduk, L. (1987). *Structural equation modeling with LISREL Essentials and Advances.* Baltimore: The Johns Hopkins University Press.

Hair, J. E., Black, W. C., Babin, B. J., Anderson, R. E. y Tatham R. L. (2006). *Multivariate Data Analysis* (6th Edition). Upper Saddle River: Pearson-Prentice Hall

Hatcher, L. (2003). A Step-by-Step Approach to using SAS for Factor Analysis and Structural Equation Modeling. Cary: SAS Institute Inc.

Jöreskog, K. G. (1973) A general method for estimating a linear structural equation system, pp. 85-112 in A. S. Goldberger and O. D. Duncan (eds.) Structural Equation Models in the Social Sciences. New York: Seminar.

Jöreskog, K. G. y Sörbom, D. (1986*). LISREL VI: Analysis of Linear Structural Relationships by Maximum Likelihood and Least Squares Methods*. Mooresville, IN: Scientific Software, Inc.

Long, J. S. (1983a). *Confirmatory Factor Analysis: A Preface to LISREL*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 007-033. Newbury Park, CA: Sage.

Long, J. S. (1983b) *Covariance Structure Models: An introduction to LISREL*. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-034. Beverly Hills and London: Sage.

Long, J. S. (1990). *Covariance Structure Models: An Introduction to LISREL*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 007-034. Newbury Park, CA: Sage.

Pardo A., Ruiz, M.A. y San Martín, R. (2009). *Análisis de datos en ciencias sociales y de la salud* (volumen I). Madrid: Síntesis.

Preacher, K. J., Rucker, D. D., y Hayes, A. F. (2007). Assessing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behavioral Research, 42*, 185-227

Ruiz, M.A. (2000). *Introducción a los modelos de ecuaciones estructurales*. Madrid, UNED.

Saris, W. E. y Stronkhorst, L. H. (1984). *Causal Modelling in Non-Experimental Research*. Amsterdam: Sociometric Research.

Schreider, J.B., Stage, F.K., King, J., Nora, A., Barlow, E.A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: a review. *The Journal of Education Research, 99* (6), 323-337.

Skrondal, A., Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling. Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman & Hall/CRC.

SPSS (1990). *SPSS® LISREL® 7 and PRELIS®: User's Guide and Reference*. Chicago, IL: SPSS.

SPSS (1993). *SPSS® LISREL® 7 and PRELIS®: User's Guide and Reference*. Chicago, IL: SPSS.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 46-56
http://www.cop.es/papeles

ESCALAMIENTO MULTIDIMENSIONAL: CONCEPTO Y APLICACIONES

MULTIDIMENSIONAL SCALING: CONCEPT AND APPLICATIONS

Constantino Arce, Cristina de Francisco e Iria Arce
Universidad de Santiago de Compostela

A través del presente artículo se ofrece una visión conceptual, a la vez que operativa, del concepto de escalamiento multidimensional. En la forma de presentación se busca, en primer lugar, que los psicólogos interesados comprendan lo que es el modelo de escalamiento multidimensional a través de varios ejemplos muy sencillos e intuitivos y, en segundo lugar, adquieran competencias que le permitan resolver distintos problemas de escalamiento multidimensional con el uso de software específico. Se pretende igualmente descargar la presentación de fórmulas y métodos matemáticos sin renunciar por ello al rigor metodológico que el tema requiere.
Palabras clave: *Escalamiento de objetos, Escalamiento de sujetos, Datos de proximidad, Datos de preferencia, Reducción de la dimensionalidad.*

The present article offers a conceptual, and at the same time operative, vision of the concept of multidimensional scaling. In the manner it is presented, the aim is, firstly, to help interested psychologists understand what the multidimensional scaling model is, using a number of simple, intuitive examples; and, secondly, for them to acquire the competence required to resolved different problems in multidimensional scaling through the use of specific software. The aim is also to download the presentation of mathematical formulae and method, without renouncing the methodological rigour that the subject demands.
Key words: Scaling, Proximity data, Preference data, Dimensionality reduction.

El escalamiento multidimensional, en su formulación más básica, pretende representar un conjunto de objetos en un espacio de baja dimensionalidad. La palabra objeto es muy genérica y se refiere, en realidad, a cualquier entidad que deseemos escalar. Otro término equivalente utilizado en Psicología es estímulo. El número de dimensiones, habitualmente reducido (dos, tres, cuatro), las decide el investigador por razones sustantivas, aunque también puede hacerse por criterios estadísticos. Los modelos y métodos de construcción de escalas unidimensionales, que fueron desarrollados en la primera mitad del siglo XX, entre los que cabe citar a Thurstone, Likert, Guttman o Coombs, constituyen los antecedentes de los modelos y métodos más modernos de escalamiento multidimensional y, en muchas ocasiones, pueden éstos últimos considerarse como generalizaciones de aquellos.

El primer autor en desarrollar un modelo y un método de escalamiento multidimensional ha sido Torgerson (1958). A su modelo se le conoce, hoy en día, con el nombre de modelo métrico clásico. La denominación de métrico tiene que ver con la escala de medida que se requiere, o asume, para los datos que es de intervalos, en la jerarquía de Stevens. Pocos años después Shepard (1962) y Kruskal (1964a, 1964b) han propuesto un modelo que permite un descenso en la escala de medida hasta el nivel ordinal. A este modelo se le denomina no-métrico clásico. Carroll y Chang (1970) lograron un avance significativo con la propuesta de un modelo que permite derivar, además del espacio de objetos, un espacio de sujetos sobre el que se representa el peso o ponderación que cada sujeto concede a cada una de las dimensiones del espacio de objetos. El modelo de Carroll y Chang, que se conoce con el nombre de modelo INDSCAL, tiene gran interés psicológico dado que permite o tiene en cuenta las diferencias individuales en la percepción del espacio de objetos. Existe un espacio de objetos común, compartido por todos los sujetos, pero permite las diferencias entre unos individuos y otros en la percepción de dicha configuración.

Existen programas informáticos específicos para cada uno de los modelos señalados anteriormente pero hoy en día es posible resolver problemas múltiples de escalamiento multidimensional con un único programa de ordenador como, por ejemplo PROXSCAL O ALSCAL, que tienen implementados numerosos modelos y forman ambos parte del paquete estadístico SPSS de uso universal.

Uno de los rasgos que más diferencia al escalamiento

Correspondencia: Constantino Arce, *Facultad de Psicología, Universidad de Santiago de Compostela, 15.782 Santiago de Compostela. España. E-mail: constantino.arce @usc.es*

multidimensional de otros modelos estadísticos de análisis de datos es la matriz de entrada. En Psicología, estamos habituados a utilizar una matriz de datos rectangular X con n sujetos en las filas y p variables en las columnas, donde un elemento x_{ij} representa la medida obtenida para un sujeto i en una variable j. En su forma más típica, la matriz de entrada para el escalamiento multidimensional es una matriz de datos cuadrada de orden p con una misma entidad representada en las filas y en las columnas: los objetos que intentamos representar en el espacio multidimensional. Un elemento en esta matriz ∅ representa la distancia o desemejanza entre dos objetos i y j. Lo que tenemos en la matriz es, en realidad, una matriz de distancias o desemejanzas entre todos los pares de objetos.

La diferencia entre distancia (concepto geométrico) y desemejanza (concepto psicológico) está en que el primero, al ser un concepto matemático, no contiene error; mientras el segundo, al ser un concepto psicológico, perceptivo o subjetivo, sí contiene error. Las desemejanzas son, en realidad, distancias que contienen error o distancias distorsionadas por los mecanismos perceptivos de los seres humanos. Los modelos y métodos de escalamiento multidimensional pueden resolver ambos tipos de problemas, con error y sin error en los datos de entrada. En Psicología es más habitual trabajar con datos que contienen error y los modelos de escalamiento multidimensional pueden tratar este problema.

DERIVACIÓN DE UNA CONFIGURACIÓN DE PUNTOS A PARTIR DE UNA MATRIZ DE DISTANCIAS

En la Tabla 1 se ofrece la matriz de distancias quilométricas entre 7 ciudades españolas: A Coruña, Bilbao, Barcelona, Cáceres, Madrid, Sevilla y Valencia.

Nos proponemos elaborar, a partir de dicha matriz, un mapa de España; es decir, obtener una representación espacial de las 7 ciudades sobre un plano, donde uno de los ejes será la dirección norte-sur y otro eje será la dirección este-oeste. Utilizamos, para ello, el procedimiento PROXSCAL, implementado en SPSS.

El resultado que nos ofrece es el que se puede observar en la Figura 1.

Dado que el mapa de España es conocido, podemos valorar subjetivamente el grado en que el mapa derivado por el programa se ajusta al mapa real. Podemos decir que el mapa conseguido es bastante bueno, aunque no perfecto. En la investigación en Psicología, es habitual trabajar con configuraciones que no tienen una contraparte objetiva conocida de antemano. Por eso, cuando el programa nos deriva una solución se vuelve muy importante tener un indicador o, incluso, varios--cuántos más mejor--, del grado en que la configuración derivada por el programa se ajusta a la ideal (desconocida). Todos los programas de escalamiento multidimensional ofrecen al usuario indicadores de ajuste para que pueda valorar lo "buena" que es la solución obtenida por el programa para su problema.

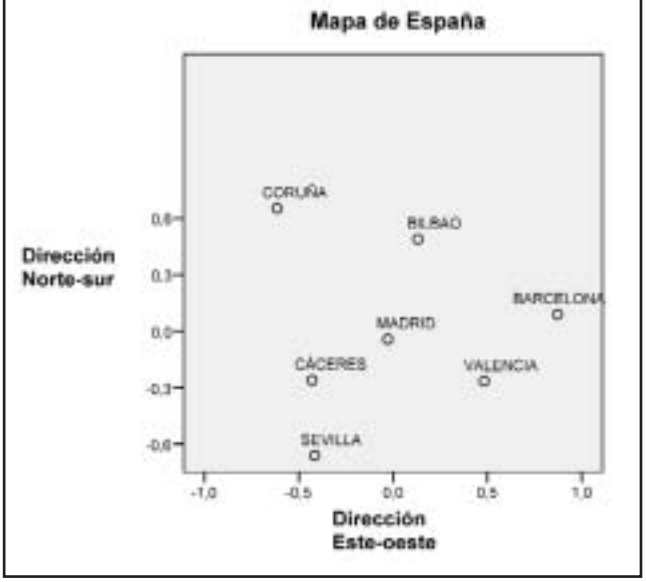
Los indicadores de bondad de ajuste ofrecidos por PROXSCAL para el mapa de España se ofrecen en la Tabla 2.

Hay dos tipos de indicadores. Aquellos para los que el cero representa un ajuste perfecto. De este primer tipo serían los indicadores Stress bruto normalizado, Stress-I, Stress-II y S-Stress. Y aquellos para los que el ajuste perfecto está representado por el 1. De este segundo tipo serían los dos últimos de la Tabla: Dispersión explicada

TABLA 1
DISTANCIAS QUILOMÉTRICAS ENTRE 7 CIUDADES ESPAÑOLAS

	A Coruña	Barcelona	Bilbao	Cáceres	Madrid	Sevilla	Valencia
A Coruña	0						
Barcelona	1050	0					
Bilbao	542	567	0				
Cáceres	617	895	591	0			
Madrid	586	600	379	294	0		
Sevilla	857	971	847	256	507	0	
Valencia	937	341	569	615	352	637	0

FIGURA 1
MAPA DE ESPAÑA DERIVADO POR PROXSCAL



(D.A.F.) y Coeficiente de congruencia de Tucker. Observando los valores de unos y otros siempre se llega a la misma conclusión: que el ajuste del modelo es bueno o muy bueno en este caso. Esto es así porque el grado de error en los datos (distancias) era muy pequeño. Las distancias utilizadas como entrada eran las distancias por carretera. Si utilizáramos las distancias lineales el ajuste sería perfecto. Los cuatro primeros índices de ajuste deberían ser iguales a 0 y los dos últimos iguales a 1.

PERCEPCIÓN DE LOS MEDIOS DE TRANSPORTE PÚBLICO

Arce (1993) se propuso obtener un mapa perceptivo de los medios de transporte utilizados por los ciudadanos de Santiago de Compostela. Para ello, elaboró una lista de todos los medios de transporte (públicos y privados) que podrían estar a su disposición en la ciudad, formó con ellos todos los pares posibles y pidió a una muestra de ciudadanos que juzgaran la desemejanza para cada par de medios de transporte.

TABLA 2
ÍNDICES DE AJUSTE OFRECIDOS POR PROXSCAL

Stress bruto normalizado	,00055
Stress-I	,02349
Stress-II	,06824
S-Stress	,00117
Dispersión explicada (D.A.F.)	,99945
Coeficiente de congruencia de Tucker	,99972

TABLA 3
LISTADO DE PARES DE MEDIOS DE TRANSPORTE, ORDENADOS POR ROTACIÓN ESTÁNDAR

1. Avión/tren 19. Autobús interurbano/moto
2. Avión/autobús interurbano 20. Autobús interurbano/ciclomotor
3. Avión/autobús urbano 21. Autobús interurbano/bicicleta
4. Avión/taxi 22. Autobús urbano/taxi
5. Avión/coche particular 23. Autobús urbano/coche particular
6. Avión/moto 24. Autobús urbano/moto
7. Avión/ciclomotor 25. Autobús urbano/ciclomotor
8. Avión/bicicleta 26. Autobús urbano/bicicleta
9. Tren/autobús interurbano 27. Taxi/coche particular
10. Tren/autobús urbano 28. Taxi/moto
11. Tren/taxi 29. Taxi/ciclomotor
12. Tren/coche particular 30. Taxi/bicicleta
13. Tren/moto 31. Coche particular/moto
14. Tren/ciclomotor 32. Coche particular/ciclomotor
15. Tren/bicicleta 33. Coche particular/bicicleta
16. Autobús interurbano/autobús urbano 34. Moto/ciclomotor
17. Autobús interurbano/taxi 35. Moto/bicicleta
18. Autobús interurbano/coche particular 36. Ciclomotor/bicicleta

Los medios de transporte estudiados fueron nueve: avión, tren, autobús interurbano, autobús urbano, taxi, coche particular, moto, ciclomotor y bicicleta. Con nueve objetos o estímulos (aquí, medios de transporte) se pueden formar 36 pares. Para averiguar el número de pares se utiliza la fórmula n(n-1)/2, donde n es el número de objetos o estímulos. Sustituyendo, en este caso, donde n = 9, nos queda 9(9-1)/2 = 36. En la Tabla 3 se ofrecen los 36 pares formados en el estudio, siguiendo un método, denominado rotación estándar, muy útil porque los datos (desemejanzas) ya quedan ordenados en la forma en que luego se van a introducir en la matriz de entrada. El método sigue la secuencia (1,2), (1,3) … (1,9), (2,3), (2,4) … (2,9), (3,4) (3,5) … (3,9) … (8,9).

Para la formación del número de pares, hemos asumido la simetría de los juicios de desemejanza, queriendo esto decir que para un par dado (p.e. avión/tren), se asume que el juicio emitido por un sujeto sería el mismo si el par se presenta en el orden avión/tren que en el orden tren/avión. Salvo raras excepciones, este supuesto es habitual en la investigación en Psicología.

Una vez que tenemos el listado con todos los pares que queremos que los sujetos nos juzguen, debemos elaborar la escala de respuesta que deben utilizar los sujetos para juzgar la desemejanza de los objetos o estímulos incluidos en cada par. En la mencionada investigación se ha utilizado una escala de nueve puntos, donde 1 indicaba que los medios de transporte incluidos en el par eran muy parecidos y 9 que eran muy distintos. A modo de ejemplo:

Avión/tren

Muy parecidos		Moderadamente parecidos				Muy distintos		
1	2	3	4	5	6	7	8	9

Los sujetos utilizaron esta escala para juzgar la desemejanza en los 36 pares formados en el estudio.

En el Figura 2 se ofrece el mapa perceptivo de los medios de transporte para un sujeto de la muestra. En la configuración de puntos obtenida ahora tenemos un problema añadido con respecto a la configuración de la Figura 1. En el problema de las distancias entre ciudades conocíamos el significado de los ejes, un eje era la dirección norte-sur y otro eje era la dirección este-oeste. Pero ¿qué significado tienen ahora los ejes de la

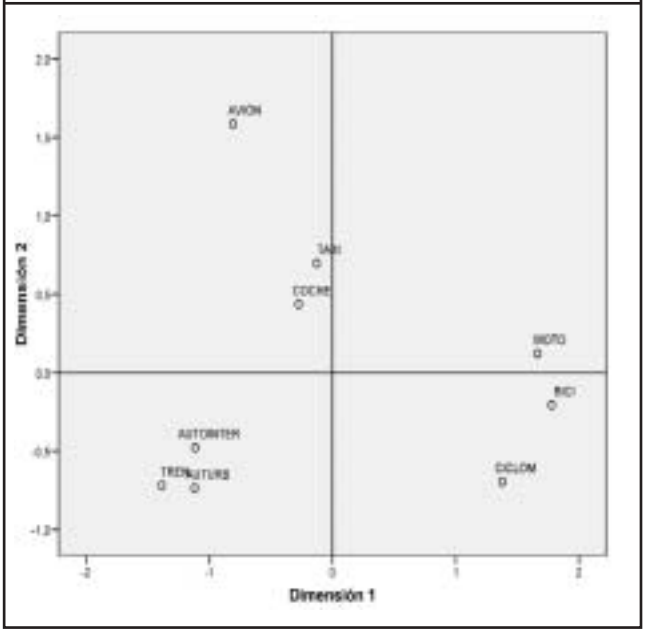
configuración que hemos obtenido? El sujeto probablemente haya utilizado en sus juicios distintos ejes o dimensiones para evaluar la desemejanza entre los medios de transporte. Por ejemplo, puede que para juzgar la desemejanza para un par dado se haya fijado en la seguridad de los medios de transporte. Para otro par pudo haberse fijado en el prestigio social, etc. Mediante el escalamiento multidimensional se busca obtener una configuración de puntos, pero también averiguar el significado de cada uno de los ejes o dimensiones de dicha configuración. Existen varios modos para enfrentarse a esta cuestión, pero el más fiable pasa por recoger más datos. En efecto, en la mencionada investigación, además de pedir a los sujetos los juicios de desemejanza, se les ha pedido que evaluaran cada uno de los medios de transporte en un serie de propiedades entre las que figuraban la seguridad, la estabilidad, la resistencia, la fuerza, el peso, la atracción, el prestigio, la puntualidad, el estatus social o la confortabilidad de los medios de transporte. Luego, se ha averiguado si existía algún tipo de relación entre alguna de estas propiedades y el posicionamiento de los medios de transporte en cada una de las dimensiones derivadas. En una primera fase exploratoria se probaron soluciones con 2, 3 y 4 dimensiones. La solución a la que se le encontró mejor significado fue la de 3 dimensiones. Análisis estadísticos de regresión múltiple, donde se tomaba como variable dependiente una propiedad dada de los medios de transporte y como variables independientes las coordenadas de los medios de transporte derivadas por los programas de escalamiento multidimensional, mostraron que la dimensión 1 (eje horizontal) representaba la seguridad percibida de los medios de transporte, la dimensión 2 (eje vertical) se refería al atractivo de los medios de transporte, y la dimensión 3 (eje de profundidad) representaba el prestigio social de los medios de transporte. En la Figura 2, donde se representan las dos primeras dimensiones de la solución tri-dimensional, los medios de transporte situados a la izquierda (tren, autobús interurbano, autobús urbano, etc.) se perciben como más seguros y los situados a la derecha (bicicleta, moto, ciclomotor) como más inseguros. De modo semejante, los medios de transporte situados más arriba (avión, taxi, coche, moto) se perciben como más atractivos y los situados más abajo como menos atractivos (autobús urbano, tren, ciclomotor, autobús interurbano, bicicleta).

EL CASO DE MÁS DE UNA MATRIZ DE ENTRADA: EL MODELO INDSCAL

En los ejemplos utilizados hasta ahora disponíamos de una matriz de entrada. En el primer problema, la matriz de distancias quilométricas entre las siete ciudades españolas; y, en el segundo problema, la matriz de desemejanzas entre los medios de transporte para un sujeto de la muestra. En el primer problema, en realidad, no existía otra posibilidad porque la matriz de distancias es única, pero en el segundo problema disponemos de múltiples sujetos y nos hubiera gustado introducir la matriz de desemejanzas de cada sujeto. De hecho, en la investigación original se ha hecho así. Hoy en día cualquier programa de escalamiento multidimensional permite la obtención de un espacio de objetos común compartido por una muestra de sujetos u otra fuente de datos.

Entre los modelos que tratan el tema de múltiples matrices de entrada, existe uno que merece especial atención porque dispone de propiedades que pueden resultar muy interesantes desde el punto de vista psicológico. Se trata del modelo INDSCAL de Carroll y Chang (1970). Este modelo permite obtener dos espacios: el espacio de objetos, común para todos los sujetos de la muestra, y el espacio de sujetos. El aspecto novedoso del modelo es realmente este último espacio. En el espacio de sujetos se representa el peso, la ponderación o importancia que

FIGURA 2
MAPA PERCEPTIVO DE LOS MEDIOS DE TRANSPORTE (DIMENSIÓN 1 VERSUS DIMENSIÓN 2)



cada sujeto de la muestra concede a cada una de las dimensiones de la configuración de objetos. Es decir, los sujetos comparten un mismo espacio de objetos pero el modelo permite que cada uno perciba dicho espacio de manera distinta; permite en definitiva las diferencias individuales entre unos y otros sujetos.

Arce (1994) pidió a dos sujetos que evaluaran la desemejanza entre 7 marcas de coches: Ferrari, Porche, BMW, Mercedes, Renault, Seat y Opel. Obtuvo dos matrices de desemejanzas y utilizó ambas como entrada para un escalamiento multidimensional. Los resultados evidenciaron que la primera dimensión perceptiva eran los rasgos deportivos de la marca de coches y la segunda su confortabilidad. En la Figura 3 se ofrece el espacio de objetos, común, compartido por los dos sujetos. Los coches situados más a la derecha (dimensión 1) se perciben como más deportivos que los situados a la izquierda y los situados más abajo (dimensión 2) como más confortables que los situados más arriba.

FIGURA 3
ESPACIO DE OBJETOS

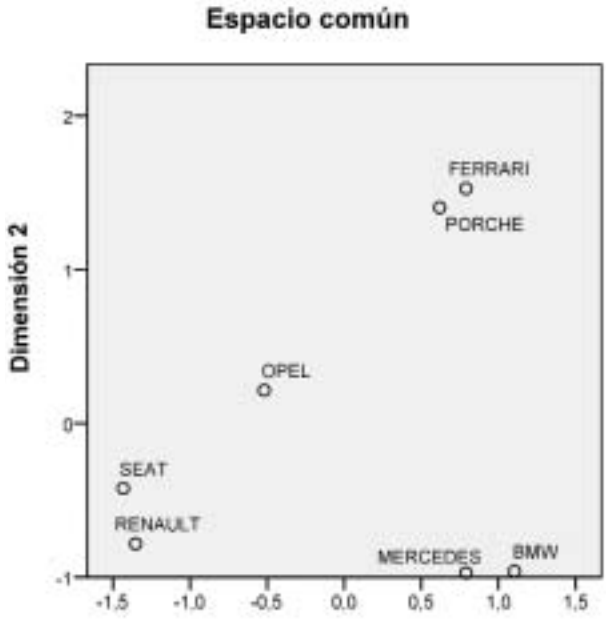
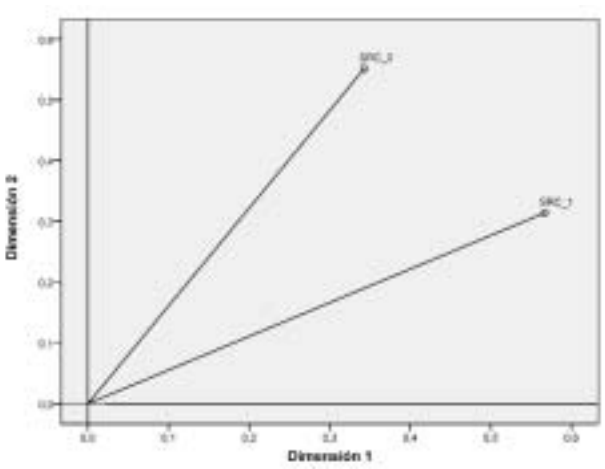


FIGURA 4
ESPACIO DE SUJETOS



En la Figura 4 se ofrece el espacio de sujetos. Mientras el espacio de objetos es común para los dos objetos, el espacio de sujetos nos indica que el sujeto 1 (SRC_1 en el gráfico) concede más importancia a la dimensión 1, los rasgos deportivos del coche, mientras el sujeto 2 (SRC_2, en el gráfico) concede más importancia a la confortabilidad de los coches. A diferencia del espacio de objetos donde cada objeto se representa por un punto, en el espacio de sujetos, el sujeto se representa por un vector (una línea). Cuanto más cerca esté el vector de una dimensión, más importancia le concede el sujeto a dicha dimensión y cuanto más alejado menos importancia. En efecto, se observa en el gráfico que el sujeto 1 está más cerca de la dimensión 1 (diseño deportivo), concediendo por tanto más importancia a esta dimensión, mientras el sujeto 2 está más cerca de la dimensión 2 indicando que en su caso es ésta la dimensión (confortabilidad de los coches) la que adquiere más peso en sus juicios sobre las marcas de coche.

ESCALAMIENTO MULTIDIMENSIONAL CON DATOS DE PREFERENCIA

Aunque el escalamiento multidimensional utiliza, en su forma más típica, una matriz de desemejanzas entre objetos como entrada, se han desarrollado modelos y métodos que permiten el escalamiento multidimensional de objetos a partir de datos de preferencia (p.e. Bennett y Hays, 1960; Carroll, 1980; Tucker, 1960). Si tenemos n objetos que queremos escalar, al sujeto simplemente se le pide que los posicione por orden de preferencia, asignándole el número 1 al objeto más preferido, el 2 al segundo más preferido y así sucesivamente hasta el último objeto, al que debe asignar el número n. Estos datos tienen la ventaja de que son más cómodos de obtener que los datos de desemejanza. La tarea suele ser mucho más simple tanto para el sujeto como para el investigador. Los datos de preferencia se ordenan, luego, en una matriz rectangular, con sujetos en las filas y objetos en las

columnas. Cada fila es un sujeto y un elemento de la fila representa el orden (o preferencia) que dicho sujeto ha concedido a un objeto dado.

A modo de ejemplo, supongamos que estuviésemos interesados en obtener un mapa perceptivo de los deportes y actividades físicas que los ciudadanos pueden practicar en su tiempo de ocio. Para hacer el ejemplo manejable, elegimos 8 deportes o actividades físicas y 16 sujetos, a los que les pedimos que nos indiquen sus preferencias marcando con un 1 el deporte o actividad física más preferida para él/ella, con un 2 el deporte o actividad física que prefiere en segundo lugar y así sucesivamente hasta marcar el que prefiere en último lugar, al que le asignaría el número 8.

Los deportes o actividades físicas elegidos en el ejemplo son: fútbol, baloncesto, tenis, atletismo, caminar, nadar, andar en bicicleta y correr.

Las preferencias indicadas por los sujetos se ofrecen en Tabla 4.

En la Figura 5 se ofrece el mapa perceptivo de los deportes y actividades físicas evaluadas por los sujetos de la muestra. Para interpretar el significado de las dimensiones, nos fijamos, en primer lugar, en las propiedades de los deportes o actividades físicas que ocupan los lugares más extremos en cada una de las dimensiones. Es probable que tengan alguna propiedad contrapuesta que nos ayude a interpretar el significado de la dimensión respectiva. Así, se puede observar que en la dimensión 1 (horizontal) a la derecha están situadas las actividades físicas no competitivas (caminar, nadar, correr y andar en bicicleta) y a la izquierda los deportes competitivos (fútbol, atletismo, tenis y baloncesto). Podría interpretarse, por tanto, la dimensión 1 como la competitividad de los deportes o actividades físicas. De modo semejante, si observamos el posicionamiento de los

FIGURA 5
MAPA PERCEPTIVO DERIVADO POR PROXSCAL PARA LOS DEPORTES Y ACTIVIDADES FÍSICAS A PARTIR DE DATOS DE PREFERENCIA

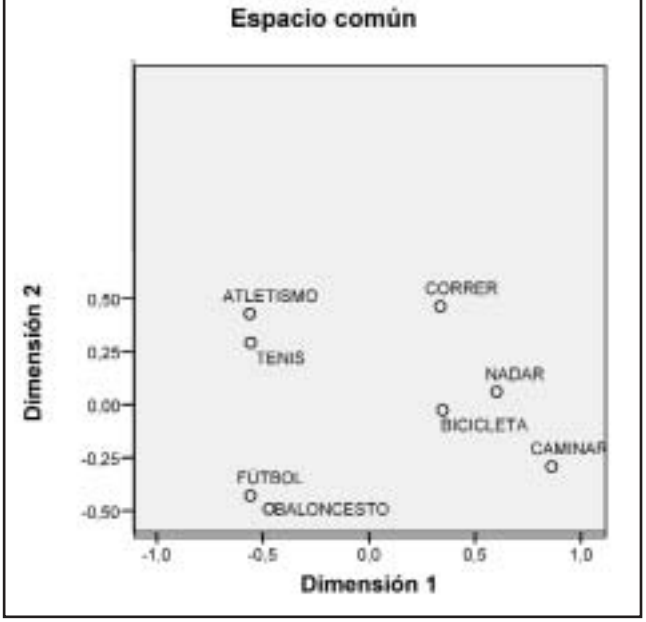


TABLA 4
PREFERENCIAS DE LOS SUJETOS

Sujeto	Fútbol	Baloncesto	Tenis	Atletismo	Caminar	Nadar	Andar Bicicleta	Correr
1	8	7	6	5	1	4	3	2
2	7	8	5	6	2	3	4	1
3	8	7	6	5	1	3	2	4
4	7	8	5	6	2	4	3	1
5	6	5	7	8	1	2	3	4
6	5	6	7	8	2	3	4	3
7	6	5	7	8	2	1	3	4
8	5	6	8	7	3	2	4	3
9	1	2	3	4	5	6	7	8
10	2	1	4	3	5	7	6	8
11	1	2	4	3	8	7	6	5
12	2	1	3	4	8	6	7	5
13	3	4	1	2	8	6	7	5
14	4	3	1	2	8	7	6	5
15	4	3	2	1	7	8	5	6
16	3	4	2	1	8	6	7	5

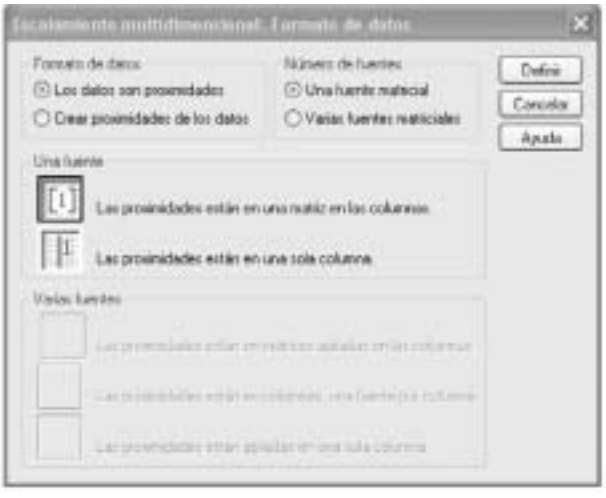
deportes y actividades físicas en la dimensión 2 (eje vertical), podemos apreciar que hacia arriba están los deportes o actividades físicas de naturaleza individual (atletismo, correr, tenis, etc.) y en la parte inferior los deportes colectivos (fútbol, baloncesto). Podría interpretarse, por tanto, esta segunda dimensión como el tipo de deporte o actividad: individual versus colectivo.

RESOLUCIÓN DE PROBLEMAS DE ESCALAMIENTO MULTIDIMENSIONAL CON SPSS

Hasta ahora hemos pasado por encima, sin detenernos, en el proceso de resolución de los problemas de escalamiento por medio de software específico. Vamos a reproducir ahora cómo hemos resuelto el problema de las distancias quilométricas con el procedimiento PROXSCAL implementado en SPSS. Simultáneamente, en ocasiones, indicaremos cuáles son las diferencias en la toma de decisiones entre este problema y los otros tres que también hemos resuelto.

FIGURA 6
FICHERO DE ENTRADA PARA SPSS CON LAS DISTANCIAS QUILOMÉTRICAS ENTRE LAS SIETE CIUDADES

FIGURA 7
ELECCIÓN DE LA FORMA DE LOS DATOS Y DEL NÚMERO DE MATRICES DE ENTRADA EN EL PROCEDIMIENTO PROXSCAL



Paso 1. Creamos el fichero de datos en SPSS con las distancias quilométricas entre las siete ciudades españolas. Debe tener la apariencia que ofrece la Figura 6. Dado que la matriz de entrada es cuadrada, las filas en la matriz tienen el mismo significado que las columnas. La fila 1 es Coruña, la fila 2 Barcelona y así sucesivamente hasta la fila 7 que es Valencia. El hecho de que aparezca el nombre de las ciudades en las columnas y no aparezca en las filas es porque el sistema SPSS permite, utilizando la pestaña Vista de variables, etiquetar o dar nombres a las columnas de la matriz de datos pero no así a las filas.

Paso 2. Elegimos el procedimiento que queremos ejecutar:
Analizar / Escalas / Escalamiento multidimensional (PROXSCAL)

Paso 3. Forma de los datos
El procedimiento PROXSCAL permite dos tipos de datos de entrada:
(a) datos de proximidad (matriz cuadrada)
(b) datos de perfil (matriz rectangular)
Las distancias quilométricas, al igual que las desemejanzas entre objetos, son datos de proximidad. Elegimos, por tanto, la opción que indica al programa que los datos son proximidades (ver Figura 7).
Si tuviésemos una matriz rectangular de entrada como, por ejemplo, cuando disponíamos de preferencias en el ejemplo de los deportes o actividades físicas, entonces tendríamos que elegir la opción que indica al programa que debe crear proximidades de los datos.

Paso 4. Número de matrices de entrada
El procedimiento PROXSCAL permite una o más de una matriz de entrada. El número de matrices de las que disponemos en el problema se indica en el recuadro denominado número de fuentes. Como, en este caso, sólo disponemos de una matriz elegimos la opción una fuente matricial (ver Figura 7). En el tercer problema que hemos resuelto aquí, el de las marcas de coche, hemos elegido la opción varias fuentes matriciales, dado que disponíamos de dos matrices, una por sujeto. Las matrices, en el fichero de entrada se sitúan unas debajo de otras, respetando el mismo formato en todas ellas.

Paso 5. Pulsamos el botón Definir (ver Figura 7)

Paso 6. Seleccionamos los objetos que queremos escalar (aquí ciudades)

Para ello marcamos las 7 ciudades en el recuadro de la izquierda de la Figura 8 y las pasamos al recuadro denominado Proximidades pulsando la flecha que está entre ambos.

Paso 7. Elección del Modelo (pulsamos el botón denominado Modelo)

Dado que tenemos una sola matriz de proximidades (distancias) de entrada, el procedimiento no permite elegir el modelo de escalamiento (ver Figura 9) que, en cualquier caso, será semejante al modelo clásico.

Si tuviésemos más de una matriz de proximidades de entrada, entonces sí que podríamos elegir el modelo de escalamiento. Los más habituales serían el modelo con replicación (denominado modelo de Identidad en el cuadro de diálogo) y el modelo INDSCAL (denominado Euclídeo ponderado en el cuadro de diálogo). En el modelo de replicación los sujetos se consideran replicaciones unos de otros, lo que significa que las diferencias que puedan existir entre ellos se atribuyen a factores aleatorios. El modelo INDSCAL, por el contrario, permite las diferencias individuales. En el problema de las marcas de coche, hemos elegido este modelo.

Paso 8. Más decisiones sobre la forma y la naturaleza de los datos

El programa hasta ahora sabe que tenemos una matriz de proximidades de entrada (cuadrada), pero todavía nos pide, bajo la denominación de Forma (ver Figura 9), que le especifiquemos si la información la tenemos en el triángulo inferior, en el superior o en toda la matriz. Como la matriz de distancias es simétrica, hemos optado por disponer la información tan sólo en la mitad inferior (matriz triangular inferior). La opción de matriz completa sólo se usa cuando la matriz de entrada es asimétrica.

Los programas de escalamiento multidimensional proyectan las desemejanzas como distancias en el espacio. Cuanto mayor sea la desemejanza entre objetos mayor será su distancia en el espacio multidimensional. Pero los datos de entrada pueden ser desemejanzas o semejanzas. Si fuesen semejanzas entonces la relación con las distancias sería inversa: cuanto mayor sea la semejanza entre dos objetos en el mundo empírico menor sería la distancia entre ellos en el espacio. Esta es, pues, una especificación sustancial que el usuario debe hacer al programa. En nuestro caso, en el recuadro Proximidades, debemos elegir Disimilaridades (sinónimo de desemejanzas). Las distancias se conciben como desemejanzas.

Bajo la denominación de Transformación de las proximidades (ver Figura 9), el procedimiento permite al usuario elegir la escala de medida para los datos de entrada. Si elegimos razón o intervalos, el modelo será métrico y si elegimos ordinal el modelo será no-métrico. Rara vez en la investigación en Psicología se elige el nivel de medida de razón; los más frecuentes son intervalos u ordinal. Aquí, en este problema, sin embargo, dado que los datos son distancias y no juicios de los sujetos, elegimos el nivel de medida más alto (razón).

En el problema de los deportes, donde utilizábamos datos de preferencia, hemos especificado que el nivel de

FIGURA 8
SELECCIÓN DE LOS OBJETOS QUE QUEREMOS ESCALAR (AQUÍ CIUDADES)

FIGURA 9
ELECCIÓN DEL MODELO DE ESCALAMIENTO

medida era ordinal. En este caso, el programa todavía nos permite hacer una especificación más, bajo la denominación de Desempatar observaciones empatadas. Esta es una decisión muy técnica. Si la elegimos, el programa asumirá que el proceso de medida es continuo y si no la elegimos que es discreto. Esta decisión tan sólo tiene repercusión para aquellos casos en que haya empates. Por defecto, el programa asume que el proceso de medida es discreto y respeta los empates en los datos. Si consideramos que el proceso de medida es continuo debemos especificar al programa que proceda a desempatar las observaciones empatadas. En nuestro problema de los deportes, hemos probado con las dos opciones sin que hayamos notado diferencias en las soluciones derivadas por el programa. De hecho, esto es lo que ocurre en la mayoría de las ocasiones. Es decir, se trata de decisiones que tienen importancia a nivel matemático pero no tanta a nivel sustantivo.

Paso 9. Número de dimensiones

Si tenemos una hipótesis clara de partida podemos elegir un número de dimensiones fijo, y si no la tenemos lo mejor es que probemos a obtener soluciones distintas y a posteriori seleccionemos la solución con el número de dimensiones que sea más interpretable desde un punto de vista sustantivo. En nuestro ejemplo de las distancias, el número de dimensiones que hemos elegido, dado que se trataba de un mapa, fueron dos (ver Figura 9).

En el ejemplo de los medios de transporte, como puede ocurrir en otras muchas investigaciones en Psicología que se realizan con carácter exploratorio, no teníamos una hipótesis tan clara en cuanto al significado de las dimensiones que podríamos obtener. En consecuencia, hemos probado a obtener soluciones en dos, tres y cuatro dimensiones. Luego, a posteriori, hemos intentado buscarles un significado. Hemos comprobado que eran interpretables tres y ésta fue la solución que hemos elegido.

La interpretación de las dimensiones puede hacerla el investigador tratando de analizar, en primer lugar, las propiedades de los objetos que ocupan posiciones más extremas en la dimensión. Cuando el procedimiento separa mucho a los objetos suele ser porque tienen alguna propiedad contrapuesta que, si la identificamos, puede ayudarnos a dar nombre a la dimensión. Este procedimiento lo hemos utilizado en el ejemplo de los deportes. No obstante, esta interpretación basada en la opinión de un experto (el investigador) puede ser discutida por otros investigadores (o expertos). Lo ideal es proceder como hemos hecho en el problema de los medios de transporte, donde además de los juicios de desemejanza entre objetos se ha pedido a los sujetos que evaluasen cada uno de los medios de transporte sobre una serie de escalas bipolares que representaban hipotéticas propiedades de los medios de transporte. Luego, por métodos estadísticos de correlación y regresión, se ha podido ofrecer evidencia de cuál era el verdadero significado de cada una de las dimensiones retenidas.

Paso 10. Restricciones y Opciones

Tanto el botón de Restricciones como el de Opciones (ver Figura 8) permiten al usuario tomar decisiones de un nivel muy avanzado. En la práctica, se suelen tomar las opciones que el procedimiento tiene implementadas por defecto. En Restricciones, por defecto, el programa asume que debe estimar todas las coordenadas de los objetos (sin restricciones). A veces, de manera excepcional, las coordenadas se conocen y lo único que se busca es proyectar nuevos objetos sobre un espacio ya definido. En este caso, habría que proporcionar al programa las coordenadas que leería en un archivo que nosotros le indiquemos.

De forma semejante, en Opciones, por defecto, el programa toma una determinada configuración inicial (simplex) que nos permite cambiar por otras alternativas (p.e. Torgerson). También nos permite cambiar los criterios para alcanzar la convergencia y el número de iteraciones que realiza el algoritmo. Rara vez se podrán obtener mejores resultados si se cambian las opciones que el programa tiene incorporadas por defecto.

Paso 11. Toma de decisiones sobre la salida

En los botones denominados Gráficos y Resultados (ver Figura 8), el programa permite al usuario elegir lo que quiere que aparezca en la salida. Podemos oscilar desde una salida muy simple, con los elementos sustanciales, hasta una salida muy sobrecargada con todo tipo de detalles técnicos. Es aconsejable, en un primer momento, obtener una salida simple y, luego, si fuese necesario, obtener nuevas salidas con más elementos informativos. Por defecto, en Gráficos, el programa nos ofrece el gráfico más relevante que es el espacio de objetos (espacio común). Adicionalmente podríamos pedirle otros gráficos que nos permitirían visualizar el grado de ajuste del modelo. Por su parte, por defecto, en Resultados, el programa nos ofrece las coordenadas de los objetos (aquí ciudades) y los índices de ajuste del modelo, tal como aparecen en la Tabla 2.

PROGRAMAS DE ORDENADOR PARA EL ESCALAMIENTO MULTIDIMENSIONAL

Existe una amplia lista de programas de ordenador para la solución de problemas de escalamiento multidimensional. Nosotros hemos resuelto todos los problemas anteriores haciendo uso del procedimiento PROXSCAL, implementado en SPSS, pero el mismo paquete estadístico dispone de otro procedimiento, denominado ALSCAL, que también permite la solución de múltiples problemas de escalamiento multidimensional. Para acceder a este procedimiento se debe seguir la secuencia Analizar/Escalas/Escalamiento multidimensional (ALSCAL). Cualquier problema de los que aquí hemos tratado se podría haber resuelto igualmente con ALSCAL.

En la Tabla 5 se ofrece una pequeña relación de programas de ordenador actualmente disponibles en el mercado. Además de los ya señalados, PROXSCAL y ALSCAL, es posible resolver problemas de escalamiento multidimensional con otros programas tales como GGVIS, PERMAP, MULTISCALE o NewMDSX. GGVIS y PERMAP comparten la propiedad de ser interactivos y de poderse adquirir gratuitamente por Internet. MULTISCALE tiene la ventaja de estar también disponible de forma gratuita, pero es de difícil manejo. Su autor, Ramsay, goza de un gran prestigio en la historia del escalamiento multidimensional. Finalmente, NewMDSX es, en realidad, un paquete de programas que permite la solución de problemas de escalamiento multidimensional y de otro tipo de problemas relacionados.

TABLA 5
PROGRAMAS DE ORDENADOR PARA EL ESCALAMIENTO MULTIDIMENSIONAL

NOMBRE	DISPONIBILIDAD	DOCUMENTALES
PROXSCAL	En SPSS http://www.spss.com/	Commandeur y Heiser (1993), Meulman, Heiser y SPSS (1999), De Leeuw y Heiser (1980)
ALSCAL	En SPSS http://www.spss.com/	Takane, Young y De Leeuw (1977)
GGVIS	Gratuito, por Internet http://www.ggobi.org E-mail: ggobi-help@ggobi.org	Buja y Swayne (2002)
PERMAP	Gratuito, por Internet http://www.ucs.louisiana.edu/~rbh8900 E-mail: ron@heady.us	Ron B. Heady, University of Louisiana, Lafayette, USA
MULTISCALE	Gratuito, por Internet ftp://ego.psych.mcgill.ca/pub/ramsay/multiscl/ o dirigiéndose al autor, profesor James O. Ramsay, e-mail: ramsay@psych.mcgill.ca	Ramsay (1977)
NewMDSX	http://www.newmdsx.com/ Coxon (2004)	Coxon (2004)

Si desea más información sobre programas de ordenador y, de manera más general, sobre el escalamiento multidimensional en relación con la historia, los modelos, los métodos y las múltiples posibilidades de aplicación en Psicología y temas relacionados pueden servirle de ayuda los libros de Borg y Groenen (2005), el manual más reciente que se haya escrito sobre escalamiento multidimensional a fecha de hoy, Kruskal y Wish (1978), Arabie, Carroll y DeSarbo (1987), Green, Carmone y Smith (1989), o Arce (1993, 1994). Para ejemplos de aplicaciones véase Wish, Deutsch y Kaplan (1976) o Sabucedo y Arce (1990).

AGRADECIMIENTOS

Esta investigación ha sido realizada con la ayuda de la Dirección Xeral de Investigación, Desenvolvemento e Innovación de la Xunta de Galicia (PGIDIT06PXIB211187PR).

REFERENCIAS

Arabie, P., Carroll, J.D., y DeSarbo, W.S. (1987). Three-way scaling and clustering. Newbury Park, CA: Sage.

Arce, C. (1993). *Escalamiento multidimensional: una técnica multivariante para el análisis de datos de proximidad y preferencia*. Barcelona: Promociones y Publicaciones Universitarias (PPU).

Arce, C. (1994). *Técnicas de construcción de escalas psicológicas*. Madrid: Editorial Síntesis.

Bennett, J.F., y Hays, W.L. (1960). Multidimensional unfolding: determining the dimensionality of ranked preference data. *Psychometrika*, 25, 27-43.

Borg, I., y Groenen, P.J.F. *Modern multidimensional scaling. Theory and applications*. Nueva York: Springer.

Buja, A., y Swayne, D.F. (2002). Visualization methodology for multidimensional scaling. *Journal of Classification*, 19, 7-44.

Carroll, J.D. (1980). Models and methods for multidimensional analysis of preferential choice (or other dominance) data. En E.D. Lantermann y H. Feger (Eds.), *Similarity and choice* (pp. 234-289). Viena: Hans Huber.

Carroll, J.D. y Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via N-way generalization of Eckart-Young decomposition. *Psychometrika*, 35, 283-319.

Commandeur, J.J.F., y Heiser, W.J. (1993). Mathematical derivations in the proximity scaling (PROXSCAL) of symmetric data matrices. Tech. Rep. No. RR-93-03. Leiden, The Netherlands: Department of Data Theory, Leiden University.

Coxon. A.P.M. (2004). Multidimensional Scaling. En M.S. Lewis-Beck, A. Bryman, y T. F. Liao (Eds.), *The Sage Encyclopedia of Social Science Research Methods*. Thousand Oaks: Sage.

De Leeuw, J., y Heiser, W.J. (1980). Multidimensional scaling with restrictions on the configuration. En P.R. Krishnaiah (Ed.), *Multivariate analysis* (Vol. V, pp. 501-522). Amsterdam, Holanda: North-Holland.

Green, P.E., Carmone, F.J., y Smith, S.M. (1989). *Multidimensional scaling. Concepts and applications*. Boston: Allyn and Bacon.

Kruskal, J.B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-27.

Kruskal, J.B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika,* 29, 115-129.

Kruskal, J.B., y Wish, M. (1978). *Multidimensional scaling*. Newbury Park, CA: Sage.

Meulman, J.J., Heiser, W.J. y SPSS (1999). *SPSS Categories 10.0*. Chicago: SPSS.

Ramsay, J.O. (1977). Maximun likelihood estimation in multidimensional scaling. *Psychometrika*, 42, 241-266.

Sabucedo, J.M. , y Arce, C. (1990). Types of political participation: a multidimensional analysis. *European Journal of Political Research*, 20, 93-102.

Shepard, R.N. (1962). The analysis of proximities: multidimensional scaling with an unknown distances function (I y II). *Psychometrika*, 27, 125-139, 219-246.

Takane, Y., Young, F.W., y De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least-squares method with optimal scaling features. *Psychometrika*, 42, 7-67.

Torgerson, W.S. (1958). *Theory and methods of scaling*. Nueva York: Wiley.

Tucker, L.R. (1960). Intra-individual and inter-individual multidimensionality, en H. Gulliksen y S. Messick (Eds.), *Psychological scaling: theory and applications* (pp. 155-167). Nueva York: Wiley.

Wish, M., Deutsch, M. y Kaplan, S.J. (1976). Perceived dimensions of interpersonal relations. *Journal of Personality and social Psychology*, 33, 409-420.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 57-66
http://www.cop.es/papeles

LAS TEORÍAS DE LOS TESTS: TEORÍA CLÁSICA Y TEORÍA DE RESPUESTA A LOS ÍTEMS

TEST THEORIES: CLASSICAL THEORY AND ITEM RESPONSE THEORY

José Muñiz
Facultad de Psicología. Universidad de Oviedo

Para una interpretación y utilización adecuada de las propiedades psicométricas de los tests es necesario ir más allá del mero cálculo empírico, y conocer los fundamentos en los que se basan esos cálculos. Con el fin de contribuir a esta comprensión más allá del mero manejo superficial de la fórmulas psicométricas, el objetivo fundamental de este trabajo es presentar de una manera no excesivamente técnica y especializada las dos grandes teorías que guían la construcción y análisis de la mayoría de los tests: la Teoría Clásica de los Tests y la Teoría de Respuesta a los Ítems. En primer lugar se hace un apunte histórico sobre los tests, indicando cómo surgen y evolucionan al hilo de los avances técnicos y estadísticos. Tras razonar acerca de la necesidad de utilizar teorías psicométricas para el análisis y construcción de los tests, se expone la lógica que subyace a la Teoría Clásica de los Tests, así como sus dos variantes más granadas, la Teoría de la Generalizabilidad y los Tests Referidos al Criterio. Luego se subrayan las limitaciones más importantes del enfoque clásico y se exponen los fundamentos de la Teoría de Respuesta a los Ítems, dentro de cuyo marco encuentran una solución satisfactoria algunos de los problemas que el enfoque clásico no había sido capaz de resolver de forma satisfactoria. Finalmente se comparan ambos enfoques, y se concluye indicando la necesidad de conocer las teorías de los tests para una mejor comprensión y utilización de los instrumentos de medida.
Palabras clave: *Tests, Teoría Clásica de los Tests, Teoría de Respuesta a los Ítems, Teorías de los tests.*

For a correct interpretation and proper use of the psychometric properties of tests it is necessary to go beyond the mere empirical calculation, and know the grounds on which these calculations are based. To contribute to this understanding beyond the superficial handling of the psychometric formulas, the main goal of this work is to present, in a not technical way, the two most important theories that guide the development and analysis of most tests: Classical Test Theory and Item Response Theory. First, a historic note about tests and testing is made, indicating the evolution of tests according to the technical and statistical advances. The importance of test theories in order to develop and analyse tests is pointed out, and Classical Test Theory, including Generalizability Theory and Criterion Referenced Tests, is presented. After underlining the limitations of the Classical Test Theory approach, Item Response Theory is presented. Within this new framework some of the limitations of the Classical Test Theory find a proper solution. Finally both approaches are compared, emphasizing the importance of test theories for a correct use and interpretation of psychometric properties of the tests.
Key words: *Tests, Classical Test Theory, Item Response Theory, Test theories.*

Los tests constituyen seguramente la tecnología más sofisticada de la que disponen los psicólogos para ejercer su profesión, por eso no es infrecuente que la sociedad identifique a los psicólogos con los tests. Naturalmente, unos psicólogos utilizan los tests más que otros, dependiendo de su campo profesional y de su forma de trabajar. Los tests son muestras de conducta que permiten llevar a cabo inferencias relevantes sobre la conducta de las personas. Bien utilizados son herramientas claves en la profesión del psicólogo. No conviene olvidar que los tests nacen con un afán de objetividad y justicia, para evaluar a las personas por lo que realmente valen, evitando evaluaciones sesgadas por aspectos tales como la cuna, la clase social, la raza, el sexo, las creencias, las cartas de recomendación, y otros sistemas de evaluación subjetivos. Unas veces estos nobles fines se han alcanzado mejor que otras, pero ésa era y sigue siendo la idea central, evaluar a todos por el mismo rasero.

Correspondencia: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España. E-mail: jmuniz@uniovi.es*

NOTA HISTÓRICA

¿Cuándo aparecen los tests por primera vez en la historia? Suele citarse como el origen remoto de los tests unas pruebas que los emperadores chinos ya hacían allá por el año 3000 antes de Cristo para evaluar la competencia profesional de los oficiales que iban a entrar a su servicio. Otras muchas huellas antiguas pueden rastrarse, pero los tests actuales tienen sus orígenes más cercanos en las pruebas senso-motoras utilizadas por Galton

(1822-1911) en su laboratorio antropométrico. Pero será James McKeen Cattell (1860-1944) el primero inutilizar el término *test mental*, en 1890. Pronto quedó claro (Wissler, 1901) que estos primeros tests senso-motores no eran buenos predictores de las capacidades cognoscitivas de las personas, y Binet y Simon (1905) darán un giro radical al introducir en su nueva escala tareas cognoscitivas para evaluar aspectos como el juicio, la comprensión y el razonamiento. Terman llevó a cabo la revisión de la escala en la Universidad de Stanford, la cual se conoce como la revisión Stanford-Binet (Terman, 1916), utilizando por primera vez el concepto de Cociente Intelectual (CI) para expresar la puntuación de las personas. La idea del CI había sido propuesta originariamente por Stern, dividiendo la Edad mental por la Edad Cronológica y multiplicando el resultado por 100 para evitar decimales.

La escala de Binet abre una tradición de escalas individuales que llega hasta nuestros días. En 1917 los tests reciben otro gran impulso al aparecer los tests colectivos Alfa y Beta a raíz de la necesidad del ejército norteamericano de reclutar rápidamente soldados para la primera guerra mundial. El test Alfa iba dirigido a la población general y el Beta a personas analfabetas o que no dominaban el inglés. Las pruebas tuvieron mucho éxito y terminada la guerra las empresas y otras instituciones adoptaron de forma entusiasta el uso de los tests para distintos menesteres. Comenzaba así una expansión creciente en el uso y creación de tests de todo tipo. La aparición de la técnica del análisis factorial va a suponer un gran avance en la construcción y análisis de los tests, permitiendo la aparición de las baterías de tests, cuyo representante más genuino serían las *Aptitudes Mentales Primarias* (PMA) de Thurstone (Thurstone, 1938; Thurstone y Thurstone, 1941). En España tuvimos la suerte de que uno de los grandes pioneros de la Psicología Española, Mariano Yela, estudiase en Chicago con Thurstone en los años 40, lo que le permitió introducir en nuestro país todos los avances de la época, e impulsar la Psicometría tanto en el mundo académico, como su implementación aplicada, colaborando activamente en el desarrollo de la empresa TEA (Pereña, 2007). La división de la inteligencia en sus distintos factores o dimensiones dio lugar a la aparición de dos grandes líneas de estructuración de las dimensiones cognoscitivas, lo que ha dado en llamarse la escuela inglesa y la escuela americana. En la primera se da más importancia a un factor central de inteligencia general, que coronaría una estructura en la que luego vendrían dos amplias dimensiones, la verbal-educativa y la mecánico-espacial, en las que se articularían otros muchos factores más específicos. El enfoque americano asume una serie de dimensiones no jerarquizadas que compondrían el perfil cognoscitivo, que por ejemplo en el caso del PMA serían: la comprensión verbal, la fluidez verbal, aptitud numérica, aptitud espacial, memoria, rapidez perceptiva y razonamiento general. Ambos enfoques son compatibles, y tienen mucho que ver con la tecnología estadística utilizada, sobre todo el análisis factorial. Toda esta línea de investigaciones psicométricas sobre la inteligencia culmina en la obra magna de Carroll (1993), donde se sintetizan los grandes avances alcanzados. En España trabajos como los de Juan-Espinosa (1997), Colom (1995), o Andrés-Pueyo (1996) recogen y analizan de forma brillante este campo de trabajo.

Pero no sólo se producen avances en el campo de los tests cognoscitivos, también los tests de personalidad se aprovechan de los avances que se producen en la psicometría. Suele citarse la hoja de datos personales utilizada por Woodworth en 1917 para detectar neuróticos graves como el pionero de los tests de personalidad. Por su parte el psiquiatra suizo Rorschach propone en 1921 su test proyectivo de manchas de tinta, al que seguirán otros muchos tests basados en el principio de la proyección, que asume que ante un estímulo ambiguo, la persona evaluada tenderá a producir respuestas que de algún modo reflejan aspectos importantes de su personalidad. El lector interesado en la historia de los tests puede consultar por ejemplo el libro de Anastasi y Urbina (1998), aquí solo tratamos de dar unas pinceladas para entender lo que sigue.

Tras esta larga andadura de unos cien años, uno puede preguntarse, por curiosidad, cuáles son en la actualidad los tests más utilizados por los psicólogos españoles, y si estos difieren de los que utilizan sus colegas europeos. Pues bien, en una encuesta reciente hecha en seis países europeos los tests más utilizados por los psicólogos españoles fueron: 16PF, WISC, WAIS, MMPI, Beck, STAI, Rorschach, Raven, Bender e ISRA. Estos datos son muy similares a los obtenidos en otros países europeos (Muñiz et al., 2001).

En suma, la historia de los tests es una historia exitosa de la que la psicología tiene que sentirse orgullosa, sin olvidar, claro está, que como ocurre con cualquier tecno-

logía de cualquier campo, en ocasiones su utilización por manos inexpertas ha dejado mucho que desear. Es por ello que en la actualidad distintas organizaciones nacionales (Colegio Oficial de Psicólogos, COP) e internacionales (Federación Europea de Asociaciones de Psicólogos, EFPA; Comisión Internacional de Tests, ITC, Asociación Americana de Psicología, APA) desarrollan numerosos proyectos y actividades para potenciar el uso adecuado de los tests (Muñiz, 1997b; Muñiz y Bartram, 2007; Prieto y Muñiz, 2000).

¿POR QUÉ HACEN FALTA TEORÍAS DE LOS TESTS?

Hemos visto en el apartado anterior una breve reseña histórica de cómo han surgido y han ido evolucionando los tests concretos, pero nada hemos dicho acerca de las teorías que posibilitan la construcción de los tests. Así contado podría pensarse que los tests se van sucediendo sin orden ni concierto, pero nada más lejos de la realidad. A la construcción y análisis de los tests subyacen teorías que guían su construcción y que condicionan y tiñen los tests según los avances teóricos y estadísticos de cada momento.

A la vista de ello uno puede preguntarse con toda razón: ¿por qué hacen falta teorías de los tests? O si se quiere de un modo más pragmático, ¿Por qué y para qué tienen los psicólogos en su carrera la asignatura de Psicometría dedicada fundamentalmente a exponer estas teorías? La razón es bien sencilla, los tests son instrumentos de medida sofisticados mediante los cuales los psicólogos llevan a cabo inferencias y toman decisiones sobre aspectos importantes de las personas. Por tanto hay que asegurarse de que esas inferencias son adecuadas y pertinentes, de lo contrario se puede perjudicar notablemente a las personas que acuden a los psicólogos por la razón que sea. Las teorías estadísticas de los tests van a permitir la estimación de las propiedades psicométricas de los tests para de ese modo garantizar que las decisiones tomadas a partir de ellos son las adecuadas. Sin esas teorías no podríamos estimar la fiabilidad y la validez de los tests, lo cual es imprescindible para poder usar los tests de forma rigurosa y científica. Por supuesto, aparte de estas teorías estadísticas sobre los tests, la construcción de una prueba debe de guiarse por un modelo o teoría psicológica sustantiva que dirige su construcción. En el trabajo de Muñiz y Fonseca-Pedrero (2008) pueden consultarse los pasos fundamentales para llevar a cabo la construcción de un test. Para un análisis más detallado del proceso de construcción de un test pueden verse por ejemplo los trabajos de Carretero y Pérez (2005), Downing y Haladyna (2006), Morales, Urosa y Blanco (2003), Muñiz (2000), Schmeiser y Welch (2006), o Wilson (2005).

Hay dos grandes enfoques o teorías a la hora de construir y analizar los tests, son la Teoría Clásica de los Tests (TCT) y el enfoque de la Teoría de Respuesta a los Ítems (TRI). No se trata aquí de llevar a cabo exposiciones detalladas de estas teorías (en español pueden verse, por ejemplo, en Muñiz, 1997a, 2000, 2005), sino de subrayar los aspectos claves, para que así los usuarios de los tests tengan una idea más cabal y comprendan en profundidad el alcance de las propiedades psicométricas de los tests que están utilizando.

TEORÍA CLÁSICA DE LOS TESTS

El enfoque clásico es el predominante en la construcción y análisis de los tests, así, por ejemplo, los diez tests más utilizados por los psicólogos españoles citados en el apartado anterior, todos ellos, sin excepción, han sido desarrollados bajo la óptica clásica. Sólo este dato ya deja bien patente la necesidad de que los profesionales entiendan perfectamente la lógica clásica, sus posibilidades y sus limitaciones.

Antes de entrar en la lógica de la teoría clásica, hay que señalar que hinca sus raíces en los trabajos pioneros de Spearman de principios del siglo XX (Spearman, 1904, 1907, 1913). Lleva por lo tanto unos cien años en el circuito, así que se ha ganado por méritos propios el adjetivo de clásica. A partir de esos años se produce un rápido desarrollo y para 1950 lo esencial ya está hecho, así que Gulliksen (1950) lleva a cabo la síntesis canónica de este enfoque. Más adelante serán Lord y Novick (1968) quienes lleven a cabo una reformulación de la teoría clásica y abran paso al nuevo enfoque de la TRI que veremos luego. Pero veamos lo esencial del enfoque clásico.

MODELO LINEAL CLÁSICO

Según mi experiencia, tras más de treinta años explicando estas cosas a los estudiantes de psicología, lo que más les cuesta entender es para qué, y por qué, se necesita un modelo o teoría para analizar las puntuaciones de los tests. Pero, ¿donde está el problema?, se preguntan, ahí está el test, ahí están las puntuaciones obtenidas por las personas en el test, unas altas, otras bajas, otras

intermedias, así que adelante, asignemos a cada cual su puntuación. Las cosas no son tan sencillas, el psicólogo, como cualquier otro profesional de otro campo, tiene que asegurarse de que el instrumento que utiliza mide con precisión, con poco error. Y eso mismo vale para cualquier instrumento de medida, bien sea un aparato de la policía para medir la velocidad de los vehículos, el metro para medir las distancias, o el surtidor de la gasolinera para medir los litros de gasolina que nos dispensa. Todos esos instrumentos han de estar homologados, requieren algún indicador del grado de precisión con el que miden, máxime los tests, ya que apoyados en ellos se toman decisiones muy importantes para las vidas de las personas. No es difícil estar de acuerdo en esto, pero el problema es que cuando un psicólogo aplica un test a una persona, o a varias, lo que obtiene son las puntuaciones empíricas que esa persona o personas obtienen en el test, pero eso nada nos dice sobre el grado de precisión de esas puntuaciones, no sabemos si esas puntuaciones empíricas obtenidas se corresponden o no con las puntuaciones que verdaderamente le corresponden a esa persona en la prueba. Bien podría ocurrir que las puntuaciones estuviesen, por ejemplo, algo rebajadas debido a que ese día la persona no está en sus mejores condiciones, o porque las condiciones físicas en las que se desarrolló la aplicación de la prueba no eran las más adecuadas, o porque las relaciones establecidas entre los aplicadores de las pruebas y las personas evaluadas dejaron mucho que desear. Los psicólogos, como les ocurre a los que construyen aparatos dispensadores de gasolina, estamos obligados a garantizar que las puntuaciones de nuestros tests sean precisas, tengan poco error, el problema es que esto no se sabe escrutando directamente las puntuaciones que obtienen las personas en los tests, esas puntuaciones vistas así de frente no nos dicen nada acerca de su grado de precisión. Como no lo podemos hacer así de frente, es por lo que tenemos que dar algunos rodeos, es decir, es por lo que tenemos que plantear algunos modelos que subyacen a las puntuaciones a fin de ser capaces de estimar el grado de precisión de éstas. El error está mezclado con la verdadera puntuación, como la sal en el agua del mar, o el polvo con la paja, y para separarlos necesitamos llevar a cabo algunos procesos y ahí es donde entran las teorías o modelos estadísticos. Modelos para esto ha habido muchos, pero uno de los que se ha mostrado más eficaz y parsimonioso es el modelo lineal clásico propuesto originalmente por Spearman. Entender la lógica y funcionamiento del modelo es muy sencillo, lo que ya es algo más latoso, aunque no difícil, es desarrollar los aspectos formales y deducciones del modelo, lo cual constituye el corpus central de la psicometría, pero para eso ya están los psicómetras, alguien tiene que hacerlo.

¿Qué propuso Spearman a principios del siglo XX que ha tenido tanto éxito en la historia de la Psicología? Spearman propone un modelo muy simple, de sentido común, para las puntuaciones de las personas en los tests, y que ha dado en llamarse modelo lineal clásico. Consiste en asumir que la puntuación que una persona obtiene en un test, que denominamos su puntuación empírica, y que suele designarse con la letra X, está formada por dos componentes, por un lado la puntuación verdadera de esa persona en ese test (V), sea la que sea, y por otro un error (e), que puede ser debido a muchas causas que se nos escapan y que no controlamos. Lo dicho puede expresarse formalmente así: $X = V + e$

Ahora bien, si se ha entendido lo dicho, está justificado decir que con esto poco hemos avanzado, pues si una persona saca en un test 70 puntos de puntuación empírica, el modelo no nos permite saber ni cual es su puntuación verdadera ni el error contenido en esa puntuación. Exactamente así es, tenemos un solo dato, la puntuación empírica (X), y dos incógnitas, la puntuación verdadera (V) y el error (e). Desde ese punto de vista no hemos avanzado nada, tenemos, eso sí, un modelo de puntuación que parece sensato y plausible, pero nada más, y nada menos, pues que el modelo sea plausible es todo lo que se puede pedir para empezar. El error cometido al medir alguna variable con un test (e) puede deberse a muchas razones, que pueden estar en la propia persona, en el contexto, o en el test, una clasificación bastante exhaustiva de las fuentes posibles de error puede consultarse en Stanley (1971). Para poder avanzar Spearman añade tres supuestos al modelo y una definición, veamos cuáles son.

El primer supuesto es definir la puntuación verdadera (V) como la esperanza matemática de la puntuación empírica, que formalmente puede escribirse así: $V = E(X)$. Lo que esto significa conceptualmente es que se define la puntuación verdadera de una persona en un test como aquella puntuación que obtendría como media si se le pasase infinitas veces el test. Se trata de una definición teórica, nadie va a pasar infinitas veces un test a nadie, por razones obvias, pero parece plausible pensar que si

esto se hiciese la puntuación media que esa persona sacase en el test sería su verdadera puntuación.

En el segundo supuesto Spearman asume que no existe relación entre la cuantía de las puntuaciones verdaderas de las personas y el tamaño de los errores que afectan a esas puntuaciones. En otras palabras, que el valor de la puntuación verdadera de una persona no tiene nada que ver con el error que afecta esa puntuación, es decir, puede haber puntuaciones verdaderas altas con errores bajos, o altos, no hay conexión entre el tamaño de la puntuación verdadera y el tamaño de los errores. De nuevo se trata de un supuesto en principio razonable, que formalmente puede expresarse así: r (v,e) = 0.

El tercer supuesto establece que los errores de medida de las personas en un test no están relacionados con los errores de medida en otro test distinto. Es decir, no hay ninguna razón para pensar que los errores cometidos en una ocasión vayan a covariar sistemáticamente con los cometidos en otra ocasión. Formalmente este supuesto puede expresarse así: $r(e_j, e_k)=0$.

Estas asunciones parecen razonables y sensatas, pero no se pueden comprobar empíricamente de forma directa, serán las deducciones que luego se hagan a partir de ellas las que permitan confirmarlas o falsearlas. Tras cien años formuladas y con muchos resultados empíricos detrás, bien podemos decir hoy que las ideas de Spearman han sido de gran utilidad para la psicología.

Además del modelo y de estos tres supuestos, se formula una definición de lo que son Tests Paralelos, entendiendo por ello aquellos tests que miden lo mismo exactamente pero con distintos ítems. Las puntuaciones verdaderas de las personas en los tests paralelos serían las mismas, y también serían iguales las varianzas de los errores de medida.

Pues bien, el modelo lineal, junto con los tres supuestos enunciados, y la definición de tests paralelos propuesta, constituyen el cogollo central de la Teoría Clásica de los Tests. Un curso sistemático de Psicometría consiste en llevar a cabo las deducciones correspondientes para a partir de esos ingredientes llegar a las fórmulas que permiten estimar el grado de error que contienen las puntuaciones de los tests, y que se denomina habitualmente Fiabilidad de los Tests, véase al respecto el trabajo de Prieto y Delgado (2010) en este mismo monográfico. También se obtienen otras fórmulas populares de la psicometría, como la de Spearman-Brown, que permite estimar la fiabilidad de un test cuando se aumenta o disminuye su longitud; o las fórmulas de atenuación que permiten estimar el coeficiente de validez de una prueba si se atenúan los errores de medida, tanto de la prueba como del criterio. Por no hablar de la fórmula que permite estimar los cambios en la fiabilidad de un test cuando varía la variabilidad de la muestra en la que se calcula. En suma, el modelo lineal clásico expuesto, junto con los supuestos asumidos y la definición de tests paralelos están a la base de todas las fórmulas clásicas utilizadas habitualmente por los psicólogos que se valen de los tests en su práctica profesional. Alguien podría decir que para usar estas fórmulas no hace falta saber de donde vienen, ni cual es su fundamento, pero tal aserto no es digno de un psicólogo que se respete a si mismo, a su ciencia, y a su profesión.

De modo que cuando los psicólogos manejan sus coeficientes de fiabilidad y validez para indicar a sus clientes o usuarios en general que los tests que utilizan son precisos, tienen poco error de medida, han de saber que esa estimación de la fiabilidad se puede hacer gracias a este sencillo modelo y a los supuestos planteados hace ya más de cien años.

TEORÍA DE LA GENERALIZABILIDAD Y TESTS REFERIDOS AL CRITERIO

Este enfoque clásico ha generado diversas variantes sobre todo en función del tratamiento dado al error de medida. Ha habido numerosos intentos de estimar los distintos componentes del error, tratando de descomponerlo en sus partes. De todos estos intentos el más conocido y sistemático es la Teoría de la Generalizabilidad (TG) propuesta por Cronbach y sus colaboradores (Cronbach, Gleser, Nanda y Rajaratnam, 1972). Se trata de un modelo de uso complejo, que utiliza el análisis de varianza para la mayoría de sus cálculos y estimaciones.

Otro desarrollo psicométrico surgido en el marco clásico ha sido el de los Tests Referidos al Criterio (TRC). Se trata de tests utilizados fundamentalmente en el ámbito educativo y en la evaluación en contextos laborales. Su objetivo es determinar si las personas dominan un criterio concreto o campo de conocimiento, por tanto no pretenden tanto discriminar entre las personas, como la mayoría de los tests psicológicos, sino evaluar en qué grado conocen un campo de conocimiento denominado criterio, de ahí su nombre. Estos tests se desarrollan a partir de la propuesta de Glaser (1963) y han tenido una gran influencia sobre todo en el ámbito educativo.

Los indicadores psicométricos clásicos desarrollados a partir del modelo lineal clásico no se adaptaban bien a la filosofía de construcción de estos nuevos tests, por lo que se ha desarrollado todo un conjunto de tecnología psicométrica específica para calcular la fiabilidad y validez, así como para establecer los puntos de corte que determinan si una persona domina o no el criterio evaluado (Berk, 1984; Cizek, 2001; Educational Measurement, 1994; Muñiz, 2000).

LIMITACIONES DEL ENFOQUE CLÁSICO

Del enfoque de la teoría clásica bien podría decirse que goza de muy buena salud, hay pocas dudas de su utilidad y eficacia, baste decir, por ejemplo, que la gran mayoría de los tests editados en España, prácticamente todos, están desarrollados y analizados dentro de este marco. Ahora bien, si es así, la pregunta obligada es por qué hacen falta otras teorías de los tests, o, en otras palabras, ¿qué problemas de medición no quedaban bien resueltos dentro del marco clásico para que se propongan nuevas teorías? Pues bien, había dos cuestiones básicas que no encontraban buena solución en la teoría clásica y que hacían que la medición psicológica no fuese homologable a la que exhibían otras ciencias empíricas.

Veamos la primera: dentro del marco clásico, las mediciones no resultan invariantes respecto al instrumento utilizado. Se preguntarán con razón qué quiere decir exactamente esa afirmación un tanto críptica. Es muy sencillo, si un psicólogo evalúa la inteligencia de tres personas distintas con un test diferente para cada persona, los resultados no son comparables, no podemos decir en sentido estricto qué persona es más inteligente. Esto es así porque los resultados de los tres tests no están en la misma escala, cada test tiene la suya propia. Esto puede sorprender a los psicólogos usuarios habituales de la teoría clásica, acostumbrados en la práctica a comparar la inteligencia de personas que han sido evaluadas con distintos tests de inteligencia. Para hacerlo se transforman las puntuaciones directas de los tests en otras baremadas, por ejemplo en percentiles, con lo que se considera que se pueden ya comparar, y de hecho así se hace. Este proceder clásico para solventar el problema de la invarianza no es que sea incorrecto, pero, amén de poco elegante científicamente, descansa sobre un pilar muy frágil, a saber, se asume que los grupos normativos en los que se elaboraron los baremos de los distintos tests son equiparables, lo cual es difícil de garantizar en la práctica. Si eso falla la comparación se viene abajo. No hay duda que lo más deseable científicamente sería que los resultados obtenidos al utilizar distintos instrumentos estuviesen en la misma escala, y todo quedaría resuelto de un plumazo, pues bien, por extraño y contra intuitivo que parezca eso es precisamente lo que va a conseguir el enfoque de la TRI. Este nuevo enfoque de la TRI va a suponer un gran avance para la medición psicológica, propiciando un gran desarrollo de nuevos conceptos y herramientas psicométricas.

La segunda gran cuestión no bien resuelta dentro del marco clásico era la ausencia de invarianza de las propiedades de los tests respecto de las personas utilizadas para estimarlas. En otras palabras, propiedades psicométricas importantes de los tests, tales como la dificultad de los ítems, o la fiabilidad del test, estaban en función del tipo de personas utilizadas para calcularlas, lo cual resulta inadmisible desde el punto de vista de una medición rigurosa. Por ejemplo, la dificultad de los ítems, o los coeficientes de fiabilidad dependen en gran medida del tipo de muestra utilizada para calcularlos. Este problema también encontrará una solución adecuada dentro del marco de la TRI.

Aparte de estas dos grandes cuestiones, había otras menores de carácter más técnico a las que la teoría clásica no daba una buena solución. Por ejemplo, cuando se ofrece un coeficiente de fiabilidad de un test en el marco clásico, como el coeficiente alfa de Cronbach (1951), se está presuponiendo que ese test mide con una fiabilidad determinada a todas las personas evaluadas con el test, cuando tenemos evidencia empírica más que suficiente de que los tests no miden con la misma precisión a todas las personas, dependiendo la precisión en gran medida del nivel de la persona en la variable medida. El nuevo marco de la TRI va a solucionar este problema ofreciendo la Función de Información, que permite estimar la fiabilidad de la prueba en función del nivel de la persona en la variable medida.

Además de estas cuestiones centrales, la TRI va a generar toda una tecnología psicométrica nueva que cambiará para siempre la forma de hacer psicometría; véase por ejemplo en este mismo número monográfico el trabajo de Olea, Abad y Barrada (2010). Ahora bien, conviene dejar muy claro que estos nuevos modelos de TRI de ninguna manera invalidan el enfoque clásico, si bien constituyen un excelente complemento que en determina-

das circunstancias dan solución a problemas mal resueltos en el marco clásico. Ambas tecnologías conviven perfectamente en la construcción y análisis de los tests, igual que coches y aviones lo hacen en el transporte, valga la analogía, unos son aconsejables en determinadas situaciones, y otros lo son en otras.

Veamos los conceptos fundamentales sobre los que se apoyan los modelos de TRI.

TEORÍA DE RESPUESTA A LOS ÍTEMS (TRI)

Como se acaba de señalar en el apartado anterior, la TRI va a resolver algunos graves problemas de la medición psicológica que no encontraban una solución adecuada dentro del marco clásico. Ahora bien, para poder hacerlo tiene que pagar el peaje de formular modelos más complejos y menos intuitivos que el modelo clásico, sin que ello suponga que entrañen dificultades especiales. Pero antes de pasar a exponer los fundamentos de estos modelos, vamos a dar unas breves pinceladas de su nacimiento histórico, para así ayudar al lector a ubicarlos en la historia de la psicología. Quienes estén interesados en una descripción detallada de los aspectos históricos pueden consultar por ejemplo el trabajo de Muñiz y Hambleton (1992), titulado medio siglo de teoría de respuesta a los ítems.

RESEÑA HISTÓRICA

En ciencia pocos avances surgen de repente, de la noche a la mañana, sin incubación, lo más habitual es que se produzca un proceso gradual que en un momento determinado cuaja en una nueva línea de trabajo. Y eso es más o menos lo que ha pasado con la TRI, sus primeros atisbos pueden rastrearse en trabajos pioneros de Thurstone allá por los años veinte (Thurstone, 1925), que se continúan en los cuarenta con las aportaciones de autores como Lawley (1943, 1944) o Tucker (1946). Como se puede ver ya en estos años de pleno dominio de la Teoría Clásica se están dando los primeros pasos de los que luego vendría a denominarse TRI. Esos son los orígenes remotos, pero será el gran psicómetra Frederic Lord (1952) quien en su tesis doctoral dirigida por Gulliksen, el gran sintetizador de la Teoría Clásica, ponga los primeros ladrillos firmes de la TRI. Birnbaum en los años cincuenta aporta nuevos avances, pero será el matemático danés Rasch (1960), quien proponga su hoy famoso modelo logístico de un parámetro. Bien podemos tomar esa fecha como el momento de despegue de la TRI, pero nótese que por estas fechas aún nos movemos a nivel meramente teórico y estadístico, muy lejos de las aplicaciones prácticas de estos nuevos modelos. El gran impulso lo darán Lord y Novick (1968) en su famoso libro, en el cual dedican cinco capítulos al tema. A partir de su libro las investigaciones sobre los modelos de TRI dominarán la psicometría, hasta nuestros días. A partir de esa fecha empiezan a aparecer los programas informáticos necesarios para utilizar los modelos de TRI, tales como BICAL y LOGIST en 1976, BILOG en 1984, MULTILOG, 1983, y otros muchos. En 1980 Lord publicará un influyente libro (Lord, 1980) dedicado a las aplicaciones de la TRI. De esas fechas hasta hoy los avances han sido notorios, y podemos decir que en nuestros días la TRI domina el panorama psicométrico. Una introducción a la TRI en español puede consultarse por ejemplo en Muñiz (1997a), en inglés es muy recomendable el libro de

GRÁFICO 1
CURVA CARACTERÍSTICA DE UN ÍTEM (CCI)

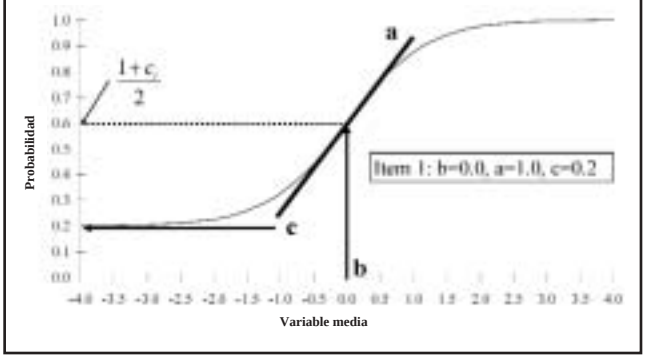
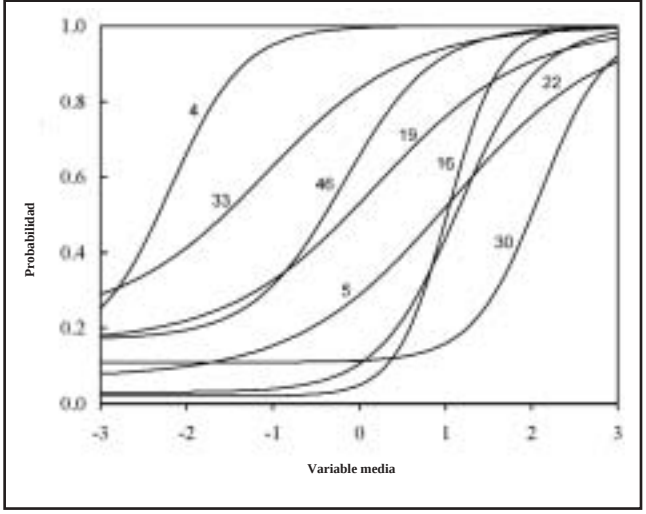


GRÁFICO 2
CURVAS CARACTERÍSTICAS DE OCHO ÍTEMS



Hambleton, Swaminathan y Rogers (1991). Veamos a continuación los supuestos y los modelos de TRI.

SUPUESTOS

Para resolver los problemas citados anteriormente que no encontraban una buena solución dentro del marco clásico, la TRI va a tener que hacer unas asunciones más fuertes y restrictivas que las hechas por la Teoría Clásica. El supuesto clave en los modelos de TRI es que existe una relación funcional entre los valores de la variable que miden los ítems y la probabilidad de acertar estos, denominando a dicha función Curva Característica del Ítem (CCI) (Muñiz, 1997a). Un ejemplo de lo dicho puede verse en el gráfico 1, nótese que al aumentar los valores de la variable medida, denominada θ, aumenta la probabilidad de acertar el ítem $P(\theta)$. Los valores de la variable medida, sea la que sea, se encuentran entre menos infinito y más infinito, mientras que en la teoría clásica los valores dependían de la escala de cada test, yendo desde el valor mínimo obtenible en el test hasta el máximo.

La forma concreta de la CCI viene determinada por el valor que tomen tres parámetros: *a, b* y *c*. Siendo *a* el índice de discriminación del ítem, *b* la dificultad del ítem y *c* la probabilidad que hay de acertar el ítem al azar. Según los parámetros tomen unos valores u otros se generan distintas formas de curvas, como se puede ver en el gráfico 2.

Naturalmente los valores de los parámetros se calculan a partir de los datos obtenidos al aplicar los ítems a una muestra amplia y representativa de personas. Para estos cálculos son necesarios sofisticados programas de ordenador, no en vano los modelos de TRI no se extendieron hasta que se dispuso de ordenadores potentes.

La mayoría de los modelos de TRI, y desde luego los más populares, asumen que los ítems constituyen una sola dimensión, son unidimensionales, por tanto antes de utilizar estos modelos hay que asegurarse de que los datos cumplen esa condición. Esto supone una restricción importante para su uso, pues es bien sabido que muchos de los datos que manejan los psicólogos no son esencialmente unidimensionales, si bien es verdad que los modelos siguen funcionando bastante bien cuando los datos no son estrictamente unidimensionales, es decir son bastante robustos a violaciones moderadas de la unidimensionalidad (Cuesta y Muñiz, 1999).

Un tercer supuesto de los modelos de la TRI es la denominada Independencia Local, que significa que para utilizar estos modelos los ítems han de ser independientes unos de otros, es decir, la respuesta a uno de ellos no puede estar condicionada a la respuesta dada a otros ítems. En realidad si se cumple la unidimensionalidad también se cumple la Independencia Local, por lo que a veces ambos supuestos se tratan conjuntamente.

TABLA 1
DIFERENCIAS ENTRE LA TEORÍA CLÁSICA Y LA TEORÍA DE RESPUESTA A LOS ÍTEMS

Aspectos	Teoría Clásica	Teoría de Respuesta a los Ítems
Modelo	Lineal	No Lineal
Asunciones	Débiles (fáciles de cumplir por los datos)	Fuertes (difíciles de cumplir por los datos)
Invarianza de las mediciones	No	Sí
Invarianza de las propiedades del test	No	Sí
Escala de las puntuaciones	Entre cero y la puntuación máxima en el test	Entre $-\infty$ y $+\infty$
Énfasis	Test	Ítem
Relación Ítem-Test	Sin especificar	Curva Característica del Ítem
Descripción de los ítems	Índices de Dificultad y de Discriminación	Parámetros a, b, c
Errores de medida	Error típico de medida común para toda la muestra	Función de Información (varía según el nivel de aptitud)
Tamaño Muestral	Puede funcionar bien con muestras entre 200 y 500 sujetos aproximadamente	Se recomiendan más de 500 sujetos, aunque depende del modelo

MODELOS

Con los supuestos señalados, según se elija para la Curva Característica de los ítems una función matemática u otra tendremos distintos modelos, por eso se suele hablar de modelos de TRI. Teóricamente habría infinitos posibles modelos, pues funciones matemáticas donde elegir hay de sobra, ahora bien las funciones más utilizadas por razones varias son la función logística y la curva normal. La función logística tiene muchas ventajas sobre la curva normal, pues da resultados similares y sin embargo es mucho más fácil de manejar matemáticamente, así que los tres modelos de TRI más utilizados son los modelos logísticos, que adoptan la función logística como Curva Característica de los ítems. Si sólo se tiene en

cuenta la dificultad de los ítems (parámetro *b*) estamos ante el modelo logístico de un parámetro, o modelo de Rasch, por haber sido propuesto por este autor en 1960 (Rasch, 1960). Si además de la dificultad se tiene en cuenta el índice de discriminación de los ítems (parámetro *a*) estamos ante el modelo logístico de dos parámetros, y si además se añade la probabilidad de acertar el ítem al azar (parámetro *c*), tenemos el modelo logístico de tres parámetros. Este modelo es el más general de los tres, en realidad los otros dos son casos particulares, así cuando el parámetro *c* es cero tenemos el modelo de dos parámetros, y cuando además el parámetro *a* es igual para todos los ítems, se convierte en el modelo de Rasch. Véase a continuación la fórmula del modelo logístico de tres parámetros, donde P(θ) es la probabilidad de acertar el ítem, θ es la puntuación en la variable medida, *a, b* y *c* son los tres parámetros descritos, *e* es la base de los logaritmos neperianos (2,72) y D es una constante que vale 1,7.

$$P(\theta) = c + (1-c)\,[e^{Da(\theta-b)} / (1+e^{Da(\theta-b)})]$$

En la actualidad hay más de cien modelos de TRI, que se utilizan según el tipo de datos manejados, así disponemos de modelos para escalas tipo Likert, para datos dicotómicos, o para datos multidimensionales. Una buena clasificación y revisión de los modelos puede consultarse en el libro de Van der Linden y Hambleton (1997).

COMPARACIÓN DE LA TEORÍA CLÁSICA CON LA TRI

En la tabla 1, tomada de Muñiz (1997a), se sintetizan las diferencias y similitudes entre el enfoque clásico y la TRI.

A MODO DE CONCLUSIÓN

El objetivo de este artículo ha sido el presentar de una manera no técnica a los psicólogos profesionales, lectores de *Papeles del Psicólogo*, las teorías más influyentes en la construcción y análisis de los tests: la Teoría Clásica de los Tests y la Teoría de Respuesta a los Ítems. Espero que estos fundamentos les ayuden a entender e interpretar un poco mejor los datos psicométricos que habitualmente se ofrecen sobre los tests. También sería bueno que ello les animase a refrescar sus conocimientos psicométricos y a profundizar en aspectos nuevos relevantes para su práctica profesional. Todo lo relativo a la medición psicológica ha evolucionado muy rápido en las últimas décadas, produciéndose importantes avances que es necesario seguir de cerca para no quedarse atrás en el ámbito de la evaluación psicológica, pues sin una evaluación precisa y rigurosa no se puede hacer un diagnóstico certero, y sin éste resulta imposible una intervención eficaz.

REFERENCIAS

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

Andrés-Pueyo, A. (1996). *Manual de psicología diferencial*. Madrid: McGraw Hill.

Berk, R. A. (Ed.) (1984). *A guide to criterion referenced test construction.* Baltimore, MD: The Johns Hopkins University Press.

Binet, A. y Simon, T. H. (1905). Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique, 11*, 191-244.

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

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Nueva York: Cambridge University Press.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives*. Londres: LEA:

Colom, B. R. (1995). *Tests, inteligencia y personalidad*. Madrid: Pirámide.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Cronbach, L.J., Gleser, G., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles.* Nueva York: Wiley.

Cuesta, M. y Muñiz, J. (1999). Robustness of item response logistic models to violations of the unidimensionality assumption. *Psicothema,* Vol. 11, 175-182

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

Educational Measurement: Issues and Practice (1994). Número monográfico dedicado a los treinta años de tests referidos al criterio. Vol. 13, nº 4.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Gulliksen, H. (1950). *Theory of mental tests*. Nueva York: Wiley.

Hambleton, R. K., Swaminathan, H., y Rogers, J. (1991). *Fundamentals of item response theory.* Beverly Hills, CA: Sage.

Juan-Espinosa, M. (1997). *Geografía de la inteligencia humana*. Madrid: Pirámide.

Lawley, D. N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edimburg, 61*, 273-287.

Lawley, D. N. (1944). The factorial analysis of multiple item tests. *Proceedings of the Royal Society of Edimburg, 62,* 74-82.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs, nº 7.*

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

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

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

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

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

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

Muñiz, J. (2005). Classical test models. En B. S. Everitt and D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley and Sons. (Vol. 1, pp. 278-282).

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

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J. R. y Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment*, 17(3), 201-211.

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

Muñiz, J. y Hambleton, R. K. (1992). Medio siglo de teoría de respuesta a los ítems. *Anuario de Psicología,* 52(1), 41-66.

Olea, J., Abad, F.J y Barrada, J.R. (2010). Tests informatizados y otros nuevos tipos de tests. *Papeles del Psicólogo*, *31(1)*, 97-107

Pereña, J. (2007). *Una tea en la psicometría española*. Madrid: Tea Ediciones.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo*, *31(1)*, 67-74

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhague: The Danish Institute for Educational Research.

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology, 18*, 161-169.

Spearman, C. (1913). Correlations of sums and differences. *British Journal of Psychology, 5,* 417-426.

Stanley, J. C. (1971). Reliability. En R. L. Thorndike (Ed.), *Educational Measurement.* Washington: American council on Education.

Terman, L. M. (1916). *The measurement of intelligence*. Boston: Houghton Mifflin.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *The Journal of Educational Psychology, 16*, 433-451.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, nº 1.

Thurstone, L. L. y Thurstone. T. G. (1941). Factorial studies of intelligence. *Psychometric Monographs*, nº 2.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. *Psychometrika, 11*, 1-13.

Van der Linden, W. J. y Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. Nueva York: Springer-Verlag.

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

Wissler, C. (1901). Correlation of mental and physical traits. *Psychological Monographs, 3*, nº 16.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 67-74
http://www.cop.es/papeles

FIABILIDAD Y VALIDEZ
RELIABILITY AND VALIDITY

Gerardo Prieto y Ana R. Delgado
Universidad de Salamanca

En este capítulo se describen conceptualmente las propiedades psicométricas de fiabilidad y validez y los procedimientos para evaluarlas. El apartado dedicado a la fiabilidad o precisión de las puntuaciones de las pruebas describe los distintos modelos, procedimientos empíricos e índices estadísticos para cuantificarla. En cuanto a la validez, la propiedad psicométrica más importante y la que ha experimentado mayores transformaciones a lo largo de la historia de la Psicometría, se resumen las principales concepciones y los debates en torno a la misma.
Se previene al lector de dos frecuentes malentendidos: en primer lugar, considerar que la fiabilidad y la validez son características de los tests cuando corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que esos tests proporcionan; en segundo lugar, tratar la fiabilidad y la validez como propiedades que se poseen o no en lugar de entenderlas como una cuestión de grado.
Palabras clave: *Fiabilidad, Psicometría, Tests, Validez.*

The psychometric properties of reliability and validity and the procedures used to assess them are conceptually described in this chapter. The part devoted to the reliability, or test score accuracy, is focused in the models, procedures and statistical indicators most usually employed. As to validity, the most important psychometric property, and the one whose conception has changed the most, we summarize its history in testing contexts.
The reader is prevented that reliability and validity are not, as usually thought, properties of the testing instruments but of the particular inferences made from the scores. Another common error is considering reliability and validity, not as questions of degree, but as absolute properties.
Key words: *Reliability, Psychometrics, Testing, Validity.*

Los psicólogos utilizan diversos procedimientos estandarizados para obtener muestras de la conducta de las personas. Estos recursos, genéricamente denominados *tests*, incluyen un procedimiento de puntuación que permite obtener medidas que pueden ser usadas con distintos propósitos: estimar el nivel de la personas en un constructo (ansiedad, calidad de vida, visualización espacial…), evaluar la competencia tras un periodo de aprendizaje, clasificar a los pacientes en categorías diagnósticas o seleccionar a los aspirantes más aptos para un puesto de trabajo. La legitimidad y eficiencia de estas prácticas depende de su fiabilidad y validez.

En este capítulo se describen, de forma conceptual, estas dos características psicométricas y los procedimientos más frecuentes para evaluarlas. De entrada, hay que prevenir al lector de dos frecuentes malentendidos. El primero consiste en considerar que la fiabilidad y la validez son características de los tests. Por el contrario, corresponden a propiedades de las interpretaciones, inferencias o usos específicos de las medidas que esos tests proporcionan. El segundo se refiere a la consideración de que la fiabilidad y la validez se poseen o no, en lugar de entenderlas como una cuestión de grado (AERA, APA y NCME, 1999).

Correspondencia: Gerardo Prieto. *Facultad de Psicología. Universidad de Salamanca. Avda. De la Merced 109-131, 37005 Salamanca. España. Email: gprieto@usal.es*

FIABILIDAD

La fiabilidad se concibe como la consistencia o estabilidad de las medidas cuando el proceso de medición se repite. Por ejemplo, si las lecturas del peso de una cesta de manzanas varían mucho en sucesivas mediciones efectuadas en las mismas condiciones, se considerará que las medidas son inestables, inconsistentes y poco fiables. La carencia de precisión podría tener consecuencias indeseables en el coste de ese producto en una ocasión determinada. De esta concepción se sigue que de la variabilidad de las puntuaciones obtenidas en repeticiones de la medición puede obtenerse un indicador de la fiabilidad, consistencia o precisión de las medidas. Si la variabilidad de las medidas del objeto es grande, se considerará que los valores son imprecisos y, en consecuencia, poco fiables. De manera semejante, si una persona contestase a un test repetidamente *en las mismas condiciones*, de la variabilidad de las puntuaciones podría obtenerse un indicador de su grado de fiabilidad. La imposibilidad de lograr que las medidas se lleven a cabo *exactamente* en las mismas condiciones es

uno de los problemas de las medición psicológica y educativa. El nivel de atención y de motivación de una persona puede variar al contestar repetidamente a la misma prueba, la dificultad de dos tests pretendidamente iguales construidos para medir el mismo constructo puede ser desigual, las muestras de examinadores que califican un examen de selectividad pueden diferir en el grado de severidad, etc. Por tanto, el esfuerzo de los evaluadores ha de centrarse en estandarizar el procedimiento de medición para minimizar la influencia de aquellas variables extrañas que pueden producir inconsistencias no deseadas. La estandarización del procedimiento implica obtener las medidas en todas las ocasiones en condiciones muy semejantes: con el mismo tiempo de ejecución, las mismas instrucciones, similares ejemplos de práctica, tareas de contenido y dificultad equivalentes, similares criterios de calificación de los evaluadores de exámenes, etc.

El estudio de la fiabilidad parte de la idea de que la puntuación observada en una prueba es un valor concreto de una variable aleatoria consistente en todas las posibles puntuaciones que podrían haber sido obtenidas por una persona en repeticiones del proceso de medida en condiciones semejantes (Haertel, 2006). Obviamente, no es posible repetir la medición un número muy grande de veces a los mismos participantes. Por tanto, la distribución de las puntuaciones es hipotética y sus propiedades deben ser estimadas indirectamente. La media de esa distribución, que reflejaría el nivel de una persona en el atributo de interés, es denominada *puntuación verdadera* en la Teoría Clásica de los Tests (TCT). La TCT es un conjunto articulado de procedimientos psicométricos desarrollados fundamentalmente en la primera mitad del siglo pasado, que se ha utilizado extensivamente para la construcción, análisis y aplicación de los tests psicológicos y educativos. Aunque la TCT surgió en el contexto de la medición de las aptitudes humanas, sus propuestas se extienden a otras áreas. Se asume que la puntuación verdadera de una persona no cambia entre ocasiones, por lo que la variabilidad de las puntuaciones observadas se debe a la influencia de un *error de medida* aleatorio, no sistemático (producido por causas desconocidas e incontrolables en esa situación). La cantidad de error en cada caso sería la diferencia entre una puntuación observada y la puntuación verdadera. La desviación típica de los errores, denominada *error típico de medida* (ETM), indica la precisión de las puntuaciones de una persona, es decir, su variabilidad en torno a la puntuación verdadera. El ETM refleja el error que puede esperarse en una puntuación observada. Por ejemplo, si el error típico de medida del peso de un objeto fuese de dos gramos, se puede aventurar que el peso observado diferirá del peso verdadero en más de dos gramos solo la tercera parte de las veces. Aunque la TCT permite estimar el ETM para personas situadas en distintos rangos de la variable (denominados errores típicos de medida *condicionales*), suele emplearse un único valor aplicable de forma general a todas las puntuaciones de las personas de una población. Obviamente, la valoración del ETM depende de la magnitud de los objetos que se están midiendo: dos gramos es un error despreciable si se pesan objetos muy pesados como sacos de cereales, pero es un error notable si se pesan objetos más livianos como los diamantes. Es decir, el valor del ETM está en las mismas unidades que los objetos medidos y carece de un límite superior estandarizado que facilite su valoración. Por ello, se ha propuesto un índice estandarizado de consistencia o precisión denominado *coeficiente de fiabilidad* que puede oscilar entre 0 y 1. De la TCT se deriva que este coeficiente es el cociente entre la varianza de las puntuaciones verdaderas y la varianza de las puntuaciones observadas en una población de personas. En consecuencia, indica la proporción de la variabilidad de las puntuaciones observadas que *no* puede atribuirse al error de medida; por ejemplo, si el coeficiente de fiabilidad es de 0,80, se considera que el 20% de la variabilidad observada es espuria.

Para estimar empíricamente los estadísticos de fiabilidad (ETM y coeficiente de fiabilidad) se emplean diversos diseños de recogida de datos que reflejan distintas repeticiones del proceso de medida. Los más conocidos se denominan *test-retest* (aplicación de un test a una muestra de personas en dos ocasiones entre las que el atributo se mantiene estable), *formas paralelas* (aplicación a una muestra de personas en la misma ocasión o en distintas ocasiones de dos versiones del test equivalentes en contenido, dificultad, etc), *consistencia entre las partes de una prueba* (división del test en dos subconjuntos equivalentes de ítems o estimación a partir de las covarianzas entre los ítems de la prueba) y *consistencia de las puntuaciones de distintos calificadores* (evaluación de una muestra de conducta por calificadores independientes). La estimación del coeficiente de fiabilidad a partir de estos diseños suele basarse en la correlación entre las puntuaciones observadas obtenidas en las distintas formas de replicación. Existe una extensa bibliografía para obtener una información detallada de estos procedimientos y de los conceptos y desarrollos de la TCT. Excelentes exposiciones pueden encontrarse en este

volumen (Muñiz, 2010) y en los textos de Gulliksen (1950), Martínez-Arias, Hernández-Lloreda y Hernández-Lloreda (2006) y Muñiz (1998).

Además de la TCT, se emplean otros enfoques para cuantificar la fiabilidad de las puntuaciones de los tests: la Teoría de la Generalizabilidad (TG) y la Teoría de Respuesta al Ítem (TRI).

La TCT permite cuantificar solamente dos componentes de la varianza de las puntuaciones observadas: la varianza verdadera y la varianza de error. La TG, concebida como una extensión de la TCT, trata de especificar la contribución a la varianza observada de un número mayor de facetas: la variabilidad entre las personas, las ocasiones en que se mide, las diferentes formas del instrumento, los diferentes calificadores y las interacciones entre los componentes. La estimación de estas influencias se lleva a cabo mediante el análisis de varianza. Los componentes distintos a las diferencias entre personas (formas del test, calificadores, ocasiones, etc) se interpretan como fuentes del error de las medidas, sirviendo como evidencia de las posibles causas del error y permitiendo mejorar los procedimientos de medición. Este modelo es especialmente útil para evaluar la fiabilidad de las calificaciones otorgadas por evaluadores a los productos obtenidos en pruebas o exámenes *abiertos* (los examinados no están constreñidos por un formato cerrado, tal como los de las pruebas de elección múltiple, para emitir sus respuestas). Un tratamiento más exhaustivo puede encontrarse en los textos de Brennan (2001) y en este volumen (Martínez-Arias, 2010).

La TRI es un conjunto de modelos de medida dirigidos a estimar estadísticamente los parámetros de las personas y los ítems en un continuo latente a partir de las respuestas observables. En todos los procedimientos de estimación estadística de parámetros, se cuantifica la cantidad de error de la estimación a partir del error típico (un índice de la variabilidad de los estimadores del parámetro). Cuanto mayor sea el error típico, menor será la precisión de la estimación y mayor será la incertidumbre sobre el valor del parámetro. De forma similar, en los modelos de la TRI la incertidumbre sobre la localización de una persona o un ítem en la variable latente se cuantifica a partir del *error típico de estimación* (ETE) de la persona o del ítem. Este estadístico se diferencia del error típico de medida de las personas correspondiente a la TCT. Como ya se ha expuesto, el ETM es una medida *global* del error, un único valor aplicable de forma general a todas las puntuaciones de las personas de una población, que suele subestimar o sobrestimar el grado de error que afecta a las puntuaciones localizadas en distintos niveles de la variable. Por el contrario, el ETE varía a lo largo de la variable. Por tanto, puede ser considerado una medida *individual* de la precisión, dado que indica la magnitud del error con la que se estiman los parámetros de las personas o los ítems situados en distintas posiciones del continuo latente. La función que describe cómo cambian los valores del ETE de las personas en los distintos niveles de la variable es especialmente útil para determinar los rangos en los que un test es más fiable y para determinar la fiabilidad de los puntos de corte empleados en la clasificación de personas en categorías diagnósticas o de rendimiento.

Puesto que el ETE permite cuantificar un intervalo para estimar el parámetro de una persona, será mayor la incertidumbre sobre su localización cuanto mayor sea el intervalo. Si se adopta la perspectiva opuesta, es decir, de cuánta *certidumbre* se dispone sobre la localización de la persona, entonces se cuantifica la denominada función de información que es análoga al recíproco de la varianza de error condicional de la TCT. La función de información del test indica en qué medida éste permite diferenciar entre las personas en los distintos niveles del atributo. Véase una exposición más detallada en de Ayala (2009).

Terminaremos este apartado con algunas consideraciones prácticas acerca de la interpretación y el uso de los estadísticos de fiabilidad, comenzando por responder a una de las preguntas más frecuentes de los usuarios de las pruebas: ¿qué grado de fiabilidad deben tener las puntuaciones para que su uso sea aceptable? Sin duda, la magnitud requerida depende de las consecuencias derivadas del uso de las puntuaciones. Cuando las puntuaciones vayan a emplearse para tomar decisiones que impliquen consecuencias relevantes para las personas (p. ej., aceptación o rechazo en una selección de personal), el coeficiente de fiabilidad debería ser muy alto (al menos de 0,90). Sin embargo, si se trata de describir las diferencias individuales a nivel de grupo, bastaría con alcanzar valores más modestos (al menos 0,70). No obstante, estas convenciones deben seguirse con cautela: si la evaluación de la fiabilidad se ha llevado a cabo mediante los procedimientos derivados de la TCT, los resultados no habrán de ser necesariamente intercambiables, puesto que los diferentes diseños de recogida de datos antes mencionados (test-retest, formas paralelas, consistencia interna, etc) aprecian distintas fuentes de error: inestabilidad de las medidas, falta de equivalencia de las pruebas, heterogeneidad de los ítems, escasez de concordancia de los evaluadores, etc. Por tanto, es aconsejable disponer de estimaciones de la fiabilidad a partir

de distintos diseños para lograr una mejor comprensión del error que afecta a las puntuaciones (Prieto y Muñiz, 2000). Además, los estadísticos de fiabilidad varían entre poblaciones y están afectados por otras condiciones como la longitud de la prueba y la variabilidad de las muestras de personas. En consecuencia, se ha de evitar el error de considerar que la estimación de la fiabilidad procedente de un único estudio refleja la verdadera y única fiabilidad de la prueba. Los constructores y los usuarios de las pruebas deben informar detalladamente de los métodos de cuantificación, de las características de las muestras y de las condiciones en las que se han obtenido los datos (AERA, APA y NCME, 1999). Como hemos indicado anteriormente, el error típico de medida está expresado en las mismas unidades que las puntuaciones de la prueba. Por ello, es difícil establecer comparaciones entre la fiabilidad de las puntuaciones de distintos tests en base a este estadístico. Por el contrario, la magnitud del coeficiente de fiabilidad oscila siempre entre unos límites estandarizados (0 y 1), por lo que es muy útil para elegir el test más fiable entre los potencialmente utilizables para una aplicación específica. Sin embargo, el error típico de medida aporta más información para describir la precisión de las puntuaciones.

En ocasiones, se utilizan las puntuaciones de los tests, no simplemente para estimar la posición de una persona en la población de interés (denominada interpretación *relativa*), sino para asignarla a una categoría diagnóstica o de rendimiento (patológica/normal, apto/no apto, aceptado/excluido, etc). Para realizar este tipo *absoluto* de interpretaciones, se suelen emplear puntos de corte que guían la clasificación. Puesto que la fiabilidad de las puntuaciones no suele ser la misma en todos los niveles de la variable, conviene conocer el grado de error en las cercanías del punto de corte, dado que si es alto será elevado el número de falsos positivos y negativos en la clasificación. En este caso, es aconsejable emplear la función de error de estimación o de información derivada de los modelos de la TRI.

Terminaremos este apartado analizando la relación entre la fiabilidad y la validez de las puntuaciones, la propiedad que se describe en el siguiente apartado. En la actualidad se considera que la validez, definida como el grado en que las interpretaciones y los usos que se hacen de las puntuaciones están justificados científicamente, es la propiedad psicométrica más importante. Obviamente, la utilidad de unas puntuaciones escasamente fiables para tales fines estará seriamente comprometida. De ahí que se considere la fiabilidad como condición necesaria de la validez. Sin embargo, no será una condición suficiente si las puntuaciones verdaderas, aunque se estimen de manera muy precisa, no resultan apropiadas para conseguir el objetivo de la medida (representar un constructo, predecir un criterio de interés, etc). Es útil tener presente que la fiabilidad es una cuestión relativa a la calidad de los datos, mientras que la validez se refiere a la calidad de la inferencia (Zumbo, 2007).

VALIDEZ

El concepto de validez ha experimentado transformaciones importantes durante el último siglo, provocadas por los diversos objetivos a los que se han destinado los tests. De acuerdo con Kane (2006), entre 1920 y 1950 el uso principal de las pruebas consistió en predecir alguna variable de interés denominada *criterio* (por ejemplo, el rendimiento laboral o académico). En la actualidad este enfoque sigue siendo de suma importancia cuando se emplean las pruebas para seleccionar a los candidatos más aptos para un empleo, en los programas de admisión, en la adscripción de pacientes a tratamientos, etc. En estos casos, la evaluación de la utilidad de la prueba suele cuantificarse mediante la correlación entre sus puntuaciones y las de alguna medida del criterio (*coeficiente de validez*). Sin embargo, el éxito de este tipo de justificación depende de la calidad de la medida del criterio, especialmente de su representatividad (por ejemplo, ¿los indicadores para medir el criterio son suficientes y representativos del puesto de trabajo a desempeñar?). De ahí que el énfasis se desplazase a la justificación de que la puntuación en el criterio procedía de una muestra de indicadores que representase de forma apropiada el dominio o *contenido* a medir (la totalidad de los indicadores posibles). Por tanto, esta fase inicial de desarrollo del concepto terminó con la propuesta de dos vías regias para establecer la validez de las pruebas: la validación de criterio (la correlación entre las puntuaciones del test y las puntuaciones en el criterio) y la validación de contenido (la justificación de que los ítems para medir el criterio son una muestra representativa del contenido a evaluar).

La validación de contenido se extendió desde el análisis del criterio al de la validez de los tests predictores: una prueba no puede considerarse válida si los ítems que la componen no muestrean adecuadamente el contenido a evaluar. La validación de contenido es un enfoque especialmente fértil cuando las facetas del dominio a medir pueden identificarse y definirse claramente. Es éste el caso de los tests dirigidos a evaluar el rendimiento académico que puede especificarse en función de los objetivos de la instrucción (conceptos y ha-

bilidades que un alumno ha de poseer). La metodología de validación descansa fundamentalmente en la evaluación de expertos acerca de la pertinencia y la suficiencia de los ítems, así como de la adecuación de otras características de la prueba como las instrucciones, el tiempo de ejecución, etc. Sin embargo, especificar con precisión el contenido de las manifestaciones de constructos como la extraversión, la memoria de trabajo o la motivación de logro es una tarea más difícil. De ahí que tanto la validación de contenido como la de criterio se considerasen insuficientes para justificar el uso de pruebas dirigidas a evaluar aptitudes cognitivas o atributos de la personalidad. Esta insatisfacción se concretó en el influyente artículo de Cronbach y Meehl (1955) en el que se propone la validación de *constructo* como el modo principal de validación. Cronbach (1971) puntualizó que en un test para medir un rasgo de personalidad no hay únicamente un criterio relevante que predecir, ni un contenido que muestrear. Se dispone, por el contrario, de una teoría acerca del rasgo y de sus relaciones con otros constructos y variables. Si se hipotetiza que la puntuación del test es una manifestación válida del atributo, se puede contrastar la asunción analizando sus relaciones con otras variables. En consecuencia, la validación de constructo puede concebirse como un caso particular de la contrastación de las teorías científicas mediante el método hipotético-deductivo. Aunque el usuario no sea, en general, consciente de ello, las técnicas de medida implican teorías (que se suponen suficientemente corroboradas en el momento de usarlas para contrastar hipótesis científicas o prácticas), por lo que deben venir avaladas ellas mismas por teorías cuyo grado de sofisticación dependerá del momento en que se encuentre el programa de investigación en el que han surgido (Delgado y Prieto, 1997). Dado que una teoría postula una red de relaciones entre constructos y atributos observables, no podremos asumir que las puntuaciones son válidas si la teoría es formalmente incorrecta, las predicciones derivadas de la teoría no se cumplen en los datos empíricos o se han violados otros supuestos auxiliares. Así, desde finales del siglo pasado se ha impuesto la concepción de que la validación de constructo constituye un marco integral para obtener pruebas de la validez, incluyendo las procedentes de la validación de criterio y de contenido (Messick, 1989). El marco de validación se define a partir de teorías en las que se especifican el significado del constructo a evaluar, sus relaciones con otros constructos, sus manifestaciones y sus potenciales aplicaciones e interpretaciones. Además de las pruebas necesarias para justificar una adecuada representación del constructo, Messick incluyó en el marco de validación la justificación de las *consecuencias* del uso de los tests (las implicaciones individuales y sociales). Como se comentará más adelante, la inclusión de la denominada *validación de las consecuencias* es aún objeto de debate. Este breve resumen de la historia del concepto de validez, de la que hemos mencionado algunos hitos importantes, permite comprender los conceptos actuales de validez y validación, de los que destacaremos a continuación sus principales características.

En la actualidad se considera que la *validez* se refiere al grado en que la evidencia empírica y la teoría apoyan la interpretación de las puntuaciones de los tests relacionada con un uso específico (AERA, APA y NCME, 1999). La *validación* es un proceso de acumulación de pruebas para apoyar la interpretación y el uso de las puntuaciones. Por tanto, el objeto de la validación no es el test, sino la interpretación de sus puntuaciones en relación con un objetivo o uso concreto. El proceso de validación se concibe como un *argumento* que parte de una definición explícita de las interpretaciones que se proponen, de su fundamentación teórica, de las predicciones derivadas y de los datos que justificarían científicamente su pertinencia. Dado que las predicciones suelen ser múltiples, una única prueba no puede sustentar un juicio favorable sobre la validez de las interpretaciones propuestas. Son necesarias pruebas múltiples y convergentes obtenidas en diferentes estudios. Por ello, se considera que la validación es un proceso dinámico y abierto. Obviamente, los usos y las interpretaciones relacionadas pueden ser muy variados. Por ello, las fuentes de validación son múltiples y su importancia varía en función de los objetivos. Los *Standards for educational and psychological testing* (AERA, APA y NCME, 1999) se refieren a las más importantes: el contenido del test, los procesos de respuesta, la estructura interna de la prueba, las relaciones con otras variables y las consecuencias derivadas del uso para el que se proponen. Antes de resumir estos enfoques metodológicos, hemos de puntualizar que reflejan distintas facetas de la validez que las engloba como un único concepto integrador. Por tanto, no es riguroso utilizar términos, como *validez predictiva, validez de contenido, factorial*, etc, que inducirían a considerar distintos tipos de validez.

Validación del contenido del test

Los tests están compuestos por un conjunto de ítems destinados a obtener una puntuación que represente el nivel de una persona en un constructo (extraversión,

competencia en matemáticas, etc). Difícilmente se podrá justificar la calidad de las medidas si los ítems no representan de forma suficiente las diferentes facetas de las manifestaciones del constructo. Si eso sucede, el constructo estará *infrarrepresentado* y, en consecuencia, las puntuaciones no alcanzarán el grado de validez requerido. Asimismo, la evidencia de que las respuestas a los ítems están influidas por variables ajenas al constructo de interés constituye una de las principales amenazas a la validez produciendo la denominada *varianza irrelevante al constructo*. También son objeto de la validez de contenido las instrucciones, los ejemplos de práctica, el material de la prueba, el tiempo de ejecución, etc. La consulta a expertos es la vía más usual para apreciar la calidad del contenido, especialmente en ámbitos educativos, aunque cada vez son más empleados los métodos cualitativos basados en la observación directa, las entrevistas o el análisis de archivos. Los procedimientos estandarizados de consulta facilitan la obtención de datos cuantitativos indicativos del porcentaje de ítems de calidad, el porcentaje de las facetas del dominio suficientemente evaluadas, el porcentaje de jueces que han valorado positivamente la calidad de los materiales, la concordancia entre los expertos, etc. Un tratamiento exhaustivo de la validación del contenido puede encontrarse en Sireci (1998).

Análisis de los procesos de respuesta

Debido a la influencia de la ciencia cognitiva, la validación de los tests de inteligencia, aptitudes y rendimiento debe incluir el análisis de los procesos, las estrategias de resolución de problemas y las representaciones mentales que emplean los participantes para resolver los ítems. Se obtendrá evidencia de validez cuando los procesos utilizados se ajustan a los que se postulan en las teorías relativas al constructo medido. La metodología de estudio es muy diversa: entrevistas a los examinados para que describan cómo resuelven las tareas, análisis de los movimientos oculares o tiempos de respuesta, etc. Cuando las teorías acerca del constructo han superado las etapas meramente exploratorias, se pueden construir los tests a partir de un *diseño cognitivo* que especifica ciertos subconjuntos de ítems para suscitar determinados procesos latentes. Las respuestas a los ítems permiten estimar, mediante modelos complejos de la TRI, los parámetros de la persona en los distintos componentes cognitivos de la tarea e identificar *clases* de personas que emplean distintas estrategias de procesamiento. En este enfoque se basan las tendencias más avanzadas del diagnóstico cognitivo (Yang y Embretson, 2007).

Análisis de la estructura interna del test

Algunos tests proporcionan una medida de un solo constructo, otros evalúan varios constructos incluyendo una subescala para cada uno de ellos. El análisis de la estructura interna persigue verificar empíricamente si los ítems se ajustan a la dimensionalidad prevista por el constructor de la prueba. Cuando un test construido inicialmente para evaluar a las personas de una población específica se pretende adaptar a una población diferente (de otra cultura, por ejemplo), es obligado analizar si la estructura interna de la prueba se mantiene invariante. En caso contrario, el significado de las puntuaciones diferirá entre ambas poblaciones. El análisis de la estructura interna del test suele llevarse a cabo con ayuda de los modelos de análisis factorial que se describen en detalle en el artículo de Ferrando y Anguiano (2010) de este monográfico.

Entre los métodos para evaluar la unidimensionalidad de la prueba, ocupa un lugar importante el análisis del *funcionamiento diferencial de los ítems* (DIF). Se podrá aseverar que un test tiene una validez similar en grupos de distinto sexo, cultura, lengua materna, etc., si sus ítems no presentan DIF, como puede leerse en el artículo de Gómez-Benito, Hidalgo y Guilera (2010).

Asociación de las puntuaciones con otras variables

Las relaciones de las puntuaciones del test con otras variables externas a la prueba constituyen una importante fuente de validación. Cuando se emplean las puntuaciones para seleccionar los candidatos más aptos para un empleo, en los programas de admisión, en la adscripción de pacientes a tratamientos, etc, la justificación se basa en su utilidad para predecir un criterio externo. El criterio es una medida de la variable de interés: rendimiento laboral, presencia o ausencia de un trastorno neuropsicológico, calificaciones académicas, etc. La utilidad de la prueba se suele cuantificar mediante la correlación entre sus puntuaciones y las de alguna medida del criterio (*coeficiente de validez*), o mediante otros procedimientos: diferencia en las puntuaciones entre grupos de distinto nivel en el criterio, grado de acuerdo en las clasificaciones en categorías diagnósticas realizadas mediante el test y expertos, etc. La elección de un criterio fiable y válido (suficiente, objetivo y representativo de la conducta de interés) es el punto crítico que determina la bondad del proceso de validación. En función del momento temporal en el que se evalúa el criterio, se distinguen distintos tipos de recogida de datos: *retrospectiva* (el criterio se ha obtenido antes de administrar el test, por ejemplo en base a un diagnóstico clínico anterior), *concurrente* (las puntuaciones del test y del criterio se obtienen en la misma sesión) y *predictiva* (el criterio se mi-

de en un momento posterior). Los resultados entre estos procedimientos pueden diferir: se preferirá el más adecuado al uso que se pretende (por ejemplo, el enfoque predictivo es más apropiado al pronóstico de un rendimiento laboral futuro). De suma importancia es analizar si la utilidad predictiva o diagnóstica se mantiene invariante en distintos grupos de personas. La cuestión de la variabilidad de los resultados en distintos grupos, distintos estudios, diferentes medidas del criterio, etc. afecta a la generalización de la validez de la prueba. El meta-análisis (véase el artículo de Sánchez-Meca y Botella, 2010) permite indagar cómo varían las correlaciones entre el test y el criterio en función de distintas facetas de los estudios.

Cuando las puntuaciones de los tests se usan para estimar el nivel de las personas en un constructo, sus correlaciones con las de otros tests que miden el mismo u otros constructos son de una relevancia especial. Se espera que la asociación entre pruebas que midan el mismo constructo, sean mayores (*validación convergente*) que entre tests que miden constructos diferentes (*validación discriminante*). Para obtener evidencia empírica, se emplean técnicas como el análisis factorial o la matriz multirrasgo-multimétodo (Campbell y Fiske, 1959) en la que se resumen las correlaciones de un test con *marcadores* (tests de validez comprobada) que miden varios constructos a través de distintos métodos.

Validación de las consecuencias del uso de los tests

La última versión de los *Standards for educational and psycological testing* (AERA, APA y NCME, 1999) plantea la previsión de las posibles consecuencias del uso de los tests como parte del proceso de validación. Desde esta perspectiva, el análisis y justificación de las consecuencias ocupan un lugar preponderante cuando los tests vayan a emplearse para tomar decisiones críticas para personas e instituciones: selección, contratación, graduación, promoción profesional, evaluación de programas, etc. La literatura psicométrica denomina estos usos como de *alto riesgo*. Estas prácticas no son ajenas al contexto español: selección de los candidatos a piloto, al ejército profesional y los cuerpos de seguridad, oposiciones para ingresar en diversas instituciones y empresas, exámenes universitarios, pruebas de selectividad, evaluación del profesorado universitario, evaluación del grado de dependencia, obtención del permiso de armas y del carnet de conducir, etc. En estos casos, la pertinencia del uso no se limita a la comprobación de que las puntuaciones representan adecuadamente los constructos y a la justificación teórica de la red nomológica que vincula los constructos con los criterios de interés. Las aplicaciones de alto riesgo tienen efectos colaterales de carácter personal y social. Citemos como ejemplo de los primeros el efecto en la validez de las puntuaciones del entrenamiento y aprendizaje de los tests que suelen seguir muchas de las personas que se presentan a programas de selección. ¿Hasta qué punto son sensibles las pruebas a este tipo de manipulación? Existen otros efectos de carácter institucional tales como la peculiaridad del uso de los tests en un contexto social. Piénsese en el fraude social relacionado con el uso de las pruebas psicotécnicas que se emplean en nuestro país para otorgar el permiso de armas o el de conducir. Si pensamos en las consecuencias, ¿podríamos decir que ejercen su función? Está claro que si la validez se refiere al grado en que la teoría y la evidencia empírica apoyan la interpretación de las puntuaciones de los tests en relación con un uso específico, las consecuencias no pueden ser ajenas al proceso de validación.

Aunque parece existir un cierto consenso sobre esta cuestión, también existen voces discordantes. Por ejemplo, Borsboom y Mellenberg (2007) consideran que el concepto de validez debe limitarse a un ámbito más restringido que el de la amplia definición incluida en las propuestas de Messick (1989) y en los actuales *Standards*. A su juicio, la validación debe limitarse a contrastar si existe una relación causal entre el constructo y las puntuaciones del test; las interpretaciones de las puntuaciones en contextos aplicados (selección de personal, acreditación, etc) y el impacto social del uso de las pruebas serían ajenas, *stricto sensu*, al ámbito de la validez. Si bien esta postura simplificadora parece libre de problemas, definir la validez de constructo como la validez de la inferencia causal implica identificarla con la validez interna de la evidencia a favor del constructo (para una versión actualizada de los distintos tipos de validez en los diseños experimentales véase Shadish, Cook y Campbell, 2002). Esta identificación podría, tal vez, justificarse en programas de investigación básica ya avanzados, pero imposibilitaría en la práctica la mayor parte de las aplicaciones psicológicas, y esto sin tener en cuenta los conocidos problemas del concepto de causación. De ahí que el pragmatismo nos lleve a preferir una postura más flexible, la que considera que los procedimientos de validación han de servir para apoyar la inferencia a la mejor explicación posible, incluyendo la evidencia aportada por los diversos métodos cualitativos y cuantitativos a disposición de los psicómetras en cada momento (Zumbo, 2007). Si se considera que la validación es un proceso abierto en el tiempo, la validez es necesariamente una cuestión de grado, como señalan los *Standards*, algo que, por otra parte, es común a los distintos conceptos de validez empleados por los epistemólogos.

El debate sobre la inclusión de las consecuencias en el concepto de validez no es un tecnicismo que preocupe solo a los sesudos teóricos de la psicometría. Tomar partido por la inclusión conlleva responsabilidades: ¿pueden y deben los constructores de las pruebas aventurar las consecuencias deseables e indeseables de su uso? ¿qué repertorio metodológico usar para ello? ¿en qué instancia recae el análisis y la justificación de las consecuencias? Estas y otras cuestiones relacionadas seguirán alimentando el debate y la generación de propuestas: una excelente revisión sobre la validación de las consecuencias puede consultarse en Padilla, Gómez, Hidalgo y Muñiz (2007).

Para terminar, un comentario terminológico: acorde con la trayectoria del uso de los tests en contextos anglosajones, *validation* tiene en inglés un significado legal: "declarar legalmente válido". Por el contrario, en nuestra lengua, el término validación tiene dos significados: "acción y efecto de validar", que comparte con el idioma inglés, y "firmeza, fuerza, seguridad o subsistencia de algún acto". Aunque solemos referirnos a la primera acepción, la más aséptica, es la segunda la que más se acerca al objetivo que persigue la investigación psicológica en su variante psicométrica.

REFERENCIAS

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

Borsboom, D. y Mellenberg, G.J. (2007). Test Validity in Cognitive Assessment. En J.P. Leighton y M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 85-115). Cambridge: Cambridge University Press.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer-Verlag.

de Ayala, R.J. (2009). *The Theory and Practice of Item Response Theory*. New York: The Guilford Press.

Campbell, D.T. y Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cronbach, L. J. (1971). Test validation. En R.L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L.J. y Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Delgado, A.R. y Prieto, G. (1997). *Introducción a los métodos de investigación de la psicología.* Madrid: Pirámide.

Gómez-Benito, J., Hidalgo, M.D. y Guilera, G. (2010). El sesgo de los instrumentos de medición. Tests justos. *Papeles del Psicólogo*, *31(1)*, 75-84.

Gulliksen, H. (1950). *Theory of mental tests*. New York, Wiley.

Haertel, E. H. (2006). Reliability. En R.L. Brennan (Ed.), *Educational Measurement* (pp. 65-110). Wesport, CT: American Council on Education and Praeger Publishers.

Kane, M.T. (2006). Validation. En R.L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Wesport, CT: American Council on Education and Praeger Publishers.

Martínez-Arias, M.R. (2010), Evaluación del desempeño. *Papeles del Psicólogo*, *31(1)*, 85-96.

Martinez-Arias, M.R., Hernández-Lloreda, M.J. y Hernández-Lloreda, M.V. (2006). *Psicometría*. Madrid: Alianza Editorial.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational measurement* (pp- 13-103). New York: American Council on Education.

Muñiz, J. (1998). *Teoría Clásica de los Tests*. Madrid: Pirámide.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo*, *31(1),* 57-66.

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

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-71.

Sánchez-Meca, J. y Botella, J. (2010). Revisiones sistemáticas y meta-análisis: herramientas para la práctica profesional. *Papeles del Psicólogo*, *31(1)*, 7-17.

Shadish, W.R., Cook, T.D., y Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

Sireci, S.G. (1998). The construct of content validity. En Zumbo, B.D. (Ed.), *Validity Theory and the Methods Used in Validation: Perspectives From the Social and Behavioral Sciences* (pp. 83-117). Kluwer Academic Press, The Netherlands.

Yang, X. y Embretson, S.E. (2007). Construct Validity and Cognitive Diagnostic Assessment. En J.P. Leighton y M.J. Gierl (Eds.), *Cognitive Diagnostic Assessment for Education* (pp. 119-145). Cambridge: Cambridge University Press.

Zumbo, B.D. (2007). Validity: Foundational Issues and Statistical Methodology. In C.R. Rao and S. Sinharay (Eds.) *Handbook of Statistics, Vol. 26: Psychometrics,* (pp. 45-79). Elsevier Science B.V.: The Netherlands.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 75-84
http://www.cop.es/papeles

EL SESGO DE LOS INSTRUMENTOS DE MEDICIÓN. TESTS JUSTOS

BIAS IN MEASUREMENT INSTRUMENTS. FAIR TESTS

Juana Gómez-Benito[1], M. Dolores Hidalgo[2] y Georgina Guilera[1]
[1]*Universidad de Barcelona.* [2]*Universidad de Murcia*

Las evaluaciones psicológicas deben garantizar la equidad y validez de las interpretaciones y decisiones adoptadas a partir de las mismas. Para ello es necesario la utilización de instrumentos libres de sesgo, y capaces de evaluar necesidades personales y sociales de individuos con diferentes características. El estudio sobre el posible sesgo de los tests, o de parte de sus ítems, ha ocupado un lugar relevante en la investigación psicométrica de los últimos 30 años y es previsible que siga constituyendo un importante foco de interés para los profesionales e investigadores implicados en la evaluación mediante el uso de los tests. Este trabajo pretende abordar esta perspectiva ofreciendo al psicólogo aplicado unas directrices y un bagaje de conocimientos sobre los conceptos de sesgo, funcionamiento diferencial e impacto, los procedimientos de detección de ítems o tests sesgados y la evaluación de sus posibles causas para, en conjunto, mejorar la validez de las mediciones psicológicas.
Palabras clave: *Sesgo, Funcionamiento diferencial del ítem, Procedimientos de detección, Tests justos, Validez.*

Psychological assessment must ensure the equity and validity of interpretations and of any decisions taken as a result of them. That is it necessary the use of bias-free assessment instruments those are capable of evaluating the personal and social needs of individuals with different characteristics. The study about the possible bias of tests, or some of their items, has had great relevance in psychometric research for the last 30 years and it will probably continue to be an important focus of interest for professionals and researchers involved in psychological and educational testing. The aim of this paper is providing to the applied psychologist the background about bias, differential functioning and impact concepts, item or tests bias detection procedures and evaluation of its possible causes and, therefore, for improving the validity of psychological measurement.
Key words: *Bias, Differential Item Functioning (DIF), Detection procedures, Fair tests, Validity.*

Los tests constituyen uno de los instrumentos de medida estandarizados más empleados en las ciencias sociales y de la salud, especialmente en psicología y educación. No hay que olvidar que un test se administra con un objetivo concreto, generalmente para tomar decisiones que en la mayoría de ocasiones son relevantes para la vida del individuo receptor. Así por ejemplo, en España se emplean tests para entrar a los cuerpos de seguridad, conseguir un puesto de trabajo, superar una materia en la universidad, formar parte de un programa de intervención, entre otros. Por lo tanto, es de extrema importancia que los profesionales que emplean este tipo de instrumentos se cercioren de que garantizan la igualdad de oportunidades y el tratamiento equitativo de los individuos a los que se les administra el test en cuestión, en otras palabras, que el test sea justo en las decisiones que de él se derivan.

Pero, ¿cuándo podemos afirmar que un test es justo?

Decidir hasta qué punto un test está siendo justo en su medición no es tarea fácil. Aspectos como el contexto sociocultural, el proceso de construcción y/o adaptación, las condiciones de aplicación, la interpretación de las puntuaciones y el grado de formación del profesional (Muñiz y Hambleton, 1996) pueden ocasionar que el test sea injusto en su aplicación. Como afirman estos autores, la mayoría de los problemas en torno a los tests provienen de su uso inadecuado, más que del test en sí, de su construcción, o de sus propiedades técnicas. Por lo tanto, asumiendo que las dos primeras cuestiones están solventadas, el interés se traslada a las propiedades técnicas o psicométricas del test.

Correspondencia: Juana Gómez-Benito. *Dpto. Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Universidad de Barcelona. Paseo Valle Hebrón, 171. 08035-Barcelona. España. E-mail: juanagomez@ub.edu*

SESGO, IMPACTO Y DIF

En este contexto, la presencia de un posible sesgo en los ítems que componen el test es una preocupación central en la evaluación de la validez de los instrumentos de medida, entendiendo por validez el grado en que la evidencia empírica y el razonamiento teórico apoyan la adecuación e idoneidad de las interpretaciones basadas en las puntuaciones de acuerdo con los usos propuestos por el test (Messick, 1989; Prieto y Delgado, 2010). Así pues, cuando afirmamos que un test determinado es váli-

do, lo que realmente estamos diciendo es que la puntuación obtenida tiene un significado específico, asumiendo que este significado es el mismo en los distintos grupos para los cuales el test ha sido validado. No obstante, para garantizar que una puntuación de un test presenta el mismo significado en diversos grupos, se requieren numerosos estudios que evalúen distintas evidencias de la validez del test (APA, AERA, y NCME, 1999). La existencia de sesgo en los instrumentos de medida psicológicos puede representar una seria amenaza contra la validez de dichos instrumentos en los que algunos de sus ítems están beneficiando a ciertos grupos de la población en detrimento de otros de igual nivel en el rasgo que interesa medir. De modo complementario, el hecho de que no haya sesgo de los ítems representa una evidencia del grado de generalización de las interpretaciones basadas en las puntuaciones del tests para distintos subgrupos de una o varias poblaciones.

El tema del sesgo ha acaparado la atención de los investigadores y profesionales, especialmente desde la polémica generada por los estudios de Jensen (1969, 1980). Este autor consideró que la inteligencia era hereditaria y que, por tanto, las diferencias que se observaban entre grupos raciales eran atribuibles a la genética. Evidentemente esta afirmación activó efusivas discusiones entre genetistas y ambientalistas. Estos últimos defendían que la explicación de las diferencias entre los grupos había que buscarla en el posible sesgo cultural de los tests de inteligencia. En ese momento, el papel de los psicómetras se centró en averiguar hasta qué punto las diferencias entre grupos eran debidas a características reales de los individuos de cada grupo o a artefactos generados por el propio instrumento. Este debate generó un nuevo conflicto semántico: ¿sesgo cultural o propiedades psicométricas distintas?

El sesgo se refiere a la injusticia derivada de uno o varios ítems del test al comparar distintos grupos que se produce como consecuencia de la existencia de alguna característica del ítem o del contexto de aplicación del test que es irrelevante para el atributo medido por el ítem, mientras que el segundo hace referencia únicamente a las características psicométricas del ítem. En la actualidad se ha llegado al acuerdo que el término sesgo asume que se conocen o se investigan las causas por las cuales determinados ítems presentan un comportamiento diferencial en función de ciertas variables, cuando en la mayoría de estudios lo único que se puede inferir es si existen diferencias en los resultados conseguidos por distintos individuos igualmente capaces. El término adecuado para este último tipo de resultados, que únicamente hace referencia a las propiedades psicométricas, es funcionamiento diferencial del ítem (Differential Item Functioning, DIF), denominación adoptada a raíz de la publicación de Holland y Thayer (1988) con la finalidad de distinguir entre ambos conceptos.

Formalmente se afirma que un determinado ítem presenta DIF si a nivel psicométrico se comporta diferencialmente para diversos grupos, es decir, el DIF indica una diferencia del funcionamiento del ítem (o test) entre grupos comparables de examinados, entendiendo por comparables aquellos grupos que han sido igualados respecto al constructo o rasgo medido por el test (Potenza y Dorans, 1995). En otras palabras, un ítem presenta DIF cuando grupos igualmente capaces presentan una probabilidad distinta de responderlo con éxito o en una determinada dirección en función del grupo al que pertenecen. En la terminología propia del DIF se denomina *grupo focal* al conjunto de individuos, generalmente minoritario, que representa el foco de interés del estudio y que normalmente es el grupo desaventajado, mientras que el *grupo de referencia*, generalmente mayoritario, se refiere a un grupo de individuos estándar respecto al cual se compara el grupo focal. Sin embargo, el hecho que un instrumento de medida obtenga resultados sistemáticamente inferiores en un grupo en comparación a otro no necesariamente implica la presencia de DIF, sino que pueden existir diferencias reales entre los grupos en el rasgo medido por el test en cuestión. En este caso se habla de impacto (Camilli y Shepard, 1994) o diferencias válidas (van de Vijver y Leung, 1997).

Una vez aclarada la diferencia entre sesgo, DIF e impacto, imaginemos que estamos estudiando un ítem potencialmente sesgado en contra de un grupo minoritario. ¿Cómo podemos evaluar la presencia de DIF? La lógica probablemente nos llevaría a comparar directamente las puntuaciones del ítem del grupo minoritario frente al resto de examinados, y si se observasen diferencias diríamos que el ítem está siendo injusto con uno de los grupos. Sin embargo, no podemos tener la certeza de si las diferencias provienen del sesgo del ítem o realmente el nivel de habilidad de un grupo y otro son distintos. El concepto de DIF pretende abordar esta cuestión, por este motivo los análisis de DIF comparan las respuestas al ítem entre los grupos únicamente cuando éstos han sido igualados en el nivel de habilidad o del rasgo medido mediante un criterio de igualación. En este sentido, es imprescindible disponer de un criterio libre de sesgo; no obstante, en la mayoría de situaciones la única evidencia

empírica de equiparación o igualación de que se dispone es el propio test (generalmente la puntuación total), que se encuentra contaminada por la presencia de ítems con DIF y que forman parte del criterio juntamente con los ítems sin DIF. Por lo tanto, un problema endémico a los métodos de detección del DIF reside en que adolece de una cierta circularidad en su forma de proceder ya que el ítem estudiado también contribuye a la definición de la variable de igualación de los grupos. Para reducir el efecto producido por los ítems con funcionamiento diferencial, se han propuesto algunas técnicas de purificación que, en dos etapas o iterativamente, eliminan del criterio aquellos ítems que previamente han sido detectados con DIF (French y Maller, 2007, Gómez-Benito y Navas, 1996; Hidalgo y Gómez-Benito, 2003; Holland y Thayer, 1988; Navas-Ara y Gómez-Benito, 2002; Wang, Shih y Yang, 2009).

TIPOS DE DIF

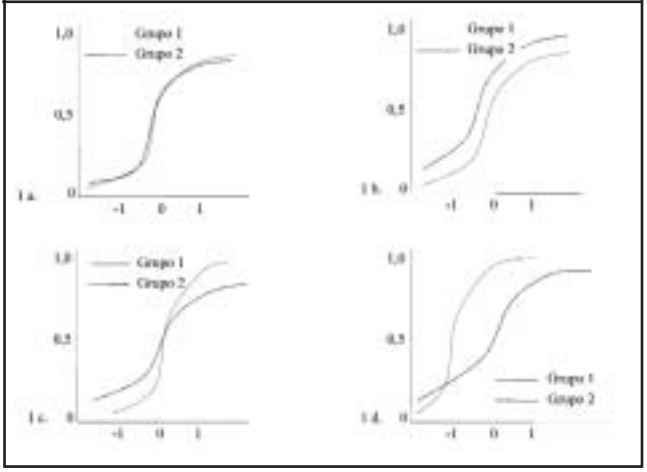
Aunque existen diversas taxonomias del DIF (ver Hessen, 2003), una clasificación muy extendida por su simplicidad proviene de Mellenbergh (1982). Este autor distingue dos tipos de DIF en función de la existencia o no de interacción entre el nivel en el atributo medido y el grupo de pertenencia de los individuos. En el denominado *uniforme* no existe interacción entre el nivel del rasgo medido y la pertenencia a un determinado grupo ya que la probabilidad de responder correctamente (o en una determinada dirección) al ítem es mayor para un grupo que para el otro de forma uniforme a lo largo de todos los niveles del rasgo. En el caso del DIF *no uniforme* sí que existe tal interacción, por lo que la probabilidad de cada grupo de responder correctamente (o en una determinada dirección) al ítem no es la misma a lo largo de los diferentes niveles del rasgo medido.

En el marco de la teoría de respuesta al ítem (véase Muñiz (2010) en este mismo número) se propone el concepto de *curva característica del ítem* (Item Characteristic Curve, ICC), de gran utilidad para entender gráficamente los diversos tipos de DIF. En ítems de respuesta dicotómica, la ICC relaciona la probabilidad de acertar el ítem (eje de ordenadas en el gráfico) con el nivel de los individuos en la variable medida o habilidad (eje de abscisas). De esta forma, un ítem no presenta DIF si su curva característica para el grupo focal y para el grupo de referencia coinciden (figura 1a), situación que se da cuando tanto el parámetro de dificultad (posición de la ICC en la escala de habilidad) como el de discriminación (proporcional a la pendiente de la ICC) presentan el mismo valor en ambos grupos. El ítem muestra DIF *uniforme* si las respectivas ICCs no se cruzan en ningún nivel de la variable medida (figura 1b), hecho que se da cuando los parámetros de dificultad son distintos, pero los correspondientes parámetros de discriminación se mantienen iguales en ambos grupos. Finalmente, presenta DIF no uniforme si en algún punto las ICCs se cruzan. En este último caso, Swaminathan y Rogers (1990) establecen una segunda subdivisión. El DIF *no uniforme simétrico* quedaría representado por un cruzamiento central de las ICCs en el nivel de habilidad (figura 1c) y se da cuando el parámetro de dificultad se mantiene constante y el parámetro de discriminación varía entre los dos grupos, mientras que el DIF *no uniforme mixto* se da cuando los parámetros de dificultad y discriminación son distintos en los dos grupos y viene representado por un cruzamiento asimétrico de las ICCs del grupo focal y de referencia (figura 1d).

PROCEDIMIENTOS DE DETECCIÓN

Desde finales de los años 80 y durante toda la década de los 90, la elaboración y análisis de métodos y técnicas estadísticas para la detección y evaluación del DIF ha concentrado los esfuerzos de investigadores, y ha incrementado paulatinamente la sofisticación de los procedimientos utilizados. Su principal reto metodológico ha sido desarrollar procedimientos que, por un lado, sean sensibles a la detección tanto del DIF uniforme como no uniforme y, por otro lado, no confundan el DIF con el impacto. Además, como respuesta a la demanda progresiva de técnicas aplicables a ítems politómicos (como las

FIGURA 1
ICC DE UN ÍTEM SIN DIF (1a), CON DIF UNIFORME (1b), CON DIF NO UNIFORME SIMÉTRICO (1c) Y CON DIF NO UNIFORME MIXTO (1d)



escalas tipo Likert), el interés se ha trasladado también al desarrollo de procedimientos útiles para este tipo de formato de respuesta, generalmente provenientes de extensiones de sus homólogos para ítems dicotómicos.

Teniendo en cuenta esta primera distinción sobre la naturaleza de respuesta al ítem (dicotómica/politómica), Potenza y Dorans (1995) clasifican los diferentes métodos en función del tipo de criterio de igualación de los grupos (puntuación observada/variable latente) y de la relación entre la puntuación en el ítem y la variable de igualación (paramétrica/no paramétrica). Basándose en esta taxonomía, Hidalgo y Gómez-Benito (2010) ofrecen una clasificación de todos los procedimientos actuales de detección del DIF.

En primer lugar, se puede estimar el nivel de habilidad de los individuos siguiendo dos estrategias: la primera, el *método de la variable latente* utiliza una estimación de la habilidad latente en el marco de la teoría de respuesta al ítem (TRI) mientras que el *método de la puntuación observada* consiste en utilizar la puntuación total observada del test. Un segundo criterio reside en cómo se estima la puntuación al ítem en cada uno de los niveles de habilidad. Una forma de proceder consiste en utilizar una función matemática que relacione la puntuación del ítem con el nivel de habilidad, como las ICCs de la figura 1, que representan gráficamente la probabilidad de obtener una determinada puntuación en el ítem en función del nivel de habilidad de los individuos. Como se ha comentado, diferencias en las ICCs de los grupos indican DIF y para que esto ocurra los parámetros que definen las correspondientes ICCs han de ser diferentes. En consecuencia, dado que las curvas vienen determinadas por uno o más parámetros en la función matemática, esta aproximación se denomina *método paramétrico*. En cambio, la segunda estrategia no utiliza ninguna función matemática para relacionar la respuesta al ítem con el nivel de habilidad, sino que simplemente tiene en consideración la puntuación observada al ítem en cada uno de los niveles de habilidad para cada grupo. En este caso, la presencia de DIF vendrá determinada por la obtención de diferencias entre grupos en la puntuación observada, sin tener en cuenta ningún modelo matemático (y, por tanto, tampoco parámetros). Por este motivo esta aproximación se conoce como el *método no paramétrico*. En tercer lugar se considera la naturaleza del tipo de respuesta, dicotómica o politómica. Dado que en el caso de ítems politómicos el DIF puede estar presente en las diferentes categorías de respuesta de un mismo ítem, y no necesariamente en la misma dirección ni en todas las categorías, las técnicas para ítems dicotómicos son siempre más sencillas computacional y conceptualmente que las extensiones para ítems de respuesta politómica.

Las técnicas que emplean la puntuación observada en el test como variable de igualación, asumiendo que esta puntuación es una estimación adecuada de la habilidad latente del individuo, pueden resultar imprecisas en la detección del DIF principalmente cuando el test contiene ítems de discriminación diversa, mientras que los métodos de la variable latente superan este inconveniente a base de incrementar la sofisticación de los modelos matemáticos de estimación de la habilidad. Una ventaja de los métodos no paramétricos, como el Mantel-Haenszel (MH) y el SIBTEST, es que los supuestos del modelo son escasos, por lo que el DIF no suele confundirse con la falta de ajuste del modelo. En el caso de los métodos paramétricos, como los procedimientos basados en la TRI, es necesario asegurar una adecuada estimación de los parámetros del test precisamente para evitar esta confusión, por tanto, se requieren tamaños muestrales del grupo de referencia y focal mucho más elevados que con los modelos no paramétricos.

Existe un abanico de programas informáticos que permite implementar la mayoría de procedimientos de detección del DIF. La mayor parte consisten en programas que han sido diseñados específicamente para la detección de los ítems con DIF, como MHDIF (Fidalgo, 1994), EZDIF (Waller, 1998a), DIFAS (Penfield, 2005) o EASY-DIF (González, Padilla, Hidalgo, Gómez-Benito y Benítez, 2009) para el procedimiento MH y de libre distribución poniéndose en contacto con los autores del programa; DIF/DBF (Stout y Roussos, 1999) para el procedimiento SIBTEST, que se distribuye mediante Assessment System Corporation; RLDIF (Gómez-Benito, Hidalgo, Padilla, y González, 2005) para el procedimiento de Regresión Logística (RL), actualmente en proceso de comercialización; y IRTLRDIF (Thissen, 2001), TESTGRAPH (Ramsay, 2000) y LINKDIF (Waller, 1998b) para procedimientos basados en la TRI, también de libre distribución. Pueden utilizarse también recursos provenientes de programas estándares de análisis estadístico, que requieren licencia de uso, como por ejemplo SPSS (SPSS Inc., 2009) para MH y RL, LISREL (Jöreskog y Sörbom, 2006) o MPLUS (Muthén y Muthén, 1998, 2007) para procedimientos basados en modelos de ecuaciones estructurales.

Con la finalidad de estudiar su comportamiento, individual y comparativamente, numerosos trabajos se han

aproximado al estudio de las técnicas de detección del DIF mediante la simulación de datos, tanto en ítems dicotómicos como de respuesta politómica. Estos estudios básicamente analizan la variación en la tasa de falsos positivos o error Tipo I (detectar un ítem con DIF cuando en realidad no lo presenta) y de detecciones correctas o potencia estadística (identificar un ítem con DIF cuando realmente lo presenta) bajo diferentes condiciones de simulación, manipulando aquellas variables que supuestamente pueden modular las correspondientes tasas de detección (por ejemplo, el tamaño muestral, la contaminación del test o el tipo de DIF, entre otras) y observando los cambios producidos en ellas. Generalmente terminan el estudio delineando sugerencias y recomendaciones sobre las condiciones bajo las cuales el procedimiento en cuestión presenta un control de la tasa de error Tipo I y una adecuada potencia estadística. Una cuestión común a prácticamente la totalidad de estos estudios es que se centran en la detección del DIF en un único ítem. Hay que tener en cuenta que un test obviamente está compuesto por un conjunto de ítems, y que la dirección del DIF en los diversos ítems de un mismo test puede ser distinta (algunos pueden favorecer al grupo focal y otros al de referencia), de tal forma que los efectos individuales del DIF de los ítems se cancelen cuando se considera el test en global. Por tanto, en ocasiones, es interesante evaluar lo que se denomina el funcionamiento diferencial del test (Differential Test Functioning, DTF) o explorar el DIF en un subconjunto de ítems. En este contexto, algunas técnicas han procurado abordar específicamente el estudio del DIF en tests o conjuntos de ítems, como son el SIBTEST en ítems dicotómicos y el POLYSIBTEST en ítems politómicos, o la aproximación desde la TRI propuesta por Raju y su equipo de investigación (Oshima, Raju, y Nanda, 2006)

TAMAÑO DEL EFECTO

Otro tipo de estudios, también basados en la simulación de datos, aconsejan la inclusión de medidas del tamaño del efecto como complemento o alternativa a las pruebas de significación, a fin de poder evaluar la magnitud del efecto observado y comparar resultados obtenidos en diferentes estudios. Hay que tener en cuenta que detectar un ítem con DIF mediante una prueba de significación estadística no necesariamente implica que su efecto sea destacable, es decir, puede que su efecto sea de escasa relevancia. En este sentido, es importante examinar la magnitud del DIF porque los efectos de la presencia de ítems con DIF pueden ser triviales, cancelarse o pueden realmente poner en duda las decisiones basadas en el test. La mayor parte de las técnicas de detección del DIF han propuesto diversas medidas. Por poner un ejemplo, Dorans y Holland (1993) presentan el estadístico Delta-DIF para el procedimiento Mantel-Haenszel, y dentro del ámbito de la regresión logística (RL) Zumbo y Thomas (1997) han sugerido el incremento en R^2; Gómez-Benito e Hidalgo (2007) y Monahan, McHorney, Stump y Perkins (2007) han propuesto el uso de la odds-ratio como medida del tamaño del efecto usando RL para ítems dicotómicos e Hidalgo, Gómez-Benito y Zumbo (2008) para ítems politómicos. Por norma general, estos trabajos establecen directrices o proponen criterios de clasificación que permiten interpretar los valores de la magnitud del DIF (no únicamente la presencia o ausencia de DIF) siguiendo las directrices de clasificación del *Educational Testing Service* que establece tres categorías, a saber: DIF insignificante (categoría A), DIF moderado (categoría B) y DIF elevado (categoría C). Los ítems que se clasifiquen como tipo C deben ser revisados y eliminados del test, por el contrario los ítems clasificados como tipo A y/o B pueden ser mantenidos en el test.

ELECCIÓN DE TÉCNICA

Aunque los avances en el desarrollo y la optimización de los métodos de detección han sido considerables, la idoneidad de aplicar un procedimiento determinado en una situación concreta todavía está llena de interrogantes. En este entramado de trabajos y técnicas, la duda suele trasladarse a la siguiente cuestión: ¿qué procedimientos empleamos con nuestros datos? La decisión de aplicar una técnica u otra suele basarse en diversos aspectos, dado que no existe hasta el momento ningún método que sea adecuado en la totalidad de situaciones. Se suelen tener en cuenta las diferencias en las distribuciones de habilidad de los grupos de referencia y focal, el tamaño muestral de ambos grupos, el tipo de DIF, la simplicidad computacional y disponibilidad de programas informáticos, y el criterio de igualación de los grupos, entre otros. Y esta complejidad ha llevado a varios autores a pensar que la opción más conservadora consiste en aplicar diversas técnicas de detección del DIF y tomar la decisión última de mantener, reformular o eliminar el ítem en función de la convergencia o divergencia entre métodos de detección, teniendo en cuenta las características y peculiaridades de cada procedimiento. Parece evidente que si diversas técnicas coinciden en sus decisiones, se tiene más certeza de la presencia o ausencia de DIF, mientras que si existe divergencia entre técnicas

deberíamos fijarnos en las características de los procedimientos de detección empleados. En cualquier caso se trataría de acumular evidencias en una dirección u otra, como en todo procedimiento de validación de un instrumento.

Siguiendo la clasificación de métodos de detección basada en el tipo de criterio de igualación de los grupos (puntuación observada/variable latente) y la relación entre la puntuación en el ítem y la variable de igualación (paramétrica/no paramétrica), se ha visto que se establecen cuatro tipos de métodos de detección tanto para ítems dicotómicos como de respuesta politómica: i) puntuación obervada/paramétrica, ii) puntuación observada/no paramétrica, iii) variable latente/paramétrica, y iv) variable latente/no paramétrica. Ya se ha señalado más arriba que los métodos que emplean la puntuación observada como estimación de la habilidad de los sujetos pueden resultar imprecisos cuando el criterio de igualación presenta un porcentaje elevado de ítems que funcionan diferencialmente, mientras que los métodos de la variable latente pueden superar este inconveniente a base de incrementar la complejidad matemática. Pero una ventaja de los métodos no paramétricos es que los supuestos del modelo son escasos, por lo que el DIF no suele confundirse con la falta de ajuste del modelo, mientras que con los métodos paramétricos es necesario asegurar un adecuado ajuste del modelo para evitar esta confusión, por tanto, se requieren tamaños muestrales mucho más elevados que con los modelos no paramétricos. Teniendo en cuenta las ventajas e inconvenientes generales que implican los distintos tipos de técnicas de detección, una recomendación iría en la línea de tomar la decisión última en base a la aplicación de una técnica de cada uno de los cuatro tipos existentes, por ejemplo, una opción sería emplear en la detección de ítems dicotómicos RL, MH, TRI y SIBTEST. Sin embargo, habría que tener en cuenta otras consideraciones respecto a los datos que podrían proporcionar una explicación sobre las posibles divergencias entre métodos de detección.

La primera de ellas deriva del tamaño muestral. Si se trabaja con tamaños reducidos, se ha evidenciado que RL y MH funcionan adecuadamente para ítems dicotómicos (Muñiz, Hambleton, y Xing, 2001; Swaminathan y Rogers, 1990) y TRI (usando la prueba de razón de verosimilitud) para ítems politómicos (Bolt, 2002). Si se dispone de tamaños considerables se puede optar por otras técnicas, como el SIBTEST y el POLYSIBTEST para la detección de ítems dicotómicos y politómicos.

Otro consejo giraría entorno al tipo de DIF. Existen técnicas que han sido diseñadas específicamente para la detección del DIF uniforme, por lo que pueden presentar ciertas dificultades en la detección del DIF no uniforme, mientras que otras se han propuesto para la detección de ambos tipos de DIF. Cuando se intuye la presencia de DIF no uniforme, es preferible emplear técnicas que sean sensibles a este tipo de funcionamiento diferencial. De nuevo, con tamaños muestrales reducidos, se puede optar por RL en ítems dicotómicos (Hidalgo y López-Pina, 2004) y TRI (usando la prueba de razón de verosimilitud) en ítems politómicos (Bolt, 2002). Si se emplean tamaños considerables se pueden seleccionar otras técnicas como el SIBTEST en el caso de ítems dicotómicos y la regresión logística multinomial (Zumbo, 1999) o el DFIT (Oshima, Raju y Nanda, 2006) para los de respuesta politómica.

Por otro lado, constatamos que la mayoría de los métodos actuales para detectar DIF requieren que el test a analizar contenga un número elevado de ítems (p.e. mayor de 30) para que el resultado sea fiable. Por el contrario, los cuestionarios y encuestas que se suelen utilizar en el ámbito de las ciencias sociales y de la salud suelen tener un número pequeño de ítems (entre 5 y 30 ítems). Cuando trabajamos con tests tan cortos la fiabilidad de las puntuaciones es menor y por lo tanto los errores de medida mayores. Métodos tales como RL o MH, que utilizan la puntuación observada en el test como variable de equiparación en el análisis del DIF, pueden ver seriamente afectada su eficacia para detectarlo. El uso de los modelos MIMIC (Gelin y Zumbo, 2007) es una alternativa.

Dado que la mayoría de estudios postulan que con la aplicación de procedimientos de purificación del criterio se produce una reducción de la tasa de falsos positivos y un incremento de la potencia estadística de diversos métodos, es aconsejable su empleo. Finalmente, en la medida de lo posible, se recomienda acompañar las tasas de detección con alguna medida del tamaño del efecto.

TESTS JUSTOS

Ya se ha comentado cómo en la década de los sesenta en EEUU se empezó a cuestionar el uso de los tests para evaluar de modo equitativo a distintos grupos de sujetos y cómo el artículo de Jensen (1969) sobre la naturaleza hereditaria de la inteligencia agudizó la polémica entre ambientalistas y genetistas. Dicha polémica tuvo una relevante repercusión social y política, llegándose a considerar que los tests en los que se constataban diferencias en función de características socioeconómicas o raciales,

estaban sesgados y eran injustos. Esta repercusión llegó a los tribunales donde se fallaron sentencias en contra de decisiones de selección de personal o de admisión a instituciones educativas. Una de las consecuencias más relevantes fue la llamada "regla dorada" que surgió del acuerdo el Educational Testing Service (la compañía de tests más importante de EEUU) y la compañía de seguros Golden Rule, por la que se debían eliminar los ítems en los que los sujetos de raza blanca obtuvieran un resultado superior en un 15% a los de raza negra. Evidentemente, dicha regla, basada únicamente en el índice de dificultad de los ítems para distintos grupos, podía conllevar la eliminación de ítems con alto poder discriminativo respecto al rasgo medido.

En ese momento, los términos sesgo e injusticia se equipararon y no se disponía de criterios eficaces para identificar si el comportamiento diferencial del test era debido a diferencias reales en el rasgo o a diferencias artefactuales provocadas por el instrumento utilizado. Esta oposición a que los tests se utilizaran para tomar decisiones que afectaran a la vida laboral o vida académica, fue también acicate para que los psicómetras se esforzaran en ofrecer definiciones y técnicas de detección de sesgos, dando lugar a una de las líneas de investigación psicométrica más fructífera de las últimas décadas. Así, en los años setenta-ochenta, aparece el término "funcionamiento diferencial del ítem" y se le distingue del término "sesgo", se ponen de relieve las diferencias entre DIF e impacto y se proponen técnicas de detección que permitan deslindar ambos aspectos. En la década de los noventa se incide en la explicación del DIF mediante la dimensionalidad de los tests; así, Ackerman (1992) distingue entre habilidad objetivo (aquella que pretende medir el test) y habilidad ruido (que no se pretendía medir pero que puede influir en las respuestas a algunos ítems del test): el DIF se puede presentar si los ítems del test miden una habilidad ruido en la que los sujetos difieren en función del grupo. Roussos y Stout (1996) añaden un matiz más: cambian la terminología y hablan de habilidades secundarias en vez de habilidad ruido, distinguiendo entre DIF-benigno y DIF-adverso; se da DIF-benigno cuando la habilidad secundaria es una dimensión auxiliar que se pretendía medir y DIF-adverso cuando la habilidad secundaria es una habilidad ruido.

De todos modos, y si bien es crucial ofrecer procedimientos estadísticos capaces de una eficaz detección de los ítems con funcionamiento diferencial, éstos por sí mismos no ofrecen una explicación de por qué el DIF se produce y si implica o no un sesgo. No hay que olvidar que la presencia de DIF es una condición necesaria pero no suficiente para poder hablar de sesgo del ítem: el DIF existe cuando individuos con una habilidad comparable pero de grupos distintos responden diferencialmente al ítem, mientras que para que exista sesgo se requiere además que estas diferencias sean atribuibles a alguna característica del ítem ajena al atributo medido por el test.

A finales del siglo pasado y principios de éste, se ha incidido en la importancia de analizar las causas del DIF. En el contexto de la adaptación de tests, los estudios de Allalouf, Hambleton y Sireci (1999) y de Gierl y Khaliq (2001) aportan algunas posibles causas centradas en el formato y el contenido del ítem; Zumbo y Gelin (2005) recomiendan que se consideren además diversas variables contextuales. Sin embargo, Ferne y Rupp (2007), en una revisión de 27 estudios que intentan identificar causas de DIF, constatan que los avances logrados son poco relevantes. Este es quizás uno de los retos actuales de la investigación en DIF, que merecería el mismo ahínco investigador que los anteriores problemas mencionados que se han ido solventando. Para ello convendría llevar a cabo estudios expresamente diseñados para investigar las causas del DIF, y sin duda la teoría multidimensional podría orientar la búsqueda de las causas hacia las habilidades espúreas que se distribuyen de forma distinta entre los grupos comparados. Considerar un resultado de DIF como una evidencia de sesgo implica explicar por qué el rasgo es multidimensional para un subgrupo específico y elaborar un argumento razonando la irrelevancia de la fuente de DIF para este rasgo (Camilli y Shepard, 1994).

En última instancia, como cualquier otro aspecto de la validez, el análisis del DIF es un proceso de acumulación de evidencias. Valorar e interpretar dichas evidencias requiere del juicio racional de los expertos y no existe una única respuesta correcta. En este sentido hay que apostar por la responsabilidad profesional de las personas que emplean tests, sensibilizarse y formarse acerca de la relevancia de la calidad métrica de estos instrumentos de medida, con la finalidad última de garantizar un proceso adecuado y justo de medición. Como paso previo a la aplicación de los métodos de detección, Hambleton y Rogers (1995) desarrollaron una lista de indicadores que pueden hacer sospechar de la posible presencia de DIF, por ejemplo, ítems que asocian los hombres al deporte y las mujeres a las acti-

vidades de la casa, o que utilizan ciertas palabras cuyo significado es más familiar para una cultura que para otra (alimentos, juegos, enfermedades, eventos históricos, etc.), entre otros. Además, Hambleton (2006) recomienda que tanto los creadores como los usuarios de tests tengan en cuenta los estudios previos de DIF, ya que pueden proporcionar información acerca de las características comunes a los ítems con DIF, así como las particularidades que comparten los ítems sin funcionamiento diferencial. Esta información es crucial tanto para el desarrollo de nuevos ítems como para alertarnos de la posible presencia de DIF en las pruebas existentes.

Como señala Zumbo (2007), los métodos de detección del DIF y del sesgo de los ítems se utilizan típicamente en el proceso de análisis de los ítems cuando se desarrollan nuevos instrumentos de medida, se adaptan tests existentes a un nuevo contexto de evaluación o a otras poblaciones que no se tuvieron en cuenta en el momento de crear el instrumento, se adaptan pruebas ya existentes a otras lenguas o culturas, o se validan las inferencias derivadas de las puntuaciones del test. Se constata pues que el ámbito de aplicación de un análisis del DIF es extenso y que está presente en las distintas fases de creación y adaptación de un instrumento de medida. En España, donde mayoritariamente se importan tests, es especialmente importante el análisis del DIF y del sesgo en la adaptación de instrumentos estandarizados a la lengua y contexto cultural propios. Desde el Colegio Oficial de Psicólogos (COP) se ha participado en la creación de unas directrices precisamente dedicadas a la creación y adaptación de tests, y obviamente en ellas el DIF tiene un papel destacado (Muñiz y Hambleton, 1996).

También tiene un papel relevante en los últimos estándares (APA et al, 1999) que incluyen el análisis del DIF y del sesgo en el análisis de la validez, concretamente en las evidencias basadas en la estructura interna del test. En definitiva, la decisión sobre si el resultado obtenido en un estudio es o no evidencia de sesgo sólo se puede tomar desde la teoría de la validez: conociendo la teoría subyacente al test, la interpretación que se pretende hacer de las puntuaciones y el contexto en el que se utiliza el test; en este sentido la ampliación de los contenidos de la validez permite que los estudios de sesgo aborden la perspectiva social del problema como una faceta más del proceso de validación de un test. El articulo de Prieto y Delgado (2010) en este mismo número describe el proceso de validación con más detalle.

PARA SABER MÁS

Aquel lector interesado en profundizar en el conocimiento de las técnicas de detección del DIF así como las implicaciones prácticas que supone la presencia de DIF en los ítems de un test puede consultar diversas revisiones teóricas (Camilli y Shepard, 1994; Fidalgo, 1996; Gómez-Benito e Hidalgo, 1997; Hidalgo y Gómez-Benito, 1999, 2010; Osterlind y Everson, 2009; Penfield y Lam, 2000; Potenza y Dorans, 1995) que se aproximan al estudio de las distintas técnicas de forma narrativa, exponiendo los procedimientos, especificando las ventajas y desventajas de su aplicación, y delineando recomendaciones para su empleo.

AGRADECIMIENTOS

Este trabajo ha sido financiado, en parte, por el Ministerio de Ciencia e Innovación (PSI2009-07280) y, en parte, por la Generalitat de Catalunya (2009SGR00822). Los autores muestran asimismo su agradecimiento a Vicente Ponsoda por su invitación a participar en este monográfico y al resto de investigadores de esta edición por su colaboración y ayuda en que este proyecto haya salido adelante.

REFERENCIAS

Ackerman, T.A. (1992). A didactic explanation of items bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Allalouf, A., Hambleton, R. K. y Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

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

Bolt, D.M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*(2), 113-141.

Camilli, G. y Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Dorans, N. J., y Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.

Ferne, T. y Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, *4,* 113-148.

Fidalgo, A.M. (1994). MHDIF – A computer-program for detecting uniform and nonuniform differential item functioning with the Mantel-Haenszel procedure. *Applied Psychological Measurement, 18*(3), 300-300.

Fidalgo, A. M. (1996). Funcionamiento diferencial de los ítems. En J. Muñiz (Coord.), *Psicometría* (pp. 370-455), Madrid: Universitas.

French, B.F. y Maller, S.J. (2007). Iterative purification and effect size use with Logistic Regression for Differential Item Functioning Detection. *Educational and Psychological Measurement, 67*, 373-393.

Gelin, M.N. y Zumbo, B.D. (2007). Operating characteristics of the DIF MIMIC approach using Jöreskog's covariance matrix with ML and WLS estimation for short scales. *Journal of Modern Applied Statistical Methods, 6*, 573-588.

Gierl, M. J., y Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests. *Journal of Educational Measurement, 38*, 164-187.

Gómez-Benito, J., e Hidalgo, M.D. (1997). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología, 74*(3), 3-32.

Gómez-Benito, J. e Hidalgo, M.D. (2007). Comparación de varios índices del tamaño del efecto en regresión logística: Una aplicación en la detección del DIF. Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud, Barcelona, 6-9 febrero.

Gómez-Benito, J., Hidalgo, M. D., Padilla, J. L., y González, A. (2005). Desarrollo informático para la utilización de la regresión logística como técnica de detección del DIF. Demostración informática presentada al *IX Congreso de Metodología de las Ciencias Sociales y de la Salud*, Granada, España.

Gómez-Benito, J., y Navas, M.J. (1996). Detección del funcionamiento diferencial del ítem: Purificación paso a paso de la habilidad. *Psicológica, 17,* 397-411.

González, A., Padilla, J.L, Hidalgo, M.D., Gómez-Benito, J. y Benítez, I. (2009) EASY-DIF: Software for analysing differential item functioning using the Mantel-Haenszel and standardization procedures. *Applied Psychological Measurement*. (Enviado para su publicación).

Hambleton, R.K. (2006). Good practices for identifying differential item functioning. *Medical Care, 44*(11 Suppl. 3), S182-S188.

Hambleton, R.K., y Rogers, H.J. (1995). Item bias review (EDO-TM-95-9). Washington, DC: Clearinghouse on Assessment and Evaluation.

Hessen, D.J. (2003). *Differential item functioning: Types of DIF and observed score based detection methods.* Dissertation (supervisors: G.J. Mellenbergh & K. Sijtsma). Amsterdam: University of Amsterdam.

Hidalgo, M. D., y Gómez-Benito, J. (1999). Técnicas de detección del funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento, 1*(1), 39-60.

Hidalgo, M. D., y Gómez-Benito, J. (2003). Test purification and the evaluation of differential item functioning with multinomial logistic regression. *European Journal of Psychological Assessment, 19*(1), 1-11.

Hidalgo, M. D., y Gómez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd edition). USA: Elsevier - Science & Technology.

Hidalgo, M.D., Gómez-Benito, J. y Zumbo, B.D. (2008). Efficacy of R-square and Odds-Ratio effect size using Dicriminant Logistic Regression for detecting DIF in polytomous items. Paper presented at the 6th Conference of the International Test Commission, 14-16 July, Liverpool, UK.

Hidalgo, M. D., y López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64*(4), 903-915.

Holland, P., y Thayer, D. (1988). Differential item performance and the Mantel-Haenszel procedure. En H. Wainer y H. I. Braun (Eds.), *Test Validity* (pp.129-145). Hillsdale, NJ: LEA.

Jensen, A.R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39(1),* 1-123.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free Press.

Jöreskog, K.G., y Sörbom, D. (2006). *Lisrel 8 (version 8.8).* Chicago, Illinois: Scientific Software International, Inc.

Mellenbergh, G. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7,* 105-118.

Messick, S. (1989). Validity. En R. Linn (Ed.). *Educatio-*

nal measurement (3rd edition, pp. 13-104). Washington, DC: American Council on Education.

Monahan, P.O., McHorney, C.A., Stump, T.E. y Perkins, A.J. (2007). Odds-ratio, Delta, ETS classification, and standardization measures of DIF magnitude for binary logistic regression. Journal of Behavioral Statistics, 32, 1, 92-109.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo, 31(1)*, 57-66.

Muñiz, J., y Hambleton, R.K. (1996). Directrices para la traduccion y adaptacion de los tests. *Papeles del Psicólogo, 66*.

Muñiz, J., Hambleton, R. K., y Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing, 1*(2), 115-135.

Muthén, L.K., y Muthén, B.O. (1998, 2007). *MPLUS statistical analysis with latent variables. User's Guide*. Los Angeles, CA: Muthén and Muthén.

Navas-Ara, M. J. y Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment, 18*(1), 9-15.

Oshima, T. C, Raju, N. S. y Nanda, A. O. (2006). A new method for assessing the statistical significance in the differential functioning of item and tests (DFIT) framework. *Journal of Educational Measurement, 43*, 1-17.

Osterlind, S. J., y Everson, H. T. (2009). Differential item functioning (2nd edition). Thousand Oaks, California: Sage Publications, Inc.

Penfield, R. D. (2005). DIFAS: Differential Item Functioning Analysis System. *Applied Psychological Measurement, 29*(2), 150-151.

Penfield, R. D., y Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19,* 5-15.

Potenza, M., y Dorans, N. (1995). DIF assessment for politomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement, 19,* 23-37.

Prieto, G. y Delgado, A. (2010). Fiabilidad y validez. *Papeles del Psicólogo, 31(1)*, 67-74.

Ramsay, J. O. (2000). *TestGraph: A program for the graphical analysis of multiple choice and test questionnaire*. Unpublished manual.

Roussos, L. y Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

SPSS 15.0. (2009). SPSS Inc. 1989-2009.

Stout, W. y Roussos, L. (1999). *Dimensionality-based DIF/DBF package* [Computer Program]. William Stout Institute for Measurement. University of Illinois.

Swaminathan, H. y Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Thissen, D. (2001). *IRTLRDIF v2.0b. Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio Test for Differential Item Functioning*. Available on Dave Thissen's web page.

van de Vijver, F., y Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage Publications.

Waller, N. G. (1998a). EZDIF: Detection of uniform and nonuniform differential item functioning with the Mantel-Haenszel and Logistic Regression procedures. *Applied Psychological Measurement,* 22, 391.

Waller, N.G. (1998b). LINKDIF: Linking item parameters and calculating IRT measures of Differential Item Functioning of Items and Tests. *Applied Psychological Measurement,* 22, 392.

Wang, W.-C, Shih, C.-L. y Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement, 69*, 713-731.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type (Ordinal) Item Scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D. y Gelin, M. N. (2005). A matter of test bias in educational policy research: Bringing the context into picture by investigating sociological / community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5,* 1-23.

Zumbo, B. D., y Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Working paper of the Edgeworth Laboratory for Quantitative Behavioral Science, University of Northern British Columbia: Prince George, B.C.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 85-96
http://www.cop.es/papeles

LA EVALUACIÓN DEL DESEMPEÑO
PERFORMANCE ASSESSMENT

Rosario Martínez Arias
Universidad Complutense de Madrid

Las prácticas de evaluación han evolucionado desde el uso casi exclusive de tests formados por ítems de elección múltiple a la combinación de formatos múltiples, incluyendo tareas de desempeño. El objetivo del artículo es proporcionar una visión del concepto, diseño, uso y características psicométricas de los tests de desempeño. Comienza con el concepto y la justificación de su uso. En la sección 2 se presentan los principales usos actuales de este tipo de tests. La sección 3 describe algunos aspectos relativos al diseño y puntuación. La sección 4 muestra algunas cuestiones relativas a las características psicométricas. La sección 5 concluye con una valoración de los tests de desempeño, presentando sus principales fuerzas y debilidades, así como las necesidades de futuras investigaciones. Se necesita un esfuerzo continuado en modelos y métodos de medida que permitan mejorar la generalizabilidad y las evidencias de validez de los tests de desempeño.
Palabras clave: *Test de desempeño, Centros de evaluación, Guías de puntuación, Generalizabilidad, Evidencias de validez.*

Assessment practices have gradually shifted from almost exclusively objectively score multiple-choice test items to the use of a mixture of formats, including performance assessments. The purpose of this article is to provide an overview on the concept, design, use and psychometric characteristics of performance assessment. The article is divided in five sections. It begins with the concept and rationale for the use of performance assessment. Section 2 presents the main uses of these tests. Section 3 describes some questions related to the design and scoring. Section 4 shows some issues related to psychometric characteristics. Section 5 concludes with an evaluation of performance assessment tests indicating the main strengths and weaknesses and needs of future research. Continued work is needed on measurement models and methods that can improve the generalizabiliy and the evidences of validity of performance assessments. The computers can contribute to practical issues.
Key words: *Performance assessment, Assessment centers, Scoring rubrics, Generalizability, Evidences of validity.*

EL CONCEPTO DE TEST DE DESEMPEÑO

Muchas personas, incluso profesionales de la psicología, consideran el test estandarizado como sinónimo de test de elección múltiple o de respuesta construida única. Esta consideración está justificada ya que estos formatos han dominado el campo de los tests de inteligencia, aptitudes y rendimiento académico durante muchos años y por buenas razones, relacionadas sobre todo con la cobertura de contenido y las facilidades para la corrección y puntuación. No obstante, bajo la etiqueta de test estandarizado caben otros formatos que cumplen con todos los requisitos exigibles a un test y que pueden mostrar adecuadas propiedades psicométricas. Entre ellos se encuentran los que aquí denominamos "tests de desempeño"[1] (*performance assessment*), de uso cada vez más frecuente en la evaluación psicológica y educativa.

En la Tabla 1 se presenta una clasificación de los distintos tipos de formato, que pueden adoptar los tests estandarizados y que se pueden graduar a lo largo de varios continuos (Gronlund, 2006).

Dentro de los formatos anteriores, suelen considerarse evaluación del desempeño los ensayos, proyectos, simulaciones y muestras de trabajo. Puede observarse que estos formatos se encuentran más próximos a los extremos caracterizados como de mayor autenticidad, complejidad cognitiva, cobertura en profundidad y respuesta estructurada por el propio sujeto. También se caracterizan por un mayor costo.

Dada la diversidad de formatos que pueden adoptar, se presenta a continuación una definición integradora que permita recoger su diversidad: "*los tests de desempeño son procedimientos estandarizados de evaluación en los que se demanda de los sujetos que lleven a cabo tareas o procesos en los que demuestren su capacidad para aplicar conocimientos y destrezas a acciones en si-*

Correspondencia: Rosario Martínez Arias. *Departamento de Metodología de las Ciencias del Comportamiento. Universidad Complutense de Madrid. E-mail: rmnez.arias@psi.ucm.es*

[1] Se ha traducido la expresión inglesa "performance assessment" por "evaluación del desempeño" o tests de desempeño. El término procede de la evaluación educativa y de las certificaciones profesionales, no obstante, en psicología también se utilizan estos tests desde hace mucho tiempo, especialmente en el ámbito de la selección de personal. Aunque no se utiliza el término "performance assessment", las tareas de simulación y muestras de trabajo utilizadas en los *centros de evaluación* (*assessment centers*) muestran todas las características de este tipo de tests: demandan respuestas que ponen el acento en la actuación del sujeto y que requieren métodos sistemáticos para su valoración. Con la llegada de las nuevas tecnologías su uso se está extendiendo a otros ámbitos como la psicología clínica y la neuropsicología.

tuaciones simuladas o de la vida real".

Estos tests pueden ser tan diversos como escribir un ensayo, interpretar una composición musical, hacer una presentación oral, diagnosticar a un paciente estandarizado, planificar las actividades del día o proponer una solución a un problema empresarial. En todos los casos el sujeto debe producir algo durante un período de tiempo y se evalúan los procesos o productos con relación a criterios de rendimiento establecidos.

La definición es integradora en el sentido de que recoge los dos grandes grupos en los que suelen dividirse las definiciones: las que ponen el acento en el formato de la respuesta y las que lo ponen en la semejanza entre la respuesta demandada y el criterio de interés (Palm, 2008). En este grupo, la mayor parte ponen el acento en la actuación del examinado (Stiggins, 1987).

Algunas van más allá del formato de respuesta, insistiendo en sus características de *autenticidad* y de *simulación* de la situación criterio. Así, Fitzpatrick y Morrison (1971) los definen como "aquellos tests en los que se simula una situación criterio con más fidelidad y globalidad que en los usuales tests de papel y lápiz" (p.268). Kane, Crooks y Cohen (1999) destacan que "suponen una muestra de la actuación del sujeto en algún dominio, interpretando las puntuaciones resultantes en términos del rendimiento típico o esperado en dicho dominio……..siendo su característica definitoria la estrecha semejanza entre el tipo de actuación observada y la de interés" (p.7). En esta misma línea se encuentra la definición de los *Standards for Educational and Psychological Tests* (American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME), 1999), que consideran que "las evaluaciones del desempeño *emulan* el contexto o las condiciones en las que se aplican los conocimientos y destrezas que se intenta evaluar" (p.137).

Esta insistencia en la emulación del rendimiento de interés lleva a confusión con la denominada *evaluación auténtica* (Wiggins,1989). Ésta comparte muchas características con los tests de desempeño y es una de sus formas, pero implica otros aspectos que van más allá de los exigidos a estos tests.

Con frecuencia se destaca también la complejidad cognitiva, ya que exigen de los sujetos la utilización de estrategias de orden superior, como planificar, estructurar la tarea, obtener información, construir las respuestas y explicar el proceso, integrando conocimientos e información (Ryan, 2006).

Los tests de desempeño suelen clasificarse por lo que evalúan y en este sentido suele hablarse de *productos* (*products*) o resultados de la tarea y *desempeños* (*performances*), que son los procesos que sigue el examinado para llegar a la solución. Ejemplos típicos del primer tipo son los ensayos escritos, informes de laboratorio, actuaciones artísticas, etc. Entre los segundos se encuentran las presentaciones orales y las demostraciones. La mayor parte de las veces suponen una combinación de procesos y productos.

USOS DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño no representan algo nuevo; Madaus y O'Dwyer (1999) establecen sus orígenes en el 210 aC durante la Dinastía Han en China. Evaluaciones similares se utilizaron en los gremios durante la Edad Media y en las Universidades para la valoración de los estudiantes. En la psicología del trabajo tienen una larga tradición en el ejército y desde hace más de 60 años se emplean en los denominados *Centros de Evaluación* (*Assessment Centers*), hoy conocidos como *Centros de Evaluación y Desarrollo* (Thorton y Rupp, 2006), en los que se emplean muestras de trabajo y ejercicios simulados para evaluar a los sujetos en competencias difíciles de medir con tests convencionales. Su uso se remonta a 1942 en la *War Office Selection Boards* del Reino Unido para la selección de altos mandos y pronto se extendieron a los Estados Unidos y otros países, especialmente

TABLA 1
CONTINUOS DE FORMATOS DE TESTS

Muestras de Trabajo	Simulaciones	Proyectos	Ensayos	Respuesta corta	Elección Múltiple	Verdadero / Falso
Más auténtica	⟷					Menos auténtica
Cognitivamente. más compleja	⟷					Cognitivamente menos compleja
Cobertura en profundidad	⟷					Cobertura en contenido
Respuesta estructurada por el sujeto	⟷					Respuesta estructurada por el test
Mayor costo	⟷					Menor costo

de lengua alemana. Desde los años cincuenta se emplean para la selección de puestos directivos, aunque en la actualidad su uso se está generalizando a muchos tipos de puestos (Thorton y Rupp, 2006).

En la evaluación educativa, las fuertes críticas de los años sesenta y setenta al formato de elección múltiple llevaron a la inclusión de tareas de desempeño en la evaluación. Durante los años noventa se pasa del uso exclusivo de formatos de elección múltiple a formatos mixtos que incluyen tareas de desempeño, como ensayos escritos, secuencias de solución de problemas, presentaciones orales e incluso *portafolios* de los estudiantes (Hambleton, 2000). Las razones del cambio son diversas, pero básicamente tienen que ver con las limitaciones de los tests de elección múltiple para lograr algunos objetivos educativos: 1) evaluar habilidades de nivel cognitivo superior; 2) evaluar destrezas para el aprendizaje a lo largo de la vida (pensamiento independiente, persistencia, flexibilidad,..); 3) evaluación de las estrategias de resolución de problemas y dificultades; 4) alineación de destrezas y habilidades con las competencias importantes para la vida, junto con contextos realistas y 5) integrar la evaluación con la instrucción de acuerdo con las teorías del aprendizaje y la psicología cognitiva. Estos objetivos están incluidos en las reformas educativas en las que se pone el acento en la enseñanza de habilidades cognitivas superiores (Linn, 1993a) y en la unión entre evaluación e instrucción, por considerar la evaluación como un instrumento valioso para la mejora de la instrucción y del aprendizaje (Frederiksen y Collins, 1989; Stiggins, 1987). Intentan superar las denunciadas reducciones del currículo generadas por los tests de elección múltiple, en la creencia de que la evaluación determina lo que los profesores enseñan y lo que los estudiantes aprenden (Wiggins, 1989).

Los avances experimentados por la psicología cognitiva fueron un importante detonante para la inclusión de tareas de desempeño en las evaluaciones. En 1998 el *Board on Testing and Assessment* del *National Research Council* (NRC) formó un comité de 18 expertos presidido por Pellegrino y Glaser para establecer un puente entre los avances de la psicología cognitiva y los métodos de medición educativa. El producto final fue un excelente libro: "*Knowing What Students Know: The Science and Design of Educational Assessment*" (NRC, 2001). En el texto se destacan las limitaciones de los tests tradicionales para captar los conocimientos y las destrezas complejas exigidas por los nuevos estándares de rendimiento y la escasa validez de las inferencias derivadas de sus puntuaciones. El comité desarrolló un marco teórico para la evaluación, *el triángulo de la evaluación,* basado en la idea de que la evaluación es un *proceso de razonamiento desde la evidencia* (Mislevy, 2006; Mislevy, Steinberg y Almond, 2002; Mislevy, Wilson, Ercikan y Chudowsky, 2003), que se apoya en tres pilares: a) un modelo de representación del conocimiento y desarrollo de las competencias, b) tareas o situaciones que permitan observar el desempeño de los estudiantes y c) métodos de interpretación para extraer inferencias.

Por otra parte, la llegada de los ordenadores abrió la posibilidad de usar nuevos formatos de ítem y de respuesta, facilitando tanto la administración como la puntuación de estas tareas (Drasgow, Luecht y Bennett, 2006; Zenisky y Sireci, 2002).

En la actualidad los tests de desempeño están presentes en la mayor parte de las evaluaciones a gran escala, generalmente acompañados de ítems de formato estructurado. En Estados Unidos comenzaron a incluirse en el *National Assessment of Educational Progress* (NAEP) durante los años 90 y hoy muchos estados evalúan el desempeño en sus programas anuales de tests. También se incluyen en todas las evaluaciones internacionales a gran escala, como *Trends in International Mathematics and Science Study*, TIMSS (Arora, Foy, Martin y Mullis, 2009) y en el programa PISA (OECD, 2007). En España se han incorporado a las pruebas de diagnóstico desarrolladas por el Instituto de Evaluación.

En el ámbito del trabajo, las evaluaciones del desempeño tienen una fuerte representación en las certificaciones profesionales, especialmente para el ejercicio de la medicina y la abogacía. Como ejemplo de las primeras se encuentran los tests de *United States Medical Licensure Examination* (USMLE, 2009). Un ejemplo de las segundas es el *Multistate Performance Test*, utilizado en 30 estados de Estados Unidos (National Conference of Bar Examiners & American Bar Association, 2005).[2]

Una gran parte de las tareas de estos tests son similares a las que se utilizan en los *centros de evaluación* para la selección de personal. En estos sistemas el tipo de tareas adopta múltiples formas, aunque las más comunes son las siguientes: tests en la bandeja, *role-play* en interacciones, análisis de casos escritos de la organización,

[2] Descripciones detalladas de las evaluaciones del desempeño utilizadas en diversas acreditaciones profesionales pueden encontrarse en Johnson, Penny y Gordon (2009).

presentaciones orales, liderazgo en discusiones de grupo, búsqueda de hechos relevantes a partir de presentaciones orales, juegos de empresa y combinaciones de varias tareas o ejercicios. Una descripción de las tareas de los centros de evaluación puede consultarse en Thornton y Rupp (2006).

Las conductas de los sujetos se evalúan en *dimensiones* relevantes para los puestos de trabajo. Su número y tipo difiere según el objetivo del centro de evaluación (Thorton y Rupp, 2006). Algunas son comunes a la mayor parte de los centros y similares a las de las certificaciones: solución de problemas, comunicación oral, liderazgo, gestión de conflictos, búsqueda de información, planificación y organización, adaptabilidad cultural, generación de soluciones, usos de los recursos,... (Arthur, Day, McNelly y Edens, 2003; Brummel, Ruth y Spain, 2009).

DESARROLLO, ADMINISTRACIÓN Y PUNTUACIÓN DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño deben asegurar que los ejercicios o tareas estén estandarizados, sean válidos, fiables, equitativos y legalmente defendibles. Para conseguirlo deben seguir en su desarrollo los estándares y guías para la construcción y uso de los tests como los *Standards for educational and psychological tests* (AERA et al., 1999). En el caso de los ejercicios de los centros de evaluación, deben cumplir además con algunas guías específicas como los *Principles for the validation and use of personnel selection procedures* (Society for Industrial and Organizational Psychology, 2003) y con las *Guidelines and Ethical Considerations for Assessment Center Operations* (International Task Force on Assessment Center Guidelines (2000).

El proceso de desarrollo comienza con la *definición del marco*, que supone la descripción del constructo o de las tareas, el propósito de la evaluación y las inferencias que se harán con las puntuaciones. El marco conceptual guía el desarrollo de las *especificaciones,* que reflejan el contenido, los procesos, las características psicométricas de las tareas y otra información pertinente para la evaluación. Pueden seguirse dos aproximaciones, centrada en el constructo o en la tarea, aunque se recomienda la primera (Messick, 1994). El constructo guía la adecuada representación del dominio, la selección de las tareas, los criterios para establecer las puntuaciones y la detección de posible varianza irrelevante. Patz (2006) presenta una buena descripción del desarrollo de una evaluación de ciencias. En los centros de evaluación, el marco de definición de los constructos o competencias se deriva de un riguroso análisis del puesto de trabajo (Thorton y Rupp, 2006).

Para la adecuada estandarización es necesario determinar las condiciones de la administración que permitan la comparabilidad de las puntuaciones (AERA et al., 1999). Se elaboran *Guías* en las que se establecen los tiempos, ítems o tareas de ensayo, equipamiento y materiales, así como instrucciones para la aplicación (Cohen y Wollack, 2006).

La clave del éxito de estos tests y uno de los aspectos más controvertidos es la correcta *asignación de puntuaciones* a las tareas realizadas. Para ello se elaboran las G*uías para la especificación de puntuaciones* (*scoring rubrics)* en las que se establecen los criterios de valoración de las respuestas junto con un procedimiento para puntuarlas (Clauser, 2000). Deben ser claras, completas e ilustradas con ejemplos de respuestas tipo (Welch, 2006). Su objetivo es obtener puntuaciones consistentes e invariantes a través de evaluadores, tareas, localizaciones, ocasiones y otras condiciones. Combinadas con un adecuado entrenamiento de los evaluadores permiten alcanzar niveles adecuados de fiabilidad.

Hay dos tipos de guías, las *holísticas* o *globales* y las *analíticas*. En las globales los evaluadores emiten un único juicio sobre la calidad del proceso o producto, asignando una puntuación basada en descripciones de *anclaje* de los distintos niveles. En las analíticas, las descripciones del desempeño se separan en partes (aspectos, criterios evaluativos, dimensiones, dominios,...). Además de los epígrafes de las guías, se incluyen respuestas ejemplares para operacionalizar cada uno de los criterios evaluativos, denominados "*anclajes*" o puntos de referencia.

Una guía analítica especifica rasgos o aspectos detallados de las respuestas y el número de puntos que se deben atribuir a cada uno, permitiendo la ponderación. Los distintos rasgos suelen puntuarse por medio de escalas tipo Likert con varios grados. En los centros de evaluación se utiliza un procedimiento similar al que se aplica en las guías analíticas, conocido como *Behaviorally Anchored Rating Scales (BARS),* que incluyen descripciones ejemplares ("*anchored"*) de conductas y permiten valorar cada dimensión en escalas que suelen tener cinco puntos.

Una variación del sistema de puntuaciones analítico es el de las *listas de conductas* (*checklist*s) en las que cada aspecto se valora como Sí o No, según que la actuación esté o no presente. Es el procedimiento habitual en las

acreditaciones médicas y legales y en ocasiones en los centros de evaluación en lugar de las BARS.

Cuando las tareas se basan en las teorías cognitivas del aprendizaje dentro de un dominio las puntuaciones pueden reflejar criterios de progresión en el aprendizaje (Wilson, 2005).

La elección de una u otra forma depende en gran medida del constructo, el propósito de la evaluación, si lo evaluado es un proceso o producto y de las inferencias que se derivarán de las puntuaciones. El número de categorías o de puntos de la escala depende de la facilidad de diferenciación y discriminación. Lane y Stone (2006) indican que tendrá el número suficiente de categorías para diferenciar entre niveles de rendimiento y que no sean tantas que la diferenciación se haga difícil.

Uno de los aspectos más investigados son los méritos relativos de los dos sistemas de puntuaciones, analizados a partir de la fiabilidad entre jueces. Por el momento no hay respuestas claras, no pudiéndose hablar de un procedimiento superior en todas las situaciones. Parece que las guías holísticas se ven más afectadas por las fuentes de sesgo de los evaluadores que las analíticas y que las listas de conducta mejoran el acuerdo entre jueces. Johnson et al., (2009) y Arter y McTighe (2001) recomiendan las holísticas para las tareas relativamente simples, como las incluidas en las evaluaciones a gran escala. Las analíticas son más adecuadas para ejecuciones complejas con múltiples elementos, como es el caso de las licencias y certificaciones y de los centros de evaluación (Welch, 2006).

Los grandes costos derivados de la puntuación de estos tests han motivado el desarrollo de algunos sistemas informatizados para la corrección (Bennett, 2004; Livingston, 2009; Williamson, Mislevy y Bejar, 2006). Para su implementación se identifican un gran número de respuestas tipo evaluadas por calificadores expertos, que representan el rango total de las puntuaciones de la escala y posteriormente se utilizan algoritmos para la obtención de las puntuaciones que emulan a los evaluadores humanos (Williamson et al., 2006).

Otro aspecto esencial es la formación de los evaluadores con los que se intenta llegar a adecuados grados de acuerdo, corrigiendo sus sesgos. Los sesgos más frecuentes se presentan resumidos en la Tabla 2, adaptada de Johnson et al. (2009).

Un procedimiento frecuente de entrenamiento supone la inclusión de protocolos previamente corregidos por expertos, que permiten detectar evaluadores con sesgos y la monitorización con evaluadores experimentados.

CARACTERÍSTICAS PSICOMÉTRICAS DE LOS TESTS DE DESEMPEÑO

Los tests de desempeño deben cumplir con los criterios psicométricos exigibles a todo procedimiento de evaluación (Kane, 2004), para lo que se utilizan los diferentes modelos de la teoría de los tests. Ciertas características específicas exigen el uso de modelos más avanzados que la Teoría Clásica de los Tests (TCT), como la Teoría de la Generalizabilidad (TG) y la Teoría de la Respuesta al Ítem (TRI). Las nuevas concepciones de la validez también llevan a algunas diferencias con respecto a los planteamientos tradicionales (véanse los artículos de Muñiz (2010) sobre las teorías de los tests y de Prieto y Delgado (2010) sobre fiabilidad y validez, en este mismo número).

A continuación se revisan brevemente algunos aspectos psicométricos de los tests de desempeño: el tratamiento de los errores de medida y la consistencia de las puntuaciones (fiabilidad), los procedimientos para obtener estimaciones de la habilidad y las evidencias de validez. Esta clasificación convencional es difícil en estos tests, ya

TABLA 2
SESGOS MÁS FRECUENTES DE LOS EVALUADORES

Tipo de sesgo	Tendencia del evaluador a....
Apariencia	Puntuar fijándose en aspectos que considera importantes
Tendencia central	Asignar puntuaciones en torno al punto medio
Conflicto de estándares	Sus estándares personales no están de acuerdo con los de las guías
Fatiga	Estar afectado por el cansancio
Efecto halo	Atribuir puntuaciones altas por algún aspecto valioso para el evaluador
Escritura del sujeto	Asignar puntuaciones basadas en características del escrito
Arrastre de ítems	Valorar un ítem en función de lo que ha hecho en otros
Lenguaje	Valorar basándose en el lenguaje utilizado por el evaluado
Longitud	Valorar más las respuestas más largas
Indulgencia/severidad	Tendencia a puntuaciones altas/bajas
Repetición	Valorar menos porque ha visto el tópico repetidamente
Prejuicios	Puntuación baja debido a algún aspecto de la respuesta que no le gusta al evaluador
Efectos de otros tests ya corregidos	Puntuar menos una respuesta de lo que dice la guía porque va precedida de respuestas excelentes de otros examinados
Aspecto particular	Poner el acento en un aspecto y darle demasiado peso

que la generalizabilidad de las puntuaciones se trata con frecuencia como uno de los aspectos de la validez (Brennan, 2000a; Kane, Crooks y Cohen, 1999; Miller y Linn, 2000; Messick, 1996).

La fiabilidad y consistencia de las puntuaciones

En ocasiones, la fiabilidad puede tratarse con la TCT (Johnson et al., 2009), pero generalmente se necesita utilizar la TG. Este modelo, sistematizado por Cronbach, Gleser, Nanda y Rajaratnam (1972) tuvo escaso eco en la construcción de tests hasta la llegada de los tests de desempeño en los que su uso se ha generalizado. La TG es una extensión de la TCT que utiliza modelos de Análisis de la Varianza (componentes de la varianza) que permiten estimar simultáneamente efectos de diferentes fuentes de variabilidad o error (*facetas)* sobre las puntuaciones. Las facetas consideradas con mayor frecuencia son las tareas y los evaluadores, aunque en algunos estudios se incluyen ocasiones, administración y formato del test. Permite analizar los efectos principales de cada faceta, así como sus interacciones con el sujeto y entre ellas. La TG contempla dos tipos de estudios, los G o de Generalizabilidad y los D o de Decisión. En los primeros se estima la contribución relativa de cada faceta y de sus interacciones sobre la varianza error y estas estimaciones permiten optimizar el procedimiento de medida, determinando el número óptimo de tareas, evaluadores, etc., en cada aplicación para la reducción del error. En los estudios D se calculan los *coeficientes de generalizabilidad* bajo las condiciones de medida concretas utilizadas; éstos pueden ser de dos tipos, según que las decisiones sean *absolutas* o *relativas*. La posibilidad de descomponer la varianza error en diferentes fuentes es lo que hace imprescindible a la TG en los tests de desempeño. La fiabilidad mejora cuando se realizan estudios para determinar el número de tareas y de evaluadores necesarios.

Razones de espacio nos impiden extendernos más en la descripción de la TG. Un completo tratamiento puede encontrarse en Brennan (2000b). Una exposición resumida se presenta en Martínez Arias, Hernández Lloreda y Hernández Lloreda (2006) y la descripción de sus principales características en el artículo de Prieto y Delgado (2010), en este monográfico.

Las fuentes de error más investigadas son las relativas a *tarea y evaluador*. Se ha encontrado que los efectos de las tareas son los más críticos, debido al reducido número que se puede incluir para cada habilidad o competencia, encontrándose baja consistencia entre tareas e interacciones con los sujetos (Lane y Stone, 2006).

El efecto de los evaluadores es importante, tanto como efecto principal como en interacción con tareas y con sujetos. En evaluaciones de la escritura se han encontrado correlaciones entre jueces muy heterogéneas que van de .33 a .91 (Lane y Stone, 2006) y algo más altas en las certificaciones médicas, con valores entre .50 y .93 (van der Vleuten y Swanson, 1990). En cuanto al tipo de competencia evaluada, se encuentra mayor consistencia en ciencias y matemáticas y menor en escritura (Shavelson, Baxter y Gao, 1993).

En general, puede decirse que la variabilidad de las tareas contribuye más al error que el evaluador en la mayor parte de las materias (Lane y Stone. 2006; Shavelson et al., 1993).

Algunos modelos de la TRI desarrollados en el marco del modelo de Rasch (Adams, Wilson y Wang, 1997) permiten incorporar los efectos del evaluador en la estimación de las puntuaciones.

Aunque tareas y evaluadores son las fuentes de variabilidad más estudiadas, también se han investigado los efectos de otras facetas: ocasiones (tiempos de la medida), formato de la evaluación y comité de calificadores. Una faceta importante es la ocasión (Cronbach, Linn, Brennan y Haertel, 1997; Fitzpatrick, Ercikan, Yen y Ferrara, 1998), especialmente en evaluaciones periódicas en las que se examinan cambios y puntúan evaluadores diferentes.

La estimación de la competencia o habilidad

Para obtener estimaciones de la competencia o habilidad de los sujetos suelen utilizarse como marco los modelos de la TRI para respuestas politómicas ordenadas (Abad, Ponsoda y Revuelta, 2006). Recientes avances en modelos TRI multidimensionales (*Multidimensional Item Response Theory*, MIRT) permiten tratar con la complejidad de estas evaluaciones, en las que es difícil lograr la unidimensionalidad asumida (Gibbons et al., 2007; Reckase, 2009).

Un problema frecuente es la combinación de diferentes formatos de respuesta en el mismo test. El uso de modelos de TRI con software especializado para modelos politómicos permite obtener estimadores únicos de las habilidades o rasgos en estas condiciones.

En relación con la estimación de las puntuaciones surge el problema de la *equiparación*, cuando se utilizan diferentes conjuntos de ítems, en la misma evaluación o en tiempos distintos para evaluar cambios. Las características de estos tests plantean problemas especiales para la aplicación de las técnicas de equiparación estricta (Kolen y Brennan, 2004), debiendo tratarse con frecuencia mediante formas más débiles como la calibración, pre-

dicción o moderación (Linn, 1993b). Los principales problemas se deben a la frecuente multidimensionalidad, la dificultad de encontrar ítems de anclaje comunes, ser ítems politómicos y la dependencia entre ítems (Muraki, Hombo y Lee, 2000), así como los efectos del evaluador (Kolen y Brennan, 2004). Para su tratamiento se utilizan con frecuencia los modelos multigrupo de la TRI (Bock, Muraki y Pfeiffenberger, 1988). Reckase (2009) propone algunos procedimientos en el contexto de los modelos multidimensionales. Un tratamiento reciente de estos problemas puede encontrarse en Dorans, Pommerich y Holland (2007).

Las evidencias de validez de los tests de desempeño

La definición de validez de las puntuaciones de los tests de desempeño es la establecida en los *Standards for Educational and Psychological Tests* (AERA et al., 1999), similar a la de otros tipos de tests estandarizados, con la validez de constructo como concepto unificador. En el artículo de Prieto y Delgado (2010), en este monográfico, se presenta la definición y los tipos de evidencias. En los tests de desempeño se mencionan a veces otros aspectos como la autenticidad, significación para los evaluados (Linn, Baker y Dunbar, 1991) y validez sistémica (Fredericksen y Collins, 1989). Messick (1996) considera estos aspectos dentro de la representación del constructo (autenticidad) y de los aspectos sustantivos y consecuenciales de la validez (significación y validez sistémica).

A continuación se revisan brevemente las evidencias de validez, con algunas consideraciones sobre el sesgo y la equidad, que también podrían tratarse dentro de los aspectos de varianza irrelevante para el constructo o de las consecuencias.

Evidencias de validez de contenido

En los tests de desempeño se presentan con mayor frecuencia que en los convencionales las dos grandes amenazas a la validez de contenido señaladas por Messick (1989,1996): *infrarrepresentación del constructo* y *varianza irrelevante*. La primera suele deberse al reducido número de ítems que se incluyen. La segunda tiene múltiples fuentes: la elección del tema por los sujetos, la tendencia de los evaluadores a fijarse en aspectos irrelevantes o sesgos (Messick,1994, 1996; véase Tabla 2), los procedimientos de corrección automatizados (Lane y Stone, 2006) y la motivación de los sujetos, especialmente en situaciones de tests sin consecuencias (DeMars, 2000; O'Neil, Subgure y Baker, 1996).

Evidencias de validez desde los procesos de la respuesta o sustantiva

Messick (1996) destaca "*la necesidad de obtener evidencias empíricas de los procesos puestos en juego por los examinados cuando realizan la tarea*" (p.9). Dadas las expectativas puestas en estos tests para evaluar procesos cognitivos superiores es preciso justificar que efectivamente lo hacen (Hambleton, 1996; Linn et al., 1991). Por el momento son escasas las investigaciones y los resultados son poco consistentes (Ayala, Shavelson, Shue y Schultz, 2002). Algunos desarrollos inspirados en los modelos del *Latent Trait Logistic Model* (Fischer, 1973), como los de Embretson (1998) y Gorin y Embretson (2006) son prometedores en este sentido. Adams, Wilson y Wang (1997) desarrollaron una versión multidimensional, adecuada para este tipo de tests.

Un marco teórico interesante es el del *triángulo de la evaluación*, mencionado en la sección segunda de este artículo. En las evaluaciones de los aprendizajes educativos, las teorías del desarrollo y progresión del aprendizaje también permiten sustentar este tipo de validez (Briggs, Alonzo, Schwab y Wilson, 2006; Wilson, 2005).

Estructural

Según AERA et al. (1999) "el análisis de la estructura interna de un test puede indicar el grado en que las relaciones entre los ítems y los componentes del test se adecuan al constructo en el que se basan las interpretaciones de las puntuaciones" (p.13).

La evaluación de la dimensionalidad suele establecerse por medio de técnicas de análisis factorial. Hay pocos trabajos publicados sobre la estructura factorial de los tests de desempeño en educación. Las razones son variadas: 1) la complejidad de los estímulos lleva a la recomendación de análisis de contenido y análisis sustantivo (Ackerman, Gierl y Walker, 2003); 2) los esquemas de puntuación pueden tener impacto en la dimensionalidad, llevando en ocasiones a la multidimensionalidad y 3) diferentes puntos de la escala de valoración pueden reflejar diferentes combinaciones de destrezas (Reckase, 1997). Los avances en los modelos multidimensionales de la teoría de la respuesta al ítem (Gibbons et al., 2007; Reckase, 2009) pueden aportar evidencias estructurales. En el ámbito de los centros de evaluación se ha tratado más este aspecto, encontrando resultados contradictorios. Rupp et al. (2006) encuentran evidencias de dimensiones claras, pero otros autores las cuestionan (Lance, 2008).

Externa

Según AERA et al. (1999) "el análisis de las relaciones de las puntuaciones del test con variables externas es otra fuente importante de evidencia de validez" (p.13). Para obtener estas evidencias se examinan los patrones de correlaciones empíricas según las expectativas teóricas o hipótesis del constructo. Messick (1996) pone el acento en la importancia de las evidencias de validez *convergente* y *discriminante* mediante las matrices multimétodo-multirrasgo (MMMR). La "e*videncia convergente* significa que la medida está coherentemente relacionada con otras medidas del mismo constructo, así como con otras variables con las que debe estar relacionada sobre bases teóricas. La *evidencia discriminante* significa que la medida no debe estar relacionada con otros constructos" (Messick, 1996, p.12).

Debe justificarse que la varianza debida al constructo supera considerablemente a la varianza del método o de las tareas. En el ámbito educativo hay pocos trabajos sobre estas evidencias, pero es un tema muy investigado en los centros de evaluación, encontrándose en general bajas evidencias, ya que suele ser mayor la proporción de varianza ligada al contexto que al constructo (Lance, 2008). No obstante, Rupp, Thorton y Gibbons (2008) atribuyen estos resultados a deficiencias metodológicas en el diseño de las matrices multimétodo-multirrasgo.

En relación con las evidencias referidas a criterios externos, la mayor parte de la investigación se ha realizado en el ámbito de los centros de evaluación. En un meta-análisis, Arthur et al. (2003) encontraron correlaciones entre .25 y .39 según el tipo de competencias. Salgado y Moscoso (2008) revisan la fiabilidad y la validez operativa (corregidos los sesgos de falta de fiabilidad del criterio y restricción del rango) de diversos instrumentos de selección, encontrando para las simulaciones de los centros de evaluación coeficientes de fiabilidad de .70 y validez de .37, siendo esta última inferior a la de otros procedimientos (tests de aptitudes generales y razonamiento, de conocimientos del puesto y entrevista conductual estructurada). Estos datos plantean dudas sobre la utilidad de estos procedimientos frente a otros que son además más económicos.

Las consecuencias del uso de los tests

Este aspecto de la validez de constructo tiene que ver con las consecuencias deseadas y no deseadas del uso de los tests y su impacto sobre la interpretación de las puntuaciones (Messick, 1996). Estas evidencias se han estudiado en el contexto educativo donde son uno de los argumentos más frecuentes para el uso de estos tests. Entre las consecuencias positivas para los examinados se incluyen la motivación, el aprendizaje y la aplicación de lo que han aprendido.

Hasta el momento las investigaciones son escasas. Stecher et al. (2000) en una encuesta al profesorado encontraron que 2/3 de los profesores de 4º y 7º grado respondieron que los estándares del estado y los tests de desempeño les influyeron en sus estrategias docentes. El impacto del Maryland State Performance Assessment Program fue examinado por Lane y colaboradores (Lane, Parke y Stone, 2002; Parke, Lane y Stone, 2006) encontrando que tanto los equipos directivos como el profesorado consideraban que la evaluación había llevado a cambios positivos en la instrucción y en las prácticas de evaluación en clase. No obstante, este resultado puede deberse a las consecuencias de la evaluación para las escuelas (rendición de cuentas).

Sesgo y equidad

Generalmente se entiende por sesgo "la validez diferencial de una interpretación de la puntuación de un test para cualquier subgrupo definible de sujetos que responden al test" (Cole y Moss, 1989, p.205). Para evitarlo se recomienda utilizar las técnicas de detección del *funcionamiento diferencial de los ítems*, que permiten detectar tareas o ítems que potencialmente pueden contribuir al sesgo. Para una descripción más detallada, véase el artículo de Gómez Benito, Hidalgoy Guilera (2010), en este monográfico.

Se han realizado escasas investigaciones sobre el funcionamiento diferencial de los ítems de los tests de desempeño (Lane y Stone, 2006). La mayor parte de las investigaciones se limitan al análisis de las diferencias entre grupos. En los ensayos escritos se encuentran diferencias entre varones y mujeres, favorables a éstas (Ryan y DeMark, 2002) y diferencias étnicas (Engelhard, Gordon, Walker y Gabrielson, 1994). En estudios más adecuados para el análisis del sesgo, se encontraron diferencias no paralelas entre formatos de desempeño y de elección múltiple (Livingston y Rupp, 2004). Cuando hombres y mujeres muestran resultados similares en los de elección múltiple, las mujeres son superiores en los de respuesta construida; cuando son similares en los de respuesta construida, los varones son superiores a las mujeres en los de elección múltiple.

Se considera que los tests de desempeño pueden mostrar más factores irrelevantes, que favorecen el funcionamiento diferencial (Penfield y Lamm, 2000). Su detección es más difícil, debido a las dificultades ya mencionadas a propósito de la equiparación.

CONCLUSIONES

Los tests de desempeño hoy forman parte del repertorio de las técnicas de evaluación y su uso es creciente. Han generado muchas expectativas por su validez aparente y sus potenciales ventajas: mayor autenticidad mediante la emulación de situaciones reales, posibilidad de medir habilidades y competencias difíciles de evaluar con otros formatos, medición de los procesos además de los productos, su valor educativo y formativo y la detección de los progresos de aprendizaje. Todo ello hace que se consideren imprescindibles en las evaluaciones, normalmente combinados con tests o tareas de formatos tradicionales. Las innovaciones derivadas del uso de nuevas tecnologías ayudan a su aplicación, posibilitando evaluar nuevas competencias y dimensiones.

No obstante, a pesar de sus innegables ventajas y de lo extendido de su uso, presentan todavía numerosos retos a la investigación psicométrica. Sus principales limitaciones son las siguientes:

- ✔ Dificultades para la representación adecuada del dominio por el número limitado de tareas que se pueden incluir.
- ✔ Problemas de generalizabilidad, debidos sobre todo a la varianza debida a las tareas y a la interacción de éstas con sujetos y evaluadores.
- ✔ Inconsistencias y sesgos de los evaluadores, que obligan al desarrollo de Guías de Puntuaciones muy claras, elaboradas y costosas. También requieren costosos procesos de entrenamiento de los evaluadores, para obtener puntuaciones consistentes entre calificadores y ocasiones.
- ✔ Los costos de corrección son altos, requiriendo a veces demasiado tiempo, lo que dificulta el uso formativo de los resultados.
- ✔ La complejidad de las tareas lleva a menudo a estructuras multidimensionales, que dificultan el empleo de los modelos de TRI unidimensionales para la estimación, equiparación o calibración y funcionamiento diferencial.
- ✔ Se requiere más investigación sobre el funcionamiento diferencial de las tareas de desempeño en relación con otros formatos y la influencia de factores motivacionales.
- ✔ Aunque su validez aparente está clara, las diferentes evidencias de validez psicométrica deben investigarse más. Dentro de este bloque son muy importantes ciertos aspectos irrelevantes en los que se fijan los evaluadores, para eliminar su influencia. Las evidencias sustantivas referidas a los procesos deben continuar investigándose, por medio del uso de modelos de medida que permitan su evaluación, así como la progresión de los aprendizajes. También parece necesario el examen de las evidencias de relaciones con otras variables.

Los desarrollos actuales de modelos psicométricos tanto dentro de la TRI, como en otros marcos (Mislevy, 2006), que permiten la introducción de componentes ligados a los procesos representan un importante avance. Los modelos de la TRI multidimensionales y jerárquicos también permitirán tratar con algunas de las limitaciones anteriores. Se necesita más investigación sobre la combinación adecuada de tareas de formato de elección múltiple y respuestas cortas con tareas de desempeño para optimizar la información.

La introducción de nuevas tecnologías puede mejorar muchas limitaciones. La presentación y respuesta por ordenador mediante tests adaptativos permite reducir considerablemente el tiempo del test, mejorando la representación del dominio. Permiten además el uso de tareas dinámicas, como las que se usan en el diagnóstico de pacientes, mejoran la autenticidad, que puede verse incrementada por la inclusión de múltiples recursos (gráficos, vídeo, audio, materiales de referencia,…). También mejoran el registro de los procesos mediante seguimiento, poniendo de relieve evidencias de validez sustantiva. Las respuestas emitidas a través del ordenador pueden corregir algunos aspectos irrelevantes relacionados con la escritura y forma de exposición. Por otra parte, debe continuarse en el desarrollo de los sistemas automatizados de corrección que reducirán considerablemente los costos.

Finalmente, podríamos preguntarnos si los tests de desempeño deben sustituir a los formatos tradicionales como los de elección múltiple y la respuesta es negativa, ya que hay muchos aspectos de las evaluaciones que pueden evaluarse adecuadamente con estos formatos más económicos en tiempo y dinero. Lo ideal es la combinación adecuada de los distintos tipos.

En este artículo se ha presentado una visión general y limitada de los tests de desempeño. Las personas interesadas en el tema pueden encontrar un tratamiento extenso en las referencias citadas de Johnson et al., (2009) sobre aplicaciones en educación y en acreditaciones; en el libro de Thorton y Rupp (2006) se trata con bastante profundidad el tema de los centros de evaluación. Por último, ejemplos de tareas de tests de desempeño típicas de la evaluación educativa se encuentran entre los ítems hechos públicos del estudio PISA (http://www.pisa.oecd.org). Información sobre tests de desempeño en las certificaciones y acreditaciones pueden encontrarse en las páginas web

de la American Board of Pediatric Dentistry (http://www.abdp.org/pamphlets/oral_handbook.pdfvv, en la National Board of Medical Examiners (http://www.usmle.org/Examinations/step2/step2ck_content.html) y en la ya mencionada de la National Conference of far Examiners (http://www.ncbex.org/multistate-tests/mbe).

REFERENCIAS

Abad, F.J., Ponsoda, V. y Revuelta, J. (2006). *Modelos politómicos de respuesta al ítem.* Madrid: La Muralla.

Ackerman, T.A., Gierl, M.J. y Walker, C.M. (2003). An NCME instructional module on using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice, 22,* 37-53.

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1–23.

Arora, A., Foy, P., Martin, M.O. y Mullis,I.V.S. (Eds.) (2009). *TIMSS Advanced 2008: technical report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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

Arter, J. y McTighe, J. (2001). *Scoring rubrics in the classroom.* Thousand Oaks, CA: CorwinPress.

Arthur, W., Day, E.A., NcNelly, T.L. y Edens, P.S. (2003). A meta-analysis of the criterion-related validity of assessment center dimension. *Personnel Psychology, 56,* 125-154.

Ayala, C.C., Shavelson,R.J., Yue, Y., y Schultz, S.E. (2002). Reasoning dimensions underlying science achievement: the case of performance assessment. *Educational Assessment, 8,* 101-121.

Bennett, R.E. (2004). *Moving the field forward: Some thoughts on validity and automated scoring* (ETS RM 04-01). Princeton, NJ: Educational Testing Service.

Bock, R.D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*, 275-285.

Brennan, R. L. (2000a). *Generalizability theory.* New York: Springer-Verlag.

Brennan, R. (2000b). Performance assessment from the perspective of the generalizability theory. *Applied Psychological Measurement, 24,* 339- 353.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33–63.

Brummel, B.J., Rupp, D.E. & Spain, S.M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62,* 137-170.

Clauser, B. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310–324.

Cohen, A. y Wollack, J. (2006). Test administration, security, scoring and reporting. En R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 355-386). Wesport, CT: American Council on Education/Praeger.

Cole, N.S. y Moss, P.A. (1989). Bias in test use. En R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 201-220).

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam,N. (1972). *The dependability of behavioural measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cronbach, L., Linn, R., Brennan, R., & Haertel, E. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement 57*, 373–399.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55- 77.

Dorans, N.J., Pommerich, M. y Holland, P.W. (Eds.). (2007). *Linking and aligning scores and scales.* New-York: Springer.

Drasgow, F., Luecht, R.M. y Bennett, R.E. (2006). Technology and testing. En R.L. Brennan (Ed.), *Educational Measurement.* Pp.471-515.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300–396.

Engelhard, G., Gordon, B., Walker, E.V y Gabrielson, S. (1994). Writing tasks and gender: Influences on writing quality of black and white students. *Journal of Educational Research, 87,* 197-209.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374.

Fitzpatrick, R., Ercikan, K., Yen, W.M. y Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education, 11,* 95-208.

Fitzpatrick, R., & Morrison, E. (1971). Performance and product evaluation. In R. Thorndike (Ed.), *Educational measurement* (pp. 237–270). Washington, DC: American Council of Education.

Frederiksen, J.R. y Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18,* 27-32.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007a). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31*, 4-19.

Gómez-Benito, J., Hidalgo, M.D. y Guilera, G. (2010). El sesgo de los instrumentos de medida. Tests justos. *Papeles del Psicólogo, 31 (1),* 75-84.

Gorin, J.S. y Embretson, S.E. (2006). Item difficulty modeling of paragraph comprehension ítems. *Applied Psychological Measurement, 30,* 394-411.

Hambleton, R.K. (1996). Advances in assessment models, methods, and practices. In D.C. Berliner.

and R.C. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 899-925). New York:Macmillan.

Hambleton,R.K. (2000). Advances in performance assessment methodology. *Applied Psychological Measurement, 24,* 291-293.

International Taskforce on Assessment Center Guideliness (2000). Guideliness and ethical considerations for assessment center operations. *Public Personnel Management, 29,* 315-331.

Johnson, R.L., Penny, J.A. y Gordon, B. (2009). *Assessing performance: designing, scoring, and validating performance tasks.* New York: Guilford Press.

Kane, M.T. (2004). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2*, 13-170

Kane, M. T., Crooks, T.,&Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18 (2),* 5–17.

Kolen, M.J. y Brennan, R.L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd Ed.). New York: Springer.

Lance, C.E. (2008). Why assessment centers do not work the way they are supposed. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1,* 84-97.

Lane, S., Parke, C.S. y Stone, C.A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment, 8,* 279-315.

Lane, S. y Stone, C.A. (2006). Performance assessment. En Brennan (Ed), *Educational Measurement*, (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger.

Linn, R.L. (1993a). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis, 15,* 1-16.

Linn, R. L. (1993b). Linking results of distinct assessments. *Applied Measurement in Education, 6,* 83-102.

Linn, R.L., Baker, E. L. y Dunbar, S.B. (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20 (8),* 15-21.

Livingston, S.A. (2009). *Constructed-response test questions: Why we use them; how to score them.*(R & D Connections, nº 11). Princeton, NJ: Educational Testing Service.

Livingston, S.A. y Rupp, S.L. (2004). *Performance of men and women on multiple-choice and constructed-response tests for beginning techers.* (ETS Research Report No.RR.04-48). Princeton, NJ: Educational Testing Service.

Martínez Arias, R., Hernández Lloreda, MV y Hernández Lloreda, MJ. (2006). *Psicometría*. Madrid: Alianza.

Madaus, G., & O'Dwyer, L. (1999). A short history of performance assessment. *Phi Delta Kappan, 80*(9), 688–695.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement,* (3rd ed., pp. 13–103). Washington, DC: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Messick, S. (1996). Validity of performance assessments. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics.

Miller, D.M. y Linn, R.L. (2000). Validation of performance assessments. *Applied Psychological Measurement, 24, 367-378.*

Mislevy, R.J. (2006). Cognitive psychology and educational assessment. En R.L. Brennan (Ed.), *Educational Measurement,* pp. 257-305.

Mislevy, R., Steinberg, L., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*(4), 477–496.

Mislevy, R., Wilson, M., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 489–532). Boston: Kluwer Academic.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica

y Teoría de la Respuesta a los Ítems. *Papeles del Psicólogo, 31*,

Muraki, E., Hombo, C.M. y Lee, Y.W. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24,* 325-33.

National Conference of Bar Examiners (NCBE) y American Bar Association (ABA). (2005). *Bar admission requirements*. Disponibleen http://www.ncnex.org/tests.htm. ECD (2007). *PISA 2006 Science Competencies for Tomorrow's World*. París: OECD.

O'Neil, H.F., Subgure, E. y Baker,E.L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress Mathematics Performance. *Educational Assessment, 3,* 135-157.

Palm, T. (2008). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation, 13*, nº 4. Disponible en http://pareonline.net/getvn.asp?v=13&n=4

Parke, C.S., Lane, S. y Stone, C.A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12,* 239-269

Patz, R.J. (2006). Building NCLB science assessments: Psychometric and practical considerations. *Measurement, 4*,199-239.

Penfield, R.D. y Lamm, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practices, 19,* 5-15.

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

Reckase, M.D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Reckase, M.D. (2009). *Multidimensional ítem response theory.* New York: Springer.

Rupp, D. E., Gibbons, A.M., Baldwin, A. M., Snyder, L. A., Spain, S. M., Woo, S. E., etal. (2006). An initial validation of developmental assessment centers as accurate assessments and effective training interventions. *Psychologist-Manager Journal*, *9*, 171–200.upp, D.E, Thorton, G.C. y Gibbons, A.M. (2008). The construct validity of the assessment center method and usefulness of dimensions as focal constructs. *Industrial and Organizational Psychology, 1,* 116-120.

Ryan, T. (2006). Performance assessment: Critics, criticism, and controversy. *International Journal of Testing, 6*(1), 97–104.

Ryan, J. M., y DeMark, S. (2002). Variation in achievement scores related to gender, item format and content area tested. En G. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: validity, technical adequacy, and implementations issues*, (pp. 67–88). Mahwah, NJ: Lawrence Erlbaum Associates.

Salgado, J.F. y Moscoso, S. (2008). Selección de personal en la empresa y las AAPP: de la visión tradicional a la visión estratégica. *Papeles del Psicólogo, 29,* 16-24. http://www.cop.es/papeles

Shavelson,R.J., Baxter, G.P. y Gao, X. (1993). Sampling Variability of performance assessments. *Journal of Educational Measurement, 30,* 215-232.

Stecher, B., Klein, S., Solano-Flores, G., McCaffrey, D.m Robyn, A., Shavelson, R. y col. (2000). The effects of content, format and inquiry level on science performance assessment scores. *Applied Measurement in Education, 13*, 139–160.

Society for Industrial and Organizational Psychology (2003). Principles for the validation and use of personnel selection procedures (4th ed.). Bowling Green, OH: Author.

Stiggins, R. (1987). Design and development of performance assessments. *Educational Measurement: Issues and Practices, 6*(3), 33–42.

Thorton, G.C. y Rupp, D.E. (2006). *Assessment centers in human resource management.* Mahwah, NJ: Erlbaum.

United States Medical Licensure Examination (2009). *Examinations.* Disponible en http://www.usmle.org/examinations/index.html Van der Vleuten, C. y Swanson, D. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and learning in Medicine, 2*, 58-76.

Welch, C. (2006). Item and prompt development in performance testing. In S. Downing y T. Haladyna (Eds.), *Handbook of test development* (pp. 303–327). Mahwah, NJ: Erlbaum.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*, 703-713.

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

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

Zenisky, A.L. y Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15,* 337-362.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 94-107
http://www.cop.es/papeles

TESTS INFORMATIZADOS Y OTROS NUEVOS TIPOS DE TESTS

COMPUTERIZED TESTS AND OTHER NEW TYPES OF TEST

Julio Olea[1], Francisco J. Abad[1] y Juan R. Barrada[2]
[1]*Universidad Autónoma de Madrid.* [2]*Universidad Autónoma de Barcelona*

Recientemente se ha producido un considerable desarrollo de los tests adaptativos informatizados, en los que el test se adapta progresivamente al rendimiento del evaluando, y de otros tipos de tests: a) los test basados en modelos (se dispone de un modelo o teoría de cómo se responde a cada ítem, lo que permite predecir su dificultad), b) los tests ipsativos (el evaluado ha de elegir entre opciones que tienen parecida deseabilidad social, por lo que pueden resultar eficaces para controlar algunos sesgos de respuestas), c) los tests conductuales (miden rasgos que ordinariamente se han venido midiendo con autoinformes, mediante tareas que requieren respuestas no verbales) y d) los tests situacionales (en los que se presenta al evaluado una situación de conflicto laboral, por ejemplo, con varias posibles soluciones, y ha de elegir la que le parece la mejor descripción de lo que el haría en esa situación). El artículo comenta las características, ventajas e inconvenientes de todos ellos y muestra algunos ejemplos de tests concretos.
Palabras clave: *Test adaptativo informatizado, Test situacional, Test comportamental, Test ipsativo y generación automática de ítems.*

The paper provides a short description of some test types that are earning considerable interest in both research and applied areas. The main feature of a computerized adaptive test is that in despite of the examinees receiving different sets of items, their test scores are in the same metric and can be directly compared. Four other test types are considered: a) model-based tests (a model or theory is available to explain the item response process and this makes the prediction of item difficulties possible), b) ipsative tests (the examinee has to select one among two or more options with similar social desirability; so, these tests can help to control faking or other examinee's response biases), c) behavioral tests (personality traits are measured from non-verbal responses rather than from self-reports), and d) situational tests (the examinee faces a conflictive situation and has to select the option that best describes what he or she will do). The paper evaluates these types of tests, comments on their pros and cons and provides some specific examples.
Key words: *Computerized adaptive test, Situational test, Behavioral test, Ipsative test and y automatic item generation.*

Hace un par de años que varios historiadores de la Psicología Española (Quintana, Bitaubé y López-Martín, 2008) rescataron y editaron unos "Apuntes para un curso de Psicología aplicada a la selección profesional", elaborados en 1924 por el doctor Rodrigo Lavín como material docente de su cátedra de Psicología Experimental. Esta auténtica joya representa una de las primeras veces que en España se habla sistemáticamente de los tipos y usos de los tests. Decía ya entonces el autor: "*Como la observación nos da muy pocos datos utilizables y la conversación o entrevista no basta para descubrir las habilidades de los solicitantes, es necesario recurrir a los tests. Se puede decir que estamos en el comienzo de los tests y, a pesar de eso, hay un desarrollo extraordinario de ellos; ello indica lo que sucederá andando el tiempo*". Hablaba el autor de que existían entonces tests de capacidades o habilidades, tanto generales como específicas, y que en la selección profesional eran de especial importancia los tests de fuerza, de resistencia a la fatiga, de control motor y de capacidades mentales (atención, sensación y percepción, imaginación e inteligencia general).

Andado el tiempo hasta hoy, el desarrollo de los tests ha sido extraordinario, como anticipaba Lavín, tanto en la variedad como en la complejidad. Prueba de ello es que las clasificaciones simples de los tipos de tests (por ejemplo, la que distinguía entre "tests impresos" y "tests manipulativos" o las que se referían al diferente contenido de las pruebas) se han quedado obsoletas por la presencia de nuevos tipos de tests que eran difíciles de prever en el pasado. Todo ello se ha debido a distintos factores:

- ✔ *Avances técnicos.* El desarrollo de los modelos psicométricos que sustentan las propiedades métricas de los tests y la evolución y abaratamiento de la tecnología informática nos ha permitido incorporar nuevos atributos psicológicos al catálogo de lo medible; también ha permitido incrementar la eficiencia de las aplicaciones e incluir nuevas funcionalidades, como son la generación automática de ítems, la aplicación adaptativa de un test o la corrección automática de respuestas complejas.

Correspondencia: Julio Olea. *Facultad de Psicología. Universidad Autónoma de Madrid. Calle Iván Pavlolv 6. 28049 Madrid. España. E-mail: Julio.olea@uam.es*

- ✔ *Nuevas demandas sociales.* En España, aunque todavía con cierta lejanía respecto a otros países, tanto los profesionales de la Psicología como otros responsables de organizaciones públicas y privadas confían cada vez más en la utilidad de los tests para conseguir ciertos objetivos aplicados, como lo prueba el artículo de Muñiz y Fernández-Hermida (2010) de este mismo número. Pero no sólo se incrementa el uso de los tests "clásicos" como el WAIS o el 16PF. En una sociedad cada vez más sensible a la evaluación de los resultados de las intervenciones y a la acreditación de competencias individuales e institucionales, se ha ampliado mucho el tipo de atributos psicológicos que se precisa medir. Mientras que hace unos años las aplicaciones fundamentales se ceñían a tests de capacidades cognitivas o pruebas de personalidad, cada vez son más los profesionales que exigen buenos tests para objetivos específicos.
- ✔ *Mayor exigencia de calidad.* Cada vez son más importantes las consecuencias que para las personas y las organizaciones tienen las puntuaciones en los tests. Por ello, también es mayor la exigencia psicométrica a la que sometemos a las puntuaciones de los tests. El ineludible requisito de "medir bien" y la necesidad de afrontar problemas singulares en ciertos contextos de evaluación (como puede ser el falseamiento de las respuestas en contextos de selección) está impulsando el desarrollo de nuevos tipos de tests y nuevos modelos psicométricos para estudiar las garantías que ofrecen sus aplicaciones.

TESTS INFORMATIZADOS

Se van incrementando progresivamente los tests cuyos ítems se presentan, se responden y puntúan en un ordenador, lo que ha representado cambios y avances importantes en contextos aplicados de evaluación psicológica y educativa. Para Davey (2005): *"En las últimas dos décadas los tests informatizados han pasado de ser un procedimiento experimental a ser empleado por cientos de programas de evaluación que evalúan a millones de personas cada año"* ... *"ser evaluado mediante un ordenador puede pronto llegar a ser incluso más natural que ser evaluado en papel"* (p. 358).

Estrictamente hablando, un test informatizado debe cumplir dos requisitos (Olea, Ponsoda y Prieto, 1999): a) que se conozcan las propiedades psicométricas de los ítems que lo integran, estimadas a partir de un modelo matemático, y b) que los ítems se presenten y respondan en un ordenador. El primero de estos requisitos excluye de la consideración como "test informatizado" a muchos de los tests que sin las oportunas garantías se ofrecen en Internet.

El ordenador permite aplicar los tests de diversos modos. Existen en primer lugar los *tests fijos informatizados.* En estos tests los ítems se aplican en la misma secuencia a todos los evaluados. Un segundo tipo son *los tests adaptativos informatizados,* que permiten presentar los mejores ítems para cada evaluado. Por su importancia, dedicaremos a este tipo de tests un apartado propio.

En general, informatizar un test supone ciertas ventajas:

- ✔ Ayuda a estandarizar mejor las condiciones de aplicación de los tests para todos los evaluados: instrucciones comunes, control del tiempo de aplicación, reducción de la posibilidad de copia y de la transmisión del contenido de los tests, eficiencia en la corrección de respuestas, etc.
- ✔ Resulta necesario para la aplicación de los complejos procedimientos de estimación que se requieren en Teoría de la Respuesta al Ítem (TRI) (véase en este número Muñiz, 2010), con lo que ha permitido aplicar nuevos modelos psicométricos y hacer operativas sus eventuales ventajas.
- ✔ Permite proporcionar de forma inmediata información cuantitativa, verbal y gráfica sobre la posición de un evaluado respecto a un grupo en un baremo concreto, es decir, permite la elaboración de informes automáticos; es posible también proceder a una actualización continua de los baremos, incorporando a los mismos las puntuaciones de nuevos evaluados.
- ✔ El ordenador es necesario para aplicar *nuevos formatos de ítems* (por ejemplo presentaciones visuales dinámicas, ítems auditivos o secuencias simuladas grabadas en video), lo que ha representado una importante ampliación de los rasgos, competencias y comportamientos que pueden evaluarse en Psicología como, por ejemplo, la aptitud musical, el rendimiento de un controlador de tráfico aéreo, la capacidad para resolver conflictos, etc. (véase Drasgow y Olson-Buchanan, 1999). Se amplía así el rango de atributos que se pueden evaluar, aumentando la similaridad entre la tarea de evaluación y los criterios a predecir a partir de las puntuaciones en el test (por ejemplo las actividades a desempeñar por el evaluado en el puesto de trabajo). Además, puede romperse con el formato tradicional de respuesta (opción múltiple o categorías ordenadas) para plantear, por ejemplo, tareas tan diversas como marcar en un mapa determinadas localizaciones, seguir con el ratón el movimiento de un determinado objeto, rotar cier-

tos grados figuras tridimensionales, detectar y cambiar errores gramaticales de diversos textos, escribir con un editor de ecuaciones el resultado simplificado de una fórmula matemática, grabar una respuesta verbal en un micrófono, dar un diagnóstico médico después de recabar información diversa sobre los síntomas de un paciente o ubicar los componentes arquitectónicos de un edificio. En este tipo de ítems, además de los aciertos o errores, el ordenador permite registrar otro tipo de variables para medir el rendimiento (por ejemplo, los tiempos de reacción o las distancias físicas respecto a la solución óptima de una tarea visomotora).

- ✔ Algunos sistemas de evaluación informatizada permiten ya la corrección automática de la ejecución en una tarea concreta. En la figura 1, se muestra un ejemplo de un ítem de conocimientos sobre Botánica, que consiste en sombrear las zonas de distribución de una determinada especie y cuya corrección es automática (tomado de Conejo, Guzmán, Millán, Trella, Pérez de la Cruz y Ríos, 2004). Para puntuar este ítem, se usa un mapa con el sombreado correcto como plantilla. Si el estudiante marca aproximadamente la zona correcta (con un cierto margen de error) se puntúa como correcta la respuesta. Además se señala la proporción de área que es correctamente localizada. Por ejemplo: el 15.5% del área sombreada es correctamente sombreada y el 91.66% del área no sombreada es correctamente no sombreada.

El libro recientemente editado "*Automated Scoring of Complex Tasks in Computer Based Testing*" (Williamson, Mislevy y Bejar, 2006) recoge numerosos ejemplos de corrección automática en ítems de respuesta compleja. En este libro se aconseja elaborar *Diseños Centrados en la Evidencia* (DCE), en los que se plantea un esquema a seguir en este tipo de desarrollos. En la metodología DCE se parte de un modelo del evaluado (descripción exhaustiva de los constructos, habilidades o destrezas que queremos medir) y de un modelo de la tarea o familia de tareas (en el que se describen exhaustivamente las características de las tareas que permitirían generar el ítem de forma automática). El modelo de evidencia conecta ambos modelos recogiendo las relaciones entre el resultado del sujeto en la tarea y el constructo o la decisión sobre el evaluado (por ejemplo apto o no apto). En los DCE se diferencia entre reglas de evidencia o puntuación (que transforman el resultado del sujeto en las tareas en puntuaciones numéricas) y un modelo de medida (que conecta las puntuaciones numéricas con las puntuaciones en los constructos y con las decisiones que se tomen a partir de estos).

Uno de los primeros intentos importantes en el desarrollo de pruebas de corrección automática fue el ARE (Architectural Registration Examination), una batería de evaluación que desempeña un papel importante en el proceso de acreditación por el que en Canadá se otorgan licencias para ejercer como arquitecto. Algunos ítems exigen que el evaluado maneje algunas funcionalidades básicas de una herramienta informática para el diseño gráfico (ver figura 2). La tarea del evaluado es hacer un diseño de una vivienda, clínica... que cumpla con un conjunto de exigencias. Los diseños producidos

FIGURA 1

por el evaluado son puntuados automáticamente por un algoritmo atendiendo a la seguridad, la funcionalidad, la atención a las restricciones (geográficas, ambientales, climáticas…), la accesibilidad, etc. El desarrollo de estos procedimientos automáticos requiere de la colaboración de expertos y de la formación de grupos de discusión que permitan diseñar un algoritmo que proporcione puntuaciones similares a las que proporcionaría un evaluador humano. Paradójicamente, aunque los expertos proporcionan las reglas de puntuación incorporadas en el algoritmo, la corrección automática puede llegar a ser más eficiente, por una mayor sistematicidad en la aplicación de los criterios. El test ARE es un test de desempeño. El artículo de Martínez-Arias de este número los estudia en detalle.

FIGURA 2

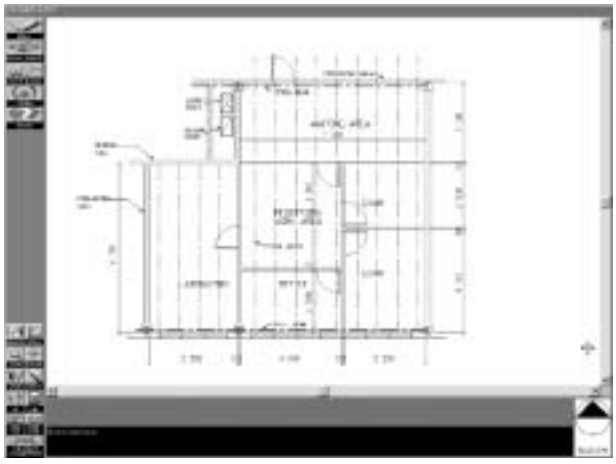
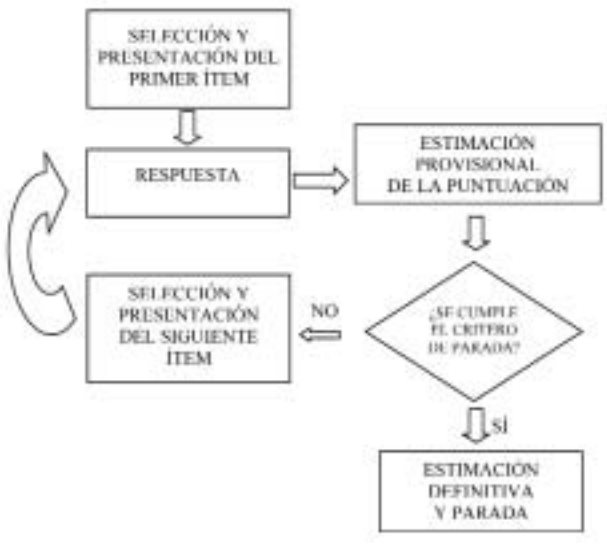


FIGURA 3



Tests adaptativos informatizados

El uso de los ordenadores combinado con la TRI permite la construcción de *tests adaptativos informatizados* (TAIs), cuya principal característica es que los ítems a administrar se van adaptando al nivel de competencia que va manifestando el evaluado, según sus respuestas a los ítems previos. Partiendo de un banco de ítems amplio, distintos ítems de ese banco son seleccionados para cada persona. Gracias a la TRI, las estimaciones del nivel de rasgo obtenidas en los distintos tests serán comparables (se encontrarán en la misma métrica).

La idea básica consiste en presentar únicamente los ítems que resultan altamente informativos para estimar el nivel de cada sujeto en un determinado rasgo. Una vez calibrado el banco de ítems, el proceso de aplicación de un TAI a un evaluando puede resumirse, de forma simplificada, en el diagrama de flujo de la Figura 3 (Olea y Ponsoda, 2003).

La aplicación de un TAI se inicia con una determinada estrategia de arranque, que consiste en establecer de alguna forma el nivel de rasgo inicial que se asigna al evaluando (por ejemplo, el nivel promedio de la población). Después de que el evaluando responde a cada ítem, se realiza una estimación de su nivel de rasgo mediante procedimientos estadísticos bayesianos o máximo-verosímiles. Se requiere también un algoritmo para la selección sucesiva de ítems. Generalmente, se emplean procedimientos basados en la medida de información, $I(\theta)$; por ejemplo, puede seleccionarse como segundo ítem el más informativo para el nivel θ estimado tras la primera respuesta. En contextos de acreditación, promoción o selección es importante que se muestreen los contenidos adecuadamente o que los evaluados reciban, en la medida de lo posible, ítems distintos. En esos casos, un algoritmo adecuado de selección deberá incluir restricciones en la tasa de exposición de los ítems (por ejemplo, que cada ítem no sea administrado en más del 20% de los tests) u otras restricciones, para que se garantice un adecuado muestreo de contenidos. Se requiere finalmente algún criterio para dar por terminada la secuencia de presentación de ítems, que normalmente tiene que ver con la consecución de cierto nivel de precisión o con haberse aplicado un número prefijado de ítems; esto último suele ser necesario para mantener el balance de contenidos en la prueba, y preferible para evitar en los usuarios del TAI la sensación de que se les

ha medido con pocos ítems. Como se representa en el diagrama, el ciclo "seleccionar ítem - aplicar ítem - recoger respuesta - estimar rasgo" se repite hasta que se satisface el criterio de parada.

Los TAIs, dada su condición adaptativa, tienen al menos tres importantes ventajas adicionales a las de cualquier test informatizado:

- ✔ Mejoran la seguridad del test, ya que gran parte de los ítems que se presentan a los evaluados son diferentes. Esta es una preocupación fundamental de los responsables de la evaluación en contextos aplicados porque, incluso cuando se decide aplicar tests convencionales, uno de los mayores obstáculos a la validez de los tests es que los evaluados puedan conocer de antemano los ítems que se les van a administrar.
- ✔ Reducen el tiempo de aplicación (a veces a menos de la mitad), ya que consiguen niveles similares de precisión que los tests convencionales con un número menor de ítems.
- ✔ Permiten además, con el mismo número de ítems que un test convencional, realizar estimaciones más precisas. Bajo condiciones similares a las de un test convencional (en tiempo requerido y número de ítems aplicados) un TAI permite mayores garantías (menor error de medida) respecto a los niveles de rasgo que se estiman y, por tanto, respecto a las decisiones que se toman a partir de las puntuaciones en los tests.

Estos tres aspectos resultan especialmente relevantes cuando se realizan aplicaciones masivas de tests de rendimiento o de conocimientos, por ejemplo en contextos de selección de personal, de evaluación educativa o en pruebas de certificación profesional o licenciatura. Por citar algunos ejemplos, en Estados Unidos existen versiones adaptativas informatizadas del TOEFL (para evaluar el nivel de inglés), del GRE (prueba de conocimientos para acceder a estudios universitarios), del GMAT (prueba de acceso a Escuelas de Negocios), del ASVAB (batería de aptitudes del Ejército) y de diversos exámenes de acreditación profesional (por ejemplo en Medicina y Enfermería) o de evaluación del nivel educativo de los estudiantes de Primaria y Secundaria. En España existen disponibles varios TAIs: el TRASI (Rubio y Santacreu, 2003) que mide la capacidad de razonamiento secuencial e inductivo; eCAT (Olea, Abad, Ponsoda y Ximénez, 2004) que mide el nivel de comprensión del inglés escrito; y CAT-Health (Rebollo, García-Cueto, Zardaín, Cuervo, Martínez, Alonso, Ferrer y Muñiz, 2009) para la evaluación de la calidad de vida relacionada con la salud. Se están elaborando otros para evaluar el dominio del catalán, euskera, otros idiomas, el ajuste emocional, la satisfacción con los servicios sanitarios, etc.

Aplicaciones vía web

La tecnología informática permite desde hace años su *aplicación a través de internet*. Por poner algunos ejemplos, se aplican a través de la web determinadas baterías neuropsicológicas, tests de conocimientos del idioma inglés, tests predictivos del rendimiento laboral, tests de conocimientos escolares, cuestionarios de personalidad aplicados en contextos clínicos o cuestionarios sobre adicciones a drogas (la información puede completarse en Bartram y Hambleton, 2006).

Tanto el test como los algoritmos de presentación y los resultados se almacenan y distribuyen desde un servidor, lo que permite un mayor control sobre los procesos de aplicación y una información inmediata sobre los resultados. La conexión a través de internet representa también importantes beneficios logísticos: una mayor accesibilidad a los evaluados (por ejemplo, en procesos de reclutamiento para la selección de personal o en casos de intervención psicológica de personas que residen lejos de los servicios de tratamiento) y, en algunos casos, un abaratamiento de costes (piénsese por ejemplo en la aplicación de tests a muestras numerosas de evaluados que viven en diferentes zonas geográficas de un país).

La aplicación a través de internet también supone ventajas para los editores de tests, ya que les permite tener acceso directo a bases de datos que permitan realizar los siempre necesarios estudios de validez de las puntuaciones y de "seguimiento" de las propiedades psicométricas de la prueba. Además, permite controlar que el "cliente" (por ejemplo, la empresa o institución que demanda la aplicación) tenga acceso únicamente a la información que resulte pertinente. Por ejemplo, ya no se requiere incluir plantillas de corrección, lo que implica una mayor garantía de seguridad.

Sin embargo, la utilización de internet como medio de transporte de los tests y de las respuestas de los evaluados requiere tener en cuenta algunas consideraciones en relación a varios riesgos:

- ✔ *Calidad.* Cualquiera puede acceder a centenares de tests que se ofrecen en todo el mundo y de los que desconocemos sus propiedades psicométricas. Como en otros muchos temas, un psicólogo competente debería saber filtrar bien los instrumentos de evaluación disponibles en la web que auténticamente han demostrado su utilidad, de aquellos que sirven únicamente como pasatiempos.

- ✔ *Seguridad.* Un importante problema es el de la seguridad del propio test, sobre todo cuando las puntuaciones en los tests tienen importantes consecuencias para los evaluados (admisión a un centro educativo, a un puesto de trabajo, acreditación profesional, etc.). En el caso del examen GRE, aplicado hace años vía internet, la empresa responsable de la prueba decidió volver a versiones de lápiz y papel tras comprobar la gran cantidad de ítems que los evaluados de ciertos países asiáticos conocían de antemano, debido a su transmisión en ciertos foros. Como es lógico, el acceso a los contenidos del test y a la información que proporcionan los evaluados debe ser seguro y controlado. A veces internet puede entrar en colisión con la Ley de Protección de Datos.
- ✔ *Control*. Otro problema importante tiene que ver con las posibilidades de suplantación de identidad, es decir, que sean otras personas las que respondan al test. Una posible solución sería la aplicación controlada por supervisores que aseguren la identidad de los evaluados, que asignen las contraseñas oportunas de acceso y que controlen el cumplimiento de las condiciones de aplicación.
- ✔ *Garantías tecnológicas.* La aplicación informatizada puede suponer una amenaza a la validez de las puntuaciones si las condiciones de evaluación no están estandarizadas. Por ejemplo, algunos tests que incluyen información dinámica y tiempos limitados de respuesta son muy susceptibles a la velocidad de transmisión de la información por la red y a las características del ordenador y conexión que tiene cada evaluado.

Por otro lado, conviene no olvidar que las propiedades de un test no dependen sólo de los ítems que se aplican sino también de cómo se aplican (por ejemplo, que el evaluador genere una situación adecuada de evaluación, que se responda a las dudas que surjan, que se garantice que el evaluado dedica el tiempo adecuado a las instrucciones, etc.). La necesidad de un supervisor directo de la aplicación puede depender del tipo de test (rendimiento óptimo vs. rendimiento típico) y de las consecuencias de la aplicación evaluación, entre otros aspectos.

Éstos y otros problemas han requerido la elaboración de directrices sobre buenas prácticas en el diseño y aplicación de tests informatizados, reservando recomendaciones específicas para los que se aplican a través de internet (ITC, 2005) y que plantean demandas adicionales en el control de calidad de este tipo de tests. Hay que determinar los requisitos mínimos de software y hardware, establecer mecanismos de prevención y detección de errores en la administración, prevenir y detectar brechas en la seguridad, determinar el nivel de supervisión en la aplicación, establecer controles de identificación del evaluado, garantizar el almacenamiento seguro de las respuestas, chequear periódicamente las propiedades psicométricas de los ítems, etc. En el ámbito más estrictamente psicométrico, las directrices establecen que un test informatizado debe incorporar la oportuna información psicométrica (fiabilidad y validez) y debe garantizar que no requiere otros conocimientos o destrezas (por ejemplo, la familiaridad con los ordenadores) diferentes a las que exige el test. Estas directrices se pueden consultar en la dirección de la ITC: http://www.intestcom.org/guidelines/index.php.

OTROS NUEVOS TIPOS DE TESTS

Tests basados en modelos

Un modo de obtener información sobre las inferencias que podemos realizar con las puntuaciones de un test es analizar los procesos, estrategias y estructuras de conocimiento que están implicados en la resolución de los ítems. Bejar (2002) emplea la denominación de *tests basados en modelos* para referirse al diseño de instrumentos de evaluación guiados por una teoría psicológica sobre el procesamiento de respuestas.

Un excelente muestrario de este tipo de tests se incluye en el libro de Irvine y Kyllonen (2002) *"Item generation for test development"* donde se recoge el progresivo acercamiento entre Psicología Cognitiva y Psicometría, lo que se ha traducido en la elaboración de tests de razonamiento cuantitativo, razonamiento analítico, visualización, analogías verbales, etc. El primer paso en la construcción de este tipo de pruebas es un análisis de los procesos cognitivos que demanda la resolución de la tarea y un estudio detallado de cuáles son las características del ítem que, en función de esos procesos, determinan su diferente nivel de demanda cognitiva y, por tanto, su dificultad. Por ejemplo, Hornke (2002) describe un test de rotación de figuras donde se manipulan variables como la cantidad de elementos a procesar, si las figuras son bi o tridimensionales, el ángulo de la rotación o el número y tipo de rotaciones (de derecha a izquierda, de arriba abajo...). Describe también un test de memoria visual donde los ítems son planos de una ciudad donde aparecen determinados iconos para representar ciertos servicios públicos, manipulándose en cada caso la cantidad de iconos, su tamaño o el nivel dispersión en el mapa.

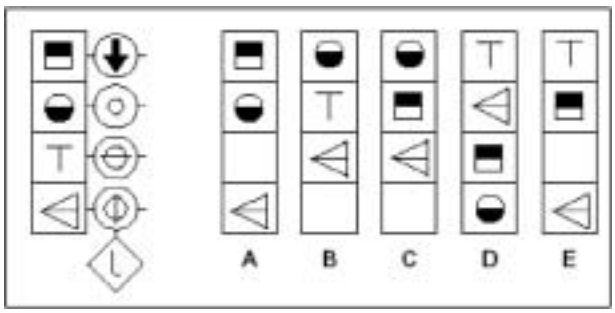
En España, Revuelta y Ponsoda (1998) desarrollaron un test basado en un modelo cognitivo para el test DA5. Los 50 ítems del test pretenden medir la capacidad de razona-

miento lógico mediante tareas que incluyen un conjunto de instrucciones (símbolos dentro de los círculos y del rombo) sobre lo que debe hacerse mentalmente con la figura adyacente correspondiente (véase figura 4). Un ítem consta de varias figuras (columna de cuatro cuadrados a la izquierda de la Figura 4, que contiene cada uno una figura), las instrucciones de los cambios que se han de hacer con cada figura (columna de círculos y rombo), y de las cinco posibles respuestas (columnas A, B … E). La tarea del evaluado es aplicar a las figuras las instrucciones y elegir la opción correcta de las cinco posibles. Las instrucciones pueden requerir, por ejemplo, girar la figura cierto número de grados, intercambiar la posición con la figura anterior, omitirla, ignorar otras instrucciones o reordenar de determinada forma todas las figuras.

Un modelo de procesamiento asume que el evaluado codifica la primera figura (la que aparece en el primer cuadrado de la primera columna) y la instrucción, aplica la instrucción sobre la figura (en ejemplo, la instrucción indica que la figura ha de desplazarse un cuadrado hacia abajo, por lo que sólo las opciones C y E podrían ser correctas), y sigue secuencialmente con las demás figuras hasta alcanzar la solución. Se estudió la aportación de cada una de las instrucciones (y de las veces que es necesario aplicarlas) a la dificultad de los ítems, mostrando más peso en la predicción de la dificultad las instrucciones que requerían reordenar las 4 figuras mentalmente.

Una aportación novedosa de esta manera de proceder es que si conocemos las variables que intervienen en los procesos de respuesta, puede establecerse un método para construir todo el universo posible de ítems gobernado por dichas variables. El procedimiento, denominado "*generación automática de ítems*" (GAI), consiste en la construcción de bancos de ítems mediante algoritmos. En la GAI se establece un conjunto de reglas explícitas, susceptibles de programarse en un ordenador, que determinan cómo deben construirse los ítems. Por ejemplo,

FIGURA 4



Revuelta y Ponsoda (1998) generaron los 4.242 ítems posibles que tienen su base en el DA5, combinando el tipo de figuras, las instrucciones a aplicar y determinados criterios para generar opciones incorrectas de respuesta. Si el modelo que describe los procesos de respuesta de los ítems es correcto, resultará posible conocer la dificultad de nuevos ítems antes de que hayan sido aplicados a persona alguna. Son muy importantes las ventajas de disponer de todo el banco posible de ítems, principalmente para garantizar que se mide con elevada precisión cualquier nivel de capacidad.

Tests ipsativos

Fundamentalmente en contextos de selección de personal, el falseamiento de respuestas a los tests de personalidad es un problema que se ha intentado resolver de varias formas. Una de las más alentadoras es la elaboración de tests ipsativos, que obligan al evaluado a elegir entre opciones de respuesta que tienen un nivel similar de deseabilidad y que se refieren a dimensiones diferentes de personalidad. Por ejemplo, el aspirante puede tener que elegir entre *"soy una persona trabajadora"* (responsabilidad) y *"soy una persona abierta"* (extraversión). El proceso de diseño de un test ipsativo es básicamente el siguiente:

a. Determinar las dimensiones a evaluar y los ítems iniciales que las definen.
b. Diseñar con estos ítems iniciales un test normativo convencional. Conviene realizar estudios factoriales para determinar empíricamente los ítems que forman cada dimensión y, en su caso, eliminar los ítems que no saturan en el factor previsto.
c. Establecer el número de opciones de cada ítem ipsativo. Lo más simple es establecer ítems binarios, cada uno formado por dos ítems iniciales.
d. Realizar un estudio empírico donde una muestra apropiada de jueces valore el nivel de deseabilidad de cada ítem inicial. A partir de estos juicios se obtienen valores en deseabilidad para cada uno de los ítems iniciales.
e. Diseñar el test ipsativo, considerando que en los ítems deben todas las posibles combinaciones de dimensiones. En cada ítem ipsativo deben incluirse opciones (ítems iniciales) de similar deseabilidad. Cada dimensión se debe comparar con cualquier otra un número similar de veces.
f. Establecer el sistema de puntuación de los evaluados, por ejemplo, contando las veces que eligen las opciones de cada una de las dimensiones.

Ejemplo del proceso de elaboración de un test ipsativo
(Abad, Olea, Ponsoda y Garrido, 2007)

1) Dimensiones a evaluar: las 5 dimensiones de personalidad definidas en el modelo Big-Five, cada una evaluada mediante 18 adjetivos.
2) Test normativo. aplicación de los 90 ítems a una muestra según un formato de 5 categorías ordenadas, pidiendo el grado en que cada uno le describe a la persona.
3) Estudio factorial: se retuvieron los 12 ítems de cada dimensión que mayor saturación manifestaron en el factor previsto, con lo que el test definitivo constaba de 60 ítems.
4) Obtención de índices de deseabilidad (ID). una muestra de personas valoró (de 1 a 4) el grado en que cada adjetivo indicaba una cualidad positiva para ser eficiente en un determinado puesto laboral. Las medias de estas valoraciones se consideraron como índices de deseabilidad de los ítems. El adjetivo de menor media fue "corriente" (ID = 1,93) y el de mayor media "organizado" (ID =3,87).
5) Diseño del test ipsativo. Se decidió construir un test de 30 ítems ipsativos, cada uno formado por dos adjetivos de dimensiones distintas y similar ID. Por ejemplo, uno de los ítems era "estable-enérgico" que, respectivamente, se refieren a las dimensiones de estabilidad emocional y extraversión, y que obtuvieron valores en ID de 3,71 y 3,43. Según este diseño, cada dimensión se comparaba 3 veces con las otras 4 dimensiones de personalidad restantes.
6) Puntuación en el test ipsativo. Para puntuar a cada sujeto en cada una de las 5 dimensiones, se sumaron las veces que en los pares de adjetivos se elegían los ítems de cada una de ellas. Por tanto, la puntuación máxima teórica en una dimensión fue 12, mientras que la mínima 0.
7) Se realizaron estudios de validez convergente y predictiva (correlaciones con calificaciones en cursos de formación), mostrando la mayor capacidad predictiva algunos ítems ipsativos que combinaban adjetivos de las dimensiones de estabilidad emocional y responsabilidad.

En las últimas décadas, los tests ipsativos han tenido momentos de auge y declive, con defensores y detractores que con igual fuerza argumentan sus beneficios o problemas. Algunos de estos problemas son:

a. El modo de puntuar ipsativamente a un sujeto en las diferentes dimensiones provoca interdepencias entre éstas: una puntuación muy alta en una dimensión necesariamente conlleva puntuaciones no elevadas en las restantes. Este problema es tanto mayor cuanto menor el número de dimensiones. De forma más general, el promedio de las correlaciones entre *m* dimensiones se acerca a *-1/(m-1*), siendo *m* el número de dimensiones (Meade, 2004). En el caso de medir dos únicas dimensiones, la correlación entre ambas sería necesariamente -1. El modo ipsativo de puntuación lleva además a que sea cero la suma de las covarianzas de las dimensiones con una variable externa (por ejemplo, un criterio) y a distorsiones en los coeficientes de fiabilidad para las puntuaciones en las dimensiones. Todo esto exige un tratamiento psicométrico específico de los datos ipsativos (no es raro, por ejemplo, que las soluciones factoriales de datos normativos e ipsativos del mismo test sean diferentes), que actualmente es objeto de investigación.
b. Conceptualmente, un test ipsativo plantea una tarea de preferencias y, por tanto, permite la comparación entre escalas dentro de una persona (por ejemplo, podría decirse que una persona es más responsable que extravertida) pero no entre distintas personas (que una persona sea más responsable que otra). Por ello, su uso está más indicado en las medidas de atributos que impliquen preferencias, como es usual en la medición de los intereses.
c. No es claro que sean resistentes al falseamiento ya que los aspirantes pueden ser conscientes de cuáles son las dimensiones deseables para el puesto.

No nos parece muy recomendable por el momento la aplicación de tests ipsativos si se pretende realizar comparaciones de rendimiento entre diferentes evaluados, dado que resulta complicado estudiar sus propiedades psicométricas mediante modelos y técnicas usuales. Sin embargo, vemos una importante potencialidad a este tipo de tests (algunos estudios han mostrado ya una mayor validez predictiva que los tests usuales de personalidad) cuando se consoliden algunos intentos que se están realizando en el ámbito de la investigación psicométrica para modelar teóricamente las respuestas a este tipo de ítems (Stark, Chernyshenko y Drasgow, 2005). En cualquier caso, la cuestión está lejos de ser resuelta.

Tests conductuales

En el contexto de la medición de la personalidad, existe una línea teórica de evaluación comportamental de la personalidad donde se estudian los estilos interactivos o tendencias de comportamiento constantes ante situaciones determinadas (Santacreu, Rubio y Hernández, 2006). Desde esta perspectiva se diseñan tests comportamentales informatizados para medir, por ejemplo, la tendencia al riesgo (propensión a elegir las opciones más recompensadas a pesar de ser poco probables) mediante simulaciones de juegos de ruleta o dados, o mediante tareas de toma de decisiones más o menos proclives a accidentes. En la figura 5 se muestra tarea consistente en decidir cuándo cruzar la calle para ir lo más rápido posible a una farmacia, cambiando en los sucesivos ensayos la ubicación de la persona y sabiendo que puede aparecer un coche del túnel. Si el peatón se

encuentra muy a la izquierda aumenta la probabilidad de que sea atropellado (menor visibilidad) pero también se reduce el tiempo para llegar a la farmacia. Lo más seguro es moverse hacia la derecha y cruzar, pero eso conlleva un mayor tiempo. Tras cada ensayo, el evaluado recibe feedback sobre el tiempo que ha tardado en llegar pero no sobre si ha sido atropellado. La tendencia al riesgo se obtiene calculando la media en los sucesivos ensayos de la distancia entre la persona y la farmacia (mayor media, menor tendencia al riesgo). Obviamente, este modo de proceder es muy distinto a aplicar tests de personalidad donde las personas informan de su tendencia a la búsqueda de sensaciones o su nivel de apertura, tal como se hace en los tradicionales tests de personalidad. Los profesionales que optan por este tipo de tests consideran que una de sus ventajas tiene que ver con la eliminación de los problemas de deseabilidad.

Tests situacionales

Consisten en describir ciertas situaciones (por ejemplo, en el ámbito laboral) y pedir a los sujetos que digan cómo creen que reaccionarían ante dichas situaciones. Parece que este tipo de pruebas añaden poder predictivo de la eficacia laboral a los tradicionales tests de capacidad cognitiva y de personalidad (por eso se aplican cada vez más frecuentemente), aunque son escasos los estudios que se han realizado sobre su eficacia para reducir el falseamiento de respuestas. Pueden realizar descripciones en un formato de respuesta abierta o, lo que es más usual, elegir entre varias posibilidades que se describen de antemano. A continuación se presenta un ejemplo de un ítem situacional sobre integridad (Becker, 2005). Entre corchetes se especifica el modo de puntuación de las respuestas, establecido a partir de las opiniones de expertos:

Tu equipo de trabajo está en una reunión debatiendo sobre cómo vender un producto nuevo. Todos parecen estar de acuerdo en que el producto debe ser ofertado a los clientes en el presente mes. Tu jefe tiene mucho interés en que sea así, y tú sabes que a él no le gustan los desacuerdos en público. Sin embargo, tú tienes reservas porque un informe reciente del departamento de investigación apunta hacia diversos problemas potenciales de seguridad. ¿Cuál crees que sería tu reacción?

A. Tratar de comprender por qué todos los demás quieren ofertar el producto a los clientes en este mes. Tal vez tus preocupaciones están fueran de lugar. [-1]

B. Expresar tus preocupaciones con el producto y explicar por qué crees que las cuestiones de seguridad necesitan ser abordadas. [1]

C. Mostrarte de acuerdo con lo que los demás quieren hacer para que todos se sientan bien acerca del equipo. [-1]

D. Después de la reunión, hablar con algunos de los otros miembros del equipo para ver si ellos comparten tus preocupaciones. [0]

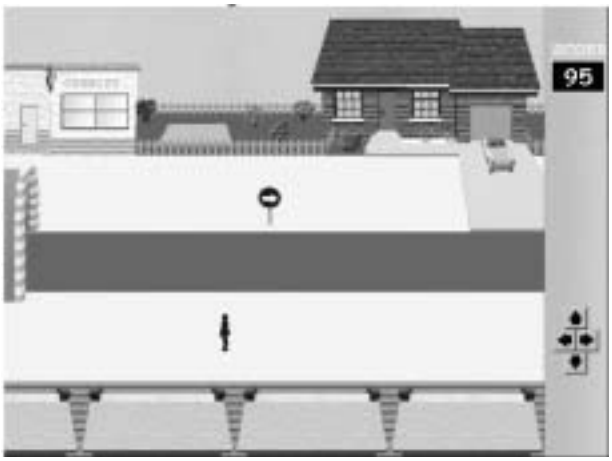
Desde un punto de vista psicométrico, un tema especialmente relevante es cómo puntuar de la mejor forma las respuestas a este tipo de ítems. Bergman, Donovan, Drasgow, Henning y Juraska (2006) estudiaron los diferentes efectos que tienen 11 modos diferentes de puntuar los ítems de un test situacional para evaluar la capacidad de liderazgo, formado por 21 ítems que se presentan mediante video y que tienen cuatro opciones diferentes de respuesta según el grado de participación en la toma de decisiones.

ALGUNOS RIESGOS ADICIONALES, ALGUNOS RECURSOS

Las importantes ventajas que tienen para el psicólogo los nuevos tipos de tests no pueden quedar empañadas por ciertos riesgos que queremos subrayar.

En primer lugar, debe enfatizarse que las nuevas tecnologías no son "per se" garantes de mejores mediciones. La eficiencia de los nuevos procedimientos de respuesta y procesamiento de la información no puede sustituir al necesario escrutinio psicométrico de las puntuaciones asignadas. La apariencia de validez de los nuevos formatos de ítems debe ir acompañada de evidencias empíricas de validez. Por ejemplo, un ítem "multimedia" puede ser más informativo que un ítem clásico de opción múltiple pero puede requerir mucho más tiempo para su resolución. Por otro lado, las demandas de precisión requerida pueden ser distintas si el objetivo es clasificar a una persona (apto vs. no apto) o si el objetivo es cuantificar su nivel de rasgo. También es necesario reflexionar sobre cuándo merece la pena aplicar un TAI y cuándo no (Wainer, 2000). Por ejemplo, si el test

FIGURA 5



es de altas consecuencias para los evaluados, se va a aplicar una o dos veces al año y el contenido no requiere una aplicación informatizada, los costes de un TAI (necesidad de crear y mantener grandes bancos de ítems, de desarrollar, evaluar y actualizar el software, disponibilidad de ordenadores para la aplicación, etc.) pueden superar a los beneficios.

Un segundo riesgo tiene que ver con el olvido de ciertas áreas de aplicación de los tests. Desarrollar nuevos tipos de tests es costoso y se corre el riesgo de avanzar casi exclusivamente en contextos aplicados (organizacionales o educativos) donde más recursos económicos se invierten o donde más se precisan soluciones tecnológicas eficientes. Los avances no deben olvidar determinados contextos de medición estrictamente propios de nuestra profesión, como son la evaluación clínica o la evaluación de programas de intervención psicosocial. En este sentido, la conjunción entre los nuevos tipos de tests, los nuevos modelos psicométricos y los modelos estadísticos para medir el cambio conseguido por las intervenciones debería ser un terreno fructífero para proyectos de I+D.

Un tercer riesgo se refiere al mal uso de los nuevos tests. Debido a su inmediata disponibilidad, los nuevos tipos de tests pueden aplicarse en contextos inadecuados, por personas no preparadas y realizando inferencias erróneas a partir de las puntuaciones que proporcionan.

¿Qué puede hacer el profesional de la Psicología para incrementar su competencia en los nuevos modos de medir? Más que nunca se necesita una formación continua a lo largo de la vida profesional para estar al tanto de las innovaciones que aceleradamente se van produciendo para mejorar la medición psicológica. Aún resultando ciertamente atrevido dar consejos, el profesional competente podría comenzar leyendo alguno de los libros recientes de Psicometría y las revistas especializadas donde se publican los avances psicométricos y las experiencias aplicadas con nuevos tipos de tests (algunas de las españolas más sensibles a estos temas son *Psicothema*, *Psicológica, Revista Electrónica de Metodología Aplicada* o *Spanish Journal of Psychology*). La revista *Psicológica* publicó en el 2000 un número monográfico sobre TAIs. Información en castellano sobre los TAIs puede encontrarse en libros (Olea y Ponsoda, 2003; Olea et al., 1998) y capítulos de libros (Olea y Ponsoda, 1996). En inglés, son muchos los libros sobre tests informatizados (Bartram y Hambleton, 2006; Eggen, 2004; Mills, Potenza, Fremer y Ward, 2002; Parshall, Spray, Kalohn y Davey, 2002; Sands, Waters y McBride, 1997; van der Linden y Glas, 2000; Wainer, Dorans y col., 2000).

Un sencillo tutorial sobre los TAIs puede encontrarse en la dirección http://edres.org/scripts/cat/catdemo.htm y una página fundamental para el investigador, que recoge una amplia información teórica y aplicada sobre los TAIs, es la siguiente: http://www.psych.umn.edu/psylabs/catcentral/.

Puede consultarse también los catálogos de tests disponibles en la web de las principales empresas editoras. Los profesionales interesados en mejorar su formación sobre estos temas pueden asistir a cursos concretos sobre estos temas. Nuestras universidades ofrecen varios. Puede también empezar a manejar algunos programas disponibles para la elaboración y análisis de tests informatizados. Si esta es su opción, no deje de consultar las prestaciones que se ofrecen en la siguiente dirección de la universidad de Málaga (http://jupiter.lcc.uma.es/siette.wiki.es/index.php/Portada) o en la principal distribuidora norteamericana de software psicométrico: http://assess.com. Además del software general para aplicar la TRI, existe software para implementar TAIs (FASTEST y POSTSIM 2.0); mientras que el primero (ASC, 2001) permite la organización de bancos de ítems, ensamblaje de pruebas y aplicación de TAIs, el segundo permite evaluar el funcionamiento psicométrico de un TAI mediante simulación y bajo distintas condiciones (de selección de ítems, de estimación del nivel de habilidad y de criterio de parada). Para tener un "primer contacto" también puede servir el programa ADTEST (Ponsoda, Olea y Revuelta, 1994).

REFERENCIAS

Abad, F.J., Olea, J., Ponsoda, V. y Garrido, L. (2007). *Test POLIPSA: Informe técnico y propiedades psicométricas.*

ASC (2001). *The FastTEST Professional Testing System, Version 1.6.* [*Computer software*]. St. Paul, MN: Author.

Bartram, D. y Hambleton, R. K. (2006). *Computer-based testing and the internet issues and advances*. Chichester, West Susex: Wiley.

Becker, T. E. (2005). Development and validation of a situational judgment test of employment integrity. International Journal of Selection and Assessment, 13(3), 225-232.

Becker TE. (2005). Development and validation of a situational judgment test of employee integrity. *International Journal of Selection and Assessment, 13,* 225–232.

Bejar, I.I. (2002). Generative testing: From conception to implementation. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NJ: LEA.

Bergman, M.E., Drasgow, F., Donovan, M.A., y Henning, J.B. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment*, *14*, 223-235.

Conejo, R., Guzmán, E., Millán, Trella, M., Pérez-de-la-Cruz, L. y Rios, A., (2004): SIETTE: A web-based tool for adaptive testing. *International Journal of Artificial Intelligence in Education*, 14, 29-61.

Davey, T. (2005). Computer-based testing, En B. S. Everitt y D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: Wiley.

Drasgow, F. y Olson-Buchanan, J. (1999). *Innovations in computerized assessment*. Mahwah, NJ: LEA.

Eggen, T. J. H. M. (2004). *Contributions to the theory and practice of computerized adaptive testing.* Arnhem, Holanda: Citogroep.

Hornke, L. F. (2002). Item-generative models for higher order cognitive functions. En S. H. Irvine y P. C. Kyllonen (Eds.), Item generation for test development. New Jersey: Lawrence Erlbaum Associates.

International Test Comission (2005). *International Guidelines on Computer-Based and Internet Delivered Testing.* Recuperado el 30 de junio de 2005, http://www.intestcom.org.

Irvine, H. y Kyllonen, P.C. (2002), *Item generation for test development*. New Jersey: LEA.

Meade, A. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531-552.

Mills, C. N., Potenza, M. T., Fremer, J. J. y Ward, W. C. (2002). *Computer-based testing: Building the foundation for future assessments.* Mahwah, NJ: LEA

Muñiz, J. (2010). Las teorías de los tests: Teoría Clásica y Teoría de Respuesta a los Ítems. *Papeles del Psicólogo, 31(1)*, 57-66.

Muñiz, J. y Fernández-Hermida, J.R. (2010) La Opinión de los Psicólogos Españoles sobre el Uso de los Tests. *Papeles del Psicólogo, 31(1)*, 108-122.

Olea, J. y Ponsoda, V. (1996). Tests adaptativos informatizados. En J. Muñiz (Coor.), *Psicometría*. Madrid: Universitas.

Olea, J. y Ponsoda, V. (2003). *Tests adaptativos informatizados*. Madrid: UNED.

Olea, J., Ponsoda, V. y Prieto, G. (1999). *Tests informatizados Fundamentos y aplicaciones.* Madrid: Pirámide.

Olea, J., Abad, F. J., Ponsoda, V. y Ximenez, M. C. (2004). A computerized adaptive test for the assessment of written English: Design and psychometric properties. *Psicothema*, *16*, 519-525.

Parshall, C. G., Spray, J. A., Kalohn, J. C., y Davey, T. (2002). *Practical considerations in computer-based testing*. Nueva York: Springer.

Ponsoda, V., Olea, J., y Revuelta, J. (1994). ADTEST: A Computer adaptive test based on the maximum information principle. *Educational and Psychological Measurement. 54,* 3, 680-686.

Quintana, J., Bitaubé, A. y López-Martín, S. (2008). *El lugar de la Psicología en la universidad española del siglo XX*. UAM Ediciones: Madrid.

Rebollo, P., García-Cueto, E., Zardaín, J.C., Martínez, I., Alonso, J., Ferrer, M. y Muñiz, J. (2009). Desarrollo del CAT-Health, primer test adaptativo informatizado para la evaluación de la calidad de vida relacionada con la salud en España. *Medicina Clínica*, 133, 7, 241-251.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema*, *10*, 3, 709-716.

Rubio, V. y Santacreu, J. (2003). *TRASI Test adaptativo informatizado para la evaluación del razonamiento secuencial y la inducción como factores de la habilidad intelectual general.* Madrid: TEA ediciones.

Sands, W. A., Waters, B. K. y McBride, J. R. (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Santacreu, J., Rubio, V.J., y Hernández, J.M. (2006). The objective assessment of personality: Cattell's T-data revisited and more. *Psychology Science*, 48, 53-68.

Stark, S, Chernyshenko, O.S., Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multiunidimensional pairwise-preference model. *Applied Psychological Measurement*, *29*, 184–203.

van der Linden, W. J. y Glas, C. A. W. (2000). *Computerized adaptive testing: Theory and practice*. Londres: Kluwer Academic.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B., Mislevy, R., Steinberg, L. y Thissen, D. (2000). *Computerized adaptive testing: A primer (2ª ed.)*. Mahwah, NJ: LEA.

Wainer, H. (2000). CAT: Wether and Whence. *Psicologica*, *21*, 121-133.

Williamson, D. M., Mislevy, R. J. y Bejar, I. I. (2006). *Automated Scoring of Complex Tasks in Computer Based Testing.* Mahwah, NJ: LEA.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 108-121
http://www.cop.es/papeles

LA OPINIÓN DE LOS PSICÓLOGOS ESPAÑOLES SOBRE EL USO DE LOS TESTS

THE OPINION OF SPANISH PSYCHOLOGISTS ON THE USE OF TESTS

José Muñiz y José Ramón Fernández-Hermida
Universidad de Oviedo

Las organizaciones nacionales e internacionales interesadas en mejorar la práctica de los tests siguen dos líneas de actuación complementarias, por un lado se trata de restringir el uso de los tests a aquellos profesionales preparados para ello, y por otro se intenta difundir todo tipo de información técnica sobre los tests y su adecuada utilización. Para una correcta aplicación de estas dos estrategias es fundamental conocer de forma rigurosa las opiniones de los psicólogos profesionales sobre la práctica de los tests. Con tal fin la Comisión de Tests de la Federación Europea de Asociaciones de Psicólogos (EFPA) ha desarrollado un cuestionario compuesto por 33 ítems. En el presente trabajo se recogen las respuestas de los psicólogos españoles a la encuesta de la EFPA. Respondieron 3.126 psicólogos profesionales, 2.235 mujeres (71,5%) y 891 hombres (28,5%), todos ellos miembros del Colegio Oficial de Psicólogos. La edad media fue de 41,92 años y la desviación típica de las edades 10,43. La media de años en la profesión fue de 12,5, con una desviación típica de 8,9. El 69,6% pertenecían al ámbito de la psicología clínica, el 13,6% a educativa, el 6,4% a trabajo y el 10,4% a otras especialidades, tales como deporte, jurídica, tráfico, o servicios sociales, el 3,8% están desempleados. Los resultados se articulan en torno a ocho grandes dimensiones, que se analizan con detalle en función de las especialidades de Clínica, Trabajo y Educativa. Los psicólogos muestran una actitud general muy positiva hacia la utilización de los tests en el ejercicio de su profesión, si bien ponen de manifiesto algunos puntos débiles y limitaciones que deben ser mejorados cara al futuro. Se finaliza comentando los resultados en detalle y analizando las perspectivas de futuro.
Palabras clave: *Tests, Uso de los tests, Opinión psicólogos, EFPA.*

National and international psychological organizations interested in improving tests and testing practices follow two complementary strategies. On one hand they try to restrict the use of tests to those professionals who have been properly trained in the field of tests and testing, and on the other, the dissemination of information on tests and testing is encouraged. In order to implement both strategies in a rigorous way it is essential to know the opinions of professional psychologists. To this end the European Federation of Psychologists' Associations (EFPA) has developed a questionnaire composed of 33 items. In this paper we present the answers of the Spanish psychologists to the EFPA questionnaire. 3.126 psychologists answered the questionnaire, 2.235 women (71,5%), and 891 men (28,5%), all of them members of the Spanish Psychological Association (COP). The mean age was 41,92 years, and standard deviation 0,43. The mean of years working as professionals was 12,5, standard deviation 8,9. In relation of the field of specialization, 69,6% work in Clinical Psychology, 13,6% in Educational Psychology, 6,4% in Work and Organizational Psychology, and 10,4% in other fields, such as sports, forensic, social services, or traffic. 3,8% are unemployed. The results are articulated around eight main dimensions, which are commented in detail comparing the results obtained in the main fields of specialization, Clinical, Educational and Work psychology. Psychologists show a very positive attitude towards tests and testing, however different aspects that most be improved in the future are pointed out as well. Finally the results are analyzed in detail, and some future perspectives commented.
Key words: *Tests, Testing practices, Opinion psychologists, EFPA.*

Hace ahora diez años, a instancias de la Comisión de Tests de la Federación Europea de Asociaciones de Psicólogos (EFPA), se llevó a cabo una encuesta en seis países europeos, entre ellos España, para conocer las opiniones de los psicólogos europeos sobre distintos aspectos relacionados con los tests (Muñiz y Fernández-Hermida, 2000; Muñiz et al., 2001). El objetivo fundamental era conocer de primera mano las opiniones de los psicólogos profesionales europeos sobre distintas cuestiones relacionadas con la práctica de los tests, para así poder organizar acciones y proyectos encaminados a mejorar su uso. Una síntesis de los proyectos y acciones llevadas a cabo en los últimos años por la EFPA y por la Comisión Internacional de Tests (ITC) pueden consultarse en el trabajo de Muñiz y Bartram (2007). Pasada una década desde la primera encuesta, la Comisión de Tests de la EFPA consideró oportuno volver a evaluar las opiniones de los psicólogos

Correspondencia: José Muñiz. *Facultad de Psicología. Universidad de Oviedo. Plaza Feijoo, s/n. 33003 Oviedo. España. E-mail: jmuniz@uniovi.es*

europeos sobre los tests. Aparte de hacer un seguimiento de los resultados obtenidos entonces, a lo largo de esta última década han irrumpido con fuerza dos avances técnicos que están teniendo una gran incidencia sobre la forma en la que ejercen su profesión los psicólogos en general, y en particular sobre la utilización de los tests, nos referimos a Internet y a las nuevas tecnologías, en especial todo lo relacionado con los avances informáticos. Por lo tanto en la nueva encuesta, aparte de mantener las cuestiones claves de hace diez años, se van a incluir varias preguntas relativas a la incidencia de estos dos factores en la práctica diaria de los profesionales en relación con los tests, tales como ¿los tests informatizados están sustituyendo a los de papel y lápiz en la práctica diaria de los psicólogos? ¿Constituye Internet un claro avance en la evaluación psicológica actual?

Pero antes de entrar en la descripción de la nueva encuesta y en los resultados obtenidos, conviene echar una ojeada y repasar lo que se ha venido haciendo en los últimos años, tanto a nivel europeo como español, para tratar de mejorar el uso de los tests. La utilización ética y deontológica de los tests se asienta en dos pilares básicos, por un lado los tests han de tener unas propiedades psicométricas adecuadas, y por otro, la utilización que se lleve a cabo debe de ser la correcta, desde su aplicación y corrección hasta el uso que se haga de las puntuaciones. Las organizaciones que dedican sus esfuerzos a mejorar el uso de los tests, tanto nacionales (COP), como internacionales (EFPA, ITC, o la Asociación Americana de Psicología, APA), llevan a cabo distintas acciones y proyectos variados que pueden articularse en torno a dos grandes estrategias que podemos denominar restrictiva e informativa.

La estrategia restrictiva se refiere a las acciones llevadas a cabo para limitar el uso de los tests a aquellos profesionales que están realmente preparados para hacerlo. Los sistemas utilizados varían de unos países a otros (Bartram, 1996; Bartram y Coyne, 1998; Muñiz, Prieto, Almeida y Bartram, 1999), si bien uno de los más habituales en varios países, incluida España, es clasificar los tests siguiendo los criterios de la APA en tres categorías (A, B, C) de menos a más especializados, siendo exclusivo de los psicólogos el uso de los tests de las categorías B (tests colectivos de carácter cognoscitivo y Personalidad) y C (escalas individuales y tests proyectivos). Otra opción también utilizada es que los profesionales obtengan una certificación específica en la que acrediten fehacientemente que conocen adecuadamente las pruebas. Aunque estas restricciones y otras son recomendables, no garantizan por sí solas un uso adecuado de los tests (Moreland, Eyde, Robertson, Primoff y Most, 1995; Simner, 1996), siendo necesario complementar esta estrategia con la difusión de información a todas las partes implicadas, tales como profesionales, usuarios, instituciones, y sociedad en general.

Las acciones llevadas a cabo en el marco de la estrategia que hemos denominado informativa, se refieren a todo tipo de iniciativas encaminadas a difundir información sobre la práctica de los tests. Se entiende que cuanta más información posean los profesionales, los usuarios, las familias, y en general todas las partes implicadas en el uso de los tests, menor será la probabilidad de que se haga un mal uso de las pruebas. En este sentido distintas organizaciones nacionales e internacionales ha desarrollado códigos éticos y deontológicos, así como directrices varias para guiar el uso adecuado de los tests. Entre los primeros cabe destacar el meta código ético de la EFPA (2005), el código desarrollado por el comité norteamericano para la buena práctica de los tests (2002), o las directrices de la asociación europea de evaluación psicológica (Fernández-Ballesteros et al., 2001). Véanse buenas revisiones en autores como Koocher y Keith-Spiegel (2007), Lindsay, Koene, Ovreeide y Lang (2008), o Leach y Oakland (2007), y sobre todo en el último número monográfico dedicado al tema por la revista *Papeles del Psicólogo* (2009). Aparte de estos códigos disponemos en la actualidad de un conjunto de directrices que marcan los pasos a seguir desde la propia construcción de la prueba, su aplicación, interpretación y aplicación de los resultados (Bartram, 1998; Brennan, 2006; Downing y Haladyna, 2006; Muñiz, 1997). Merecen mención especial los estándares técnicos desarrollados por la Asociación Americana de Psicología y otras dos organizaciones (APA, AERA y NCME, 1999), así como las directrices elaboradas por la Comisión Internacional de Tests (ITC) para la traducción y adaptación de los tests de unas culturas a otras (Hambleton, Merenda y Spielberger, 2005). Ambas directrices se encuentran en la actualidad en proceso de revisión, por lo que no tardarán en aparecer las nuevas versiones. Para consultar otras directrices sobre el uso de los tests en general, de los tests informatizados e Internet, o la utilización de los tests en el ámbito del trabajo y las organizaciones, véase, por ejemplo, el trabajo de Muñiz

y Bartram (2007), o la página web del la ITC (www.intestcom.org) y de la EFPA (www.efpa.eu). También en la página web del Colegio Oficial de Psicólogos (COP), en el apartado de la Comisión de Tests, se puede consultar información de interés (www.cop.es). Aparte de los códigos éticos y las directrices, hay dos medidas que merecen atención dentro de las acciones enmarcadas en la estrategia de la información, se trata por un lado de una nueva norma ISO que está a punto de ser publicada y que va a regular todo lo relativo a la evaluación de personas en contextos laborales, y por otro, los modelos de evaluación de tests desarrollados en distintos países, entre ellos España. Se comentan a continuación ambas propuestas.

Norma ISO 10667

Las siglas ISO se refieren a la organización internacional para la estandarización (www.iso.org), que desarrolla normativas en todos sectores industriales y de servicios, en España su correspondiente es AENOR (www.aenor.es). A iniciativa de los representantes alemanes se inició un proceso para elaborar una nueva norma ISO que regulase todo lo relativo a la evaluación de las personas en el ámbito laboral. Como es fácil de entender esta nueva norma es de gran interés para los psicólogos, dado su papel central en la evaluación de personas en contextos laborales, por ello el COP ha participado activamente en la comisión internacional que desarrolla esta norma, junto con otras asociaciones nacionales de psicología como la americana (APA) o la británica (BPS), por citar sólo dos. Tras varias reuniones, ya se dispone de un texto bastante consensuado a falta de algunos retoques. Estas normas ISO tienen una gran importancia, pues una vez aprobada las empresas e instituciones podrán certificarse garantizando que cumplen con la norma, no tiene rango legal en sentido estricto pero constituye una importante norma reguladora del mercado, pues no será lo mismo estar certificado que no estarlo. Si bien aún no se dispone de un texto definitivo publicado, el objetivo de la norma es la regulación del proceso de evaluación de las personas en contextos laborales y organizacionales, cubriendo todo el proceso de evaluación, desde el establecimiento del contrato de evaluación hasta la utilización de los resultados, pasando por la metodología de la evaluación en sí misma. Será aplicable a los procedimientos y métodos utilizados a nivel individual (selección, consejo, formación…), grupal (clima y cohesión de equipos de trabajo) y organizacional (clima laboral, cultura de empresa, satisfacción…). En la norma se describen las competencias, obligaciones y responsabilidades de los clientes y de los proveedores del servicio de evaluación, antes, durante y después del proceso evaluativo. También proporciona directrices para todas las partes implicadas en el proceso evaluador, incluida la propia persona evaluada y quienes reciban los resultados de la evaluación. En suma, una vez que se publique y empiece el proceso de certificación esta nueva norma puede suponer un importante paso para la buena práctica de la evaluación de personas en contextos laborales y organizacionales.

Evaluación de los tests

Dentro de la estrategia de difundir información sobre los tests y su práctica, los psicólogos profesionales siempre que han tenido la oportunidad han reclamado la necesidad de disponer de más información técnica sobre los tests (Muñiz y Fernández-Hermida, 2000; Muñiz et al., 2001). Esto ha motivado que a instancias de la Comisión de Tests de la EFPA se desarrollase un modelo europeo de evaluación de tests, inspirado en modelos previos ya existentes como el británico (Bartram, 1996, 1998), el holandés (Evers, 2001a, b) y el español (Prieto y Muñiz, 2000). El modelo europeo puede consultarse en la página web de la EFPA (www.efpa.eu). La idea central del modelo es evaluar de forma sistemática y cuantitativa las propiedades psicométricas de los tests y ofrecer esta información a los posibles usuarios de los tests, para que dispongan de información objetiva y actualizada llevada a cabo por expertos. En el trabajo de Prieto y Muñiz (2000) se describe el modelo español, que consta de tres grandes apartados. En el primero se lleva a cabo una descripción técnica del test, y está compuesto por 31 ítems relativos al nombre de la prueba, autor, constructo medido, ámbito de aplicación, etc. En el segundo apartado se incluye la evaluación técnica de las características del instrumento. Los expertos han de juzgar características como la fundamentación teórica, la adaptación/traducción (si el test ha sido construido en otro país), la fiabilidad, la validez, los baremos, etc. Para lograr este objetivo, se han incluido 32 ítems cerrados y 6 abiertos. En la mayor parte de los ítems cerrados se propone un sistema de cinco categorías ordenadas en función de la calidad de la característica evaluada. En los ítems abiertos se solicita una justificación razonada de

las respuestas a los ítems cerrados y una evaluación de cada característica. En último apartado, se solicita una valoración global del test y un resumen de los dos primeros apartados, al objeto de resumir toda la información en una ficha técnica (Prieto y Muñiz, 2000). Puede consultarse el modelo en el citado trabajo o e la página web del COP, en el apartado de la Comisión de Tests: www.cop.es.

Lo realmente nuevo en relación con este modelo de evaluación de tests es que en la última reunión de la Comisión de Tests del COP se tomó por unanimidad la decisión de empezar a aplicar el modelo a los tests editados en España. Se empezará por los más utilizados, con la idea de evaluar cada año unos veinte tests como mínimo, lo ideal sería que en un período no muy largo estuviesen evaluados la mayoría de los tests editados en España, tal como ya ocurre en Holanda, por ejemplo. Para su evaluación cada test será enviado de entrada a dos expertos tanto en cuestiones psicométricas como en la temática del test. Si las opiniones de los expertos resultan convergentes se hará un informe final a partir de ambos. Si hubiese divergencias entre los expertos se enviará a un tercero antes de realizar el informe final. A estas evaluaciones se les dará máxima difusión, siendo publicadas en revistas que llegan a todos los colegiados, así como en la página web del COP.

En este contexto de mejora de la práctica de los tests es donde cobra sentido pleno la encuesta de la EFPA a los psicólogos profesionales sobre distintos aspectos del uso de los tests. Conociendo sus opiniones se podrán plantear medidas encaminadas a mejorar aquellos puntos débiles detectados por los profesionales. A continuación se ofrecen los resultados obtenidos en España. En esta edición de 2009 han participado un total de unos diecisiete países, mientras que en la edición de 1999 sólo habíamos participado seis, así que desde ese punto de vista de la participación no hay duda de que se ha producido un gran avance.

PARTICIPANTES EN LA ENCUESTA

La muestra está compuesta por 3.126 psicólogos profesionales que respondieron a la encuesta enviada a los 51.545 miembros del Colegio Oficial de Psicólogos. Los datos descriptivos más relevantes aparecen en las tablas 1 y 2. Cuando se comparan algunos de los datos de la muestra con los correspondientes a la población de psicólogos encuestados (tabla 1), se observan valores muy similares, por lo que no parece que haya grandes sesgos en relación con la muestra utilizada, que representa un 6% de la población. Señalar que en la profesión de psicólogo predominan las mujeres, 78% de mujeres frente al 22% de hombres, si bien en la muestra estos porcentajes varían ligeramente, bajando las mujeres al 71,5% y subiendo los hombres al 28,5%, lo cual reflejaría una mayor predisposición a contestar de los hombres. Por especialidades, en Clínica hay un 29% de hombres, en Educativa un 22% y en Trabajo un 45%, parece claro que la especialidad de Trabajo atrae muchos más hombres que las otras dos. En cuanto a las especialidades, las tres clásicas siguen siendo dominantes (tabla 2), ocupando la Clínica un lugar destacado con el (69,6%), seguida de Educativa (13,6%) y Trabajo (6,4%), el resto (deporte, jurídica, tráfico, servicios sociales…) forman un 10,4%. El 32,6% trabajan en el sector público y el 63,6% en el privado, con un 3,8% de desempleo. En la actualidad los profesionales de la psicología en España son un colectivo relativamente joven, con el 14% entre 20 y 29 años, 28,9% entre 30 y 39, 30,8% entre 40 y 49, 21,7% entre 50 y 59, 3,9% entre 60 y 68, y un 0,7% con 70 ó más años.

TABLA 1
DESCRIPCIÓN DE LA MUESTRA Y DE LA POBLACIÓN ENCUESTADA

	Muestra	Colegio Oficial de Psicólogos
Participantes	3.126	51.545
Mujeres	71,5%	78,1%
Hombres	28,5%	21,9%
Media de Edades (DT)	41,92 (10,43)	40,58(10,13)
Años de Práctica Profesional (DT)	12,50(8,90)	10,33(8,60)

TABLA 2
DESCRIPCIÓN DE LA MUESTRA ATENDIENDO A SU DISTRIBUCIÓN POR CAMPO PROFESIONAL Y SECTOR

Campo profesional	**%**
Clínica	69,6
Educativa	13,6
Trabajo	6,4
Otras	10,4
Sector	
Público	32,6
Privado	63,6
Desempleado	3,8

CUESTIONARIO

Para recoger las opiniones de los psicólogos sobre los tests y su práctica se utilizó una encuesta de 33 ítems (ver anexo) desarrollada originalmente en inglés por la Comisión de Tests de la Federación Europea de Asociaciones de Psicólogos (EFPA). Los primeros 32 ítems son de tipo Likert con cinco categorías, puntuadas de 1 a 5, mientras que el último ítem era abierto para que los profesionales indicasen los tests que más utilizan en su práctica diaria. Para su elaboración se partió de la escala original utilizada en 1999, suprimiendo algunos ítems y añadiendo otros relativos al uso de los tests informatizados e Internet. Se mantuvieron 20 de los ítems utilizados en la encuesta de 1999, lo que permitirá comparar los resultados de entonces con los obtenidos ahora. Se tradujo al español y se volvió a traducir al inglés compro-

TABLA 3
ANÁLISIS DE COMPONENTES PRINCIPALES

	Componentes							
Ítems	I	II	III	IV	V	VI	VII	VIII
Item 25-5	0,773							
Ítem 25-2	0,770							
Ítem 25-8	0,764							
Ítem 25-7	0,750							
Ítem 25-4	0,747							
Ítem 25-3	0,740							
Ítem 25-6	0,704							
Ítem 25-1	0,463							
Ítem 23		0,868						
Ítem 22		0,862						
Ítem 21		0,728						
Ítem 24		0,539						
Ítem 3			0,700					
item12			0,689					
Ítem 8			0,624					
Ítem 11			0,616					
Ítem 19			0,599					
Ítem 17				0,705				
Ítem 20				0,680				
Ítem 13				0,625				
Ítem 10				0,564				
Ítem 1					0,833			
Ítem 6					0,736			
Ítem 2					0,618			
Ítem 7						0,680		
Ítem 5						0,610		
Ítem 15						0,593		
Ítem 9						0,402		
Ítem 14							0,739	
Ítem 18							0,615	
Item 16							0,481	
Ítem 4								0,503
% de la Varianza	13,27	8,06	7,88	7,10	5,54	4,80	4,49	3,68
% Acumulado	13,27	21,33	29,21	36,32	41,85	46,65	51,15	54,82

Nota. Se rotaron ortogonalmente los ocho componentes con valores propios mayores que la unidad. Se eliminaron los pesos inferiores a 0,45 para facilitar la lectura de la tabla, excepto cuando la variable no alcanzaba dicho peso.

bando que ambas versiones, la original y la generada a partir de la versión española eran esencialmente equivalentes, tal como recomiendan las directrices de la ITC (Hambleton et al., 2005). Con la versión española se llevaron a cabo varios estudios piloto cualitativos y cuantitativos para asegurarse que los ítems de la encuesta no ofrecían dudas de comprensión e interpretación por parte de la población a la que iban dirigidos (Wilson, 2005).

RECOGIDA Y ANÁLISIS DE LOS DATOS

Para la recogida de los datos se envió el cuestionario a todos psicólogos miembros del Colegio Oficial de Psicólogos, precedido de una carta de presentación de Francisco Santolaya, Presidente del COP, en la que se explicaban los motivos de la encuesta. También se adjuntaba un sobre con el franqueo pagado, con la instrucción de que una vez rellenada la encuesta la depositasen en cualquier buzón de correos.

Se llevaron a cabo análisis estadísticos descriptivos de los ítems y de los datos generales solicitados a los participantes. Para determinar la estructura dimensional de los ítems de la escala se realizó un análisis de componentes principales con rotación ortogonal (varimax) de los componentes con valores propios superiores a la unidad, siguiendo el criterio de Kaiser. Si bien desde un punto de vista técnico el método de Máxima Verosimilitud podría ser más recomendable (Ferrando y Anguiano, 2010), se mantiene aquí esta estrategia para permitir una mejor comparación con los resultados previos (Muñiz et al. 2001). La fiabilidad de la prueba se estimó mediante el coeficiente alfa de Cronbach (1951) y las comparaciones de las medias de los ítems se hicieron mediante análisis de varianza. Todos los análisis se llevaron a cabo con el SPSS-15.

DIMENSIONES EVALUADAS POR EL CUESTIONARIO

El coeficiente alfa de la escala fue de 0,665, lo cual indica que la consistencia interna de la escala no es muy elevada. Esto era esperable, puesto que en ningún momento se trataba de obtener una escala con una alta consistencia interna, sino que se pretende evaluar distintos aspectos implicados en la práctica de los tests.

Como se puede observar en la Tabla 3, los ítems de la escala se articulan en torno a ocho dimensiones, obtenidas mediante un análisis de componentes principales, que explican el 54,82% de la varianza total. En el primer componente se agrupan todos los ítems relacionados con los problemas de uso de los tests. El segundo componente lo forman los ítems relativos a las actitudes de los psicólogos hacia los tests. El tercer componente aparece muy claro, refiriéndose a la necesidad de regulación del uso de los tests, bien sea legalmente o por parte de las organizaciones colegiales nacionales o europeas. Estos tres primeros componentes coinciden plenamente con los obtenidos en la escala aplicada hace diez años (Muñiz y Fernández-Hermida, 2000). El cuarto componente lo forman cuatro ítems relativos al uso de Internet y de informes computerizados, esta dimensión es nueva dado que estos ítems no se habían incluido en la versión del cuestionario de 1999. El quinto componente se refiere a la formación y conocimientos sobre los tests. El sexto acoge ítems relacionados con internet y los tests informatizados, pesando también un ítem relativo al uso de los tests por no psicólogos. El séptimo, con tres ítems, se centra en la permisividad en el uso de los tests, y el octavo recoge el ítem relativo a la información técnica sobre los tests de la que disponen los profesionales. La estructura es muy clara, y refleja bien las dimensiones fundamentales a tener en cuenta a la hora evaluar el uso de los tests. Es muy similar a la estructura encontrada en la aplicación de 1999, añadiéndose ahora dos nuevas dimensiones relativas a Internet y a los tests informatizados.

OPINIONES SOBRE EL USO DE LOS TESTS

En la Tabla 4 aparecen las medias y las desviaciones típicas de las respuestas de los participantes a los ítems del cuestionario. Se ofrecen los datos de la muestra total y dividida por las especialidades profesionales de Clínica, Trabajo y Educativa. El asterisco detrás del texto del ítem indica que las diferencias entre las medias de las tres especialidades resultó estadísticamente significativa al nivel de confianza del 95%.

Los resultados en detalle se encuentran en la tabla 4, se comentan a continuación algunos de los datos más sobresalientes para cada una de las dimensiones del cuestionario (tabla 3). En la primera dimensión, relacionada con los problemas de uso de los tests, se observa que si bien la situación no es grave, obteniéndose una valoración media de 3,12, hay distintos aspectos claramente mejorables. Siguen haciéndose fotocopias de los tests (3,51), y en opinión de los psicólogos no siempre se está

TABLA 4
MEDIA Y DESVIACIÓN TÍPICA DE CADA UNO DE LOS ÍTEMS DE LA ENCUESTA POR ESPECIALIDADES (CLÍNICA, EDUCATIVA Y TRABAJO) Y GLOBAL.

Ítems	Clínica		Educativa		Trabajo		Global	
	Media	DT	Media	DT	Media	DT	Media	DT
1.- La formación recibida en la carrera de Psicología es suficiente para la utilización correcta de la mayoría de los tests	2,41	1,18	2,44	1,14	2,61	1,23	2,43	1,18
2.- La formación recibida en cursos y Másteres es suficiente para el uso correcto de la mayoría de los tests*	3,12	1,08	3,07	1,01	2,90	1,00	3,09	1,07
3.- La Federación Europea de Asociaciones de Psicólogos (EFPA) debería de establecer un sistema para acreditar la competencia de los usuarios de tests*	3,34	1,37	3,40	1,34	3,89	1,23	3,39	1,36
4.- Los profesionales disponen de suficiente información (revisiones independientes, investigaciones, documentación...) sobre la calidad de los tests editados en nuestro país	2,74	1,14	2,72	1,09	2,72	1,09	2,73	1,12
5.- En mi campo profesional los tests computerizados están reemplazando progresivamente a los tests de papel y lápiz*	2,89	1,35	2,96	1,24	3,53	1,26	2,94	1,35
6.- Mis conocimientos actuales en relación con los tests los recibí fundamentalmente durante la carrera de Psicología	2,57	1,36	2,48	1,28	2,74	1,37	2,59	1,36
7.- La aplicación de los tests por Internet tiene muchas ventajas en comparación con la aplicación clásica de papel y lápiz*	2,75	1,21	2,64	1,14	3,11	1,21	2,78	1,20
8.- El uso de los tests psicológicos debería de restringirse a psicólogos cualificados*	4,12	1,19	4,15	1,11	4,39	,98	4,12	1,17
9.- Aunque los no psicólogos podrían aplicar y puntuar los tests, la interpretación e información sobre los resultados deberían estar restringidos a los psicólogos	4,39	1,17	4,41	1,06	4,52	1,05	4,39	1,16
10.- Los informes generados automáticamente por ordenador no tienen ninguna validez	2,96	1,14	2,94	1,14	2,87	1,12	2,94	1,14
11.- Los estándares y directrices que definen las cualidades técnicas mínimas de un test deberían de ser obligatorios [por ejemplo los estándares de la Federación Europea de Asociaciones de Psicólogos (EFPA), o los de la Asociación de Psicología Americana (APA)].*	4,07	,98	4,12	,91	4,27	,87	4,10	,96
12.- Se necesita legislación para controlar los abusos más serios con los tests*	3,99	1,04	4,03	,99	4,26	,96	4,01	1,04
13.- La aplicación de los tests por Internet pone en desventaja a algunas personas evaluadas	3,53	1,10	3,51	1,13	3,46	1,11	3,54	1,10
14.- Todo aquel que sea capaz de demostrar su competencia en el uso de los tests (sea psicólogo o no) debería de ser autorizado para utilizarlos	2,11	1,34	2,07	1,29	2,22	1,35	2,10	1,32
15.- Si se utiliza adecuadamente, Internet puede mejorar mucho la calidad de la aplicación de los tests*	3,04	1,12	3,00	1,08	3,34	1,12	3,08	1,11
16.- Los controles sobre los tests deberían de ser mínimos, pues inhiben el desarrollo de nuevas ideas y nuevos procedimientos de evaluación	1,94	1,06	1,93	1,01	1,89	,99	1,93	1,04
17.- La aplicación de tests por Internet no permite proteger la privacidad de los usuarios*	2,95	1,25	3,09	1,18	2,73	1,25	2,95	1,24
18.- Habría que permitir a los editores que vendiesen cualquier test que ellos consideren adecuado	1,80	1,11	1,67	,95	1,71	1,01	1,77	1,09
19.- El Colegio Oficial de Psicólogos debería de ejercer un papel más activo para regular y mejorar el uso que se hace de los tests*	4,09	1,07	4,22	,89	4,23	1,03	4,13	1,03
20.- La aplicación de tests por Internet abre posibilidades de fraude	3,80	1,09	3,78	1,11	3,63	1,15	3,78	1,10
21.- En el desempeño de mi profesión utilizo tests habitualmente*	3,77	1,29	3,98	1,28	3,78	1,16	3,76	1,30
22.- Los tests constituyen una excelente fuente de información si se combinan con otros datos psicológicos*	4,44	,89	4,59	,79	4,49	,81	4,46	,87
23.- Utilizados correctamente, los tests son de gran ayuda para el psicólogo*	4,38	,89	4,55	,76	4,53	,74	4,41	,88
24.- Teniendo en cuenta todos los aspectos, creo que en la última década el uso de los tests ha mejorado en mi país*	3,58	,97	3,69	,87	3,42	,97	3,58	,96
25.- Estime la frecuencia de los siguientes problemas de uso de los tests en su entorno profesional (1: muy poco frecuente; 5: muy habitual)								
(1) Hacer fotocopias de materiales sujetos a copyright	3,53	1,36	3,50	1,35	3,48	1,38	3,51	1,38
(2) Hacer evaluaciones utilizando tests inadecuados*	2,62	1,31	2,47	1,23	3,13	1,25	2,64	1,31
(3) No estar al día*	3,23	1,25	3,08	1,24	3,51	1,16	3,25	1,23
(4) No contrastar las interpretaciones con otros*	3,32	1,26	3,14	1,26	3,58	1,20	3,33	1,25
(5) No tener en cuenta los errores de medida de las puntuaciones *	3,10	1,22	2,97	1,20	3,30	1,19	3,10	1,22
(6) No restringir la aplicación de los tests a personal cualificado*	2,93	1,49	2,76	1,44	3,39	1,45	2,92	1,49
(7) No tener en cuenta las condiciones locales (país, región) que pueden afectar a la validez *	3,19	1,31	3,15	1,30	3,47	1,25	3,21	1,31
(8) Hacer interpretaciones que van más allá de los límites del test*	2,96	1,37	2,86	1,32	3,24	1,36	2,97	1,36

Nota. El asterisco indica que existen diferencias estadísticamente significativas entre las medias del ítem en función de la especialidad, p<0,05.

al día (3,25), ni se contrastan las interpretaciones con otros profesionales (3,33). Se constata, por otra parte, algo que se sospechaba y es que los problemas de uso son más acentuados en el ámbito de la Psicología del Trabajo (3,39), que en Clínica (3,07) y en Educativa (2,99), o al menos así los perciben los profesionales. Este dato diferencial que ahora se confirma en España es lo que ha movido a la Comisión de Tests de la EFPA a iniciar un proyecto piloto para explorar la posibilidad de la acreditación de usuarios de tests en el ámbito del Trabajo y las Organizaciones. En este proyecto toma parte activa el COP, que ha nombrado a la Profesora Ana Hernández de la Universidad de Valencia como representante en la citada comisión europea. La idea es el establecimiento de una acreditación europea en el ámbito de los tests (Eurotest) similar al Europsy (Bartram y Roe, 2005; Lunt, 2005; Peiró, 2003). Otra iniciativa encaminada a la mejora del uso de las pruebas en este campo es la norma ISO 10667 ya comentada, que pretende regular los procesos de evaluación de personas en contextos laborales.

El segundo factor se refiere a las actitudes de los psicólogos hacia los tests, confirmándose los datos obtenidos en la encuesta de 1999 (Muñiz y Fernández Hermida, 2000), en el sentido de la favorable actitud hacia los tests cuando estos se utilizan combinados con otros datos psicológicos (4,46). Resulta de gran interés y alentador constatar que los profesionales consideran que en la última década el uso de los tests en España ha mejorado (3,58), pues aunque queden muchas cosas por hacer, parece que se camina en la dirección adecuada. En función de la especialidad son los del campo Educativo quienes más utilizan los tests en su práctica diaria (3,98), y de nuevo es en Trabajo donde la percepción de mejora del uso de los tests en la última década es más baja (3,42), si bien está por encima de la media de la escala.

La tercera dimensión se refiere a la necesidad de regulación del uso de los tests. Aquí se observa una opinión muy favorable de los profesionales a tomar medidas tanto legales como por parte de las organizaciones colegiales para mejorar el uso de los tests. No se ve con malos ojos que la EFPA estableciese un sistema de acreditación de la competencia de los usuarios de los tests, siendo los más favorables los pertenecientes al campo de Trabajo (3,89). En los cinco ítems que componen este factor son los profesionales del campo de Trabajo los que se muestran más contundentes a la hora de reclamar acciones para regular el uso de los tests, siendo favorables a la intervención de las organizaciones colegiales a nivel nacional e internacional.

El cuarto factor se centra en el uso de Internet y de los informes computerizados. Parece claro que los psicólogos se muestran bastante escépticos sobre la utilización de informes automatizados por ordenador, así como acerca de la irrupción de Internet en el campo de la evaluación. No debe de interpretarse esto como una actitud defensiva hacia las nuevas tecnologías, sino más bien de precaución hacia cuestiones como los problemas de privacidad y fraude que puede conllevar el uso de Internet, o las desventajas de algunos usuarios no familiarizados con la red. En cuanto a los informes generados automáticamente por ordenador, no les niegan su validez de plano (2,94), pero tampoco los canonizan, y es que estos informes pueden constituir una excelente ayuda para el psicólogo, pero de ninguna manera lo sustituyen, son lo que son, herramientas que el profesional debe usar con enjundia. De los cuatro ítems que pesan en este factor, sólo en uno de ellos hay diferencias estadísticamente significativas entre las especialidades, el relativo al mantenimiento de la privacidad por Internet, siendo los de Trabajo los que consideran que Internet permite mantener la privacidad en un buen grado. Se nota que este sector está más habituado que los Clínicos y Educativos a trabajar en contextos de tele-evaluación, pues en la actualidad los sistemas utilizados permiten mantener un elevado grado de privacidad cuando se utiliza la red con estos fines.

El quinto factor lo componen tres ítems relacionados con la formación y los conocimientos de los psicólogos sobre los tests. Queda patente la necesidad de formación manifestada por los profesionales, pues ni la propia carrera de psicología (2,43), ni incluso los posteriores Másters (3,09), colman las necesidades de formación. Y en esto las tres especialidades convergen en sus opiniones. Esta situación es hasta cierto punto lógica, pues si en general los conocimientos técnicos vienen a tener una vigencia de cinco años, los tests no son una excepción, demandando una formación permanente y actualizada. Surgen nuevos tests, nuevas técnicas, nuevos modelos, y lo aprendido en la carrera y en algunos Másters constituye una base imprescindible, pero debe de ser complementado y actualizado con una formación específica permanente. He ahí un reto importante para las asocia-

ciones profesionales, la universidad, amén de otras instituciones relacionadas con la profesión.

La sexta dimensión obtenida vuelve a estar relacionada, como la cuarta, con aspectos relativos a Internet y a los tests informatizados. No parece que por el momento los tests informatizados estén reemplazando a los tests de papel y lápiz, si bien se observa que es en el campo de la Psicología del Trabajo donde se produce un mayor avance. Se observa que la utilización de Internet es todavía baja entre los profesionales, mostrándose más proclives a su uso quienes se dedican al área de Trabajo. La opinión de los psicólogos sobre el uso de los tests por no psicólogos es tajante, si bien se admite la aplicación y puntuación por parte de los no titulados en psicología, la interpretación ha de ser privativa de los psicólogos (4,39). Y es que una cosa es aplicar una prueba y puntuarla y otra muy diferente es conocer con precisión las inferencias que se pueden hacer sobre la conducta humana a partir de esas puntuaciones, para lo cual todo el saber psicológico es necesario.

El séptimo es un factor relativo a la permisividad del uso de los tests. Los profesionales se manifiestan de forma clara y contundente sobre la necesidad de control para usar y editar tests, las tres especialidades son unánimes al respecto.

Finalmente un solo ítem conforma el octavo factor, relativo a la información técnica sobre los tests de la que disponen los profesionales. Hay acuerdo unánime en las tres especialidades mayoritarias en que se necesita disponer de más información de este tipo. Este dato confirma los obtenidos en la encuesta de 1999, lo que ha motivado que el COP por medio de la Comisión de Tests haya puesto en marcha un proyecto para ir evaluando los tests editados en España, poniendo a disposición de los profesionales esas evaluaciones. Se estima que las primeras evaluaciones aparecerán publicadas en 2010.

TABLA 5
LOS 25 TESTS MÁS UTILIZADOS POR LOS PSICÓLOGOS ESPAÑOLES

Nombre de la prueba		N	%
WISC*	(Weschsler Intelligence Scale for Children)	649	22,70
16PF	(16 Personality Factors)	609	22,37
MCMI	(Millon Clinical Multiaxial Inventory)	489	17,96
MMPI	(Minnesota Multiphasic Personality Inventory)	480	17,63
BDI	(Beck Depression Inventory)	372	13,66
WAIS*	(Weschsler Adult Intelligence Scale)	370	12,93
STAI	(State Trait Anxiety Inventory)	316	11,60
RORSCHACH	(Rorschach)	154	5,66
SCL-90	(Symptom Checklist 90)	143	5,25
RAVEN	(Raven Progressive Matrices)	137	5,03
TAMAI	(Test Autoevaluativo Multifactorial de Adaptación Infantil)	120	4,41
MMSE	(Mini Mental State Examination)	113	4,15
MSCA	(McCarthy Scales of Children's Abilities)	95	3,49
BADYG	(Batería de Aptitudes Diferenciales y Generales)	93	3,42
TALE	(Test de Análisis de Lecto-Escritura)	92	3,38
HTP	(House-Tree-Person Test)	88	3,23
EPQ	(Eysenck Personality Questionnaire)	84	3,08
BENDER	(Bender Visual Motor Gestalt Test)	80	2,94
ISRA	(Inventario de Situaciones y Respuesta de Ansiedad)	72	2,64
PROLEC	(Batería de Evaluación de los Procesos Lectores)	68	2,50
MACI	(Millon Adolescent Clinical Inventory)	59	2,17
BASC	(Behavior Assessment System for Children)	57	2,09
CUIDA	(Eval. de Adoptantes, Cuidadores, Tutores y Mediadores)	51	1,87
ITPA	(Illinois Test of Psycholinguistic Abilities)	51	1,87
CAQ	(Clinical Analysis Questionnaire)	48	1,80

*Bajo las siglas WISC y WAIS se incluyen las distintas versiones disponibles de ambas pruebas, como el WISC-R o el WISC-IV.

TESTS MÁS USADOS EN ESPAÑA

En la encuesta se les pedía a los participantes que indicasen los tres tests que más utilizaban en su práctica diaria. Véase en la tabla 5 los resultados obtenidos. Como se puede observar, figura en primer lugar la escala de inteligencia para niños WISC, seguida por el test de personalidad 16PF. Todos los tests más utilizados son tests clásicos psicométricos bien establecidos en Psicología, apareciendo el test proyectivo Rorschach en octavo lugar. Entre los veinticinco tests más utilizados aparecen seis de autores españoles (24%), lo cual indica el empuje cada vez mayor de la producción nacional. En la tabla 6 aparecen los diez tests más utilizados por especialidades, y como no podía ser de otro modo, las diferencias son notables, reflejando las distintas tareas de cada campo. Cabe subrayar el amplio uso de la prueba BDI (Test de Depresión de Beck), el quinto más utilizado, tratándose de una prueba no comercializada en España, lo que significa que se están utilizando fotocopias de la prueba y baremos tomados de publicaciones y estudios hechos sobre la prueba. Sería altamente aconsejable que esta prueba tan utilizada por los profesionales dispusiese de un proceso de

validación más sistemático y riguroso en nuestro país, lo cual suponemos que no se ha producido todavía por cuestiones relacionados con aspectos comerciales y de propiedad intelectual de la prueba.

En la tabla 7 aparecen las medias de los ítems comunes que se utilizaron en la encuesta del año 2000 y 2010. Como se puede observar son muy similares, no observándose grandes diferencias en estos diez años, la correlación entre ambas es de 0,986. Tal vez señalar una ligera evolución en el sentido deseado, en el ítem 4, manifestando los encuestados que en los últimos diez años ha mejorado la información de que disponen sobre la calidad de los tests, una media de 2,38 antes frente a 2,73 ahora. No es ningún consuelo, la media sigue siendo baja, pero al menos se avanza en la dirección adecuada, claro que a ese ritmo harían falta unos 50 años para llegar a una situación razonable. Es evidente que hay que hacer más cosas y más rápido.

PERSPECTIVAS DE FUTURO

Está claro que los tests, llegados a la psicología hace más de cien años, lo han hecho para quedarse, ha llovido mucho desde aquellos primeros tests de carácter sensomotor ideados por Galton a finales del siglo XIX, o desde que Binet y Simon (1905) propusieron la primera escala individual de inteligencia. Nadie ha sido capaz de predecir entonces por donde discurrirían los tests del futuro, y no pretendemos ahora hacerlo nosotros, lo que siguen son algunas reflexiones sobre la situación actual de los tests, y que previsiblemente condicionará su futuro. Puede sonar a tópico, pero la gran fuerza que está remodelando la evaluación psicológica en la actualidad son las nuevas tecnologías de la información, en especial los avances informáticos, multimedia e Internet. Autores como Bennet (1999, 2006), Breithaupt, Mills y Medican (2006) o Drasgow, Luecht y Bennet (2006) consideran que las nuevas tecnologías están influyendo sobre todos los aspectos de la evaluación psicológica, tales como el diseño de los tests, la construcción de los ítems, la presentación de los ítems, la puntuación de los tests y la evaluación a distancia. Todo ello está haciendo cambiar el formato y contenido de las evaluaciones, surgiendo la duda razonable de si los tests de papel y lápiz tal como los conocemos ahora serán capaces de resistir este nuevo cambio tecnológico. En este sentido lo dicho sobre el futuro de los libros y la prensa escrita bien se puede aplicar a los tests. Nuevas formas de evaluación emergen, como la evaluación *auténtica* en el ámbito educativo (portafolios, composiciones escritas, trabajos), aunque los tests psi-

TABLA 6
LOS DIEZ TESTS MÁS UTILIZADOS POR LOS PSICÓLOGOS ESPAÑOLES POR ESPECIALIDADES

	Clínica	Educativa	Trabajo
1	MCMI	WISC	16PF
2	16PF	BADYG	PAPI
3	MMPI	TALE	DAT
4	BDI	MSCA	TPT
5	WISC	16PF	IPV
6	WAIS	RAVEN	MMPI
7	STAI	PROLEC	IGF
8	RORSCHACH	BENDER	BFQ
9	SCL-90	ITPA	MCMI
10	MMSE	TAMAI	NEO PI

Nota. Se identifican a continuación las siglas de los tests de esta tabla que no aparecen descritos en la tabla 5: PAPI (The Personality and Preference Inventory), DAT (Differential Aptitude Test), TPT (Test de Personalidad de TEA), IPV (Inventario de Personalidad para Vendedores), IGF (Inteligencia General de TEA), BFQ (Big Five Questionnaire), NEO PI (NEO Personality Inventory).

TABLA 7
MEDIAS DE LOS ÍTEMS OBTENIDAS EN EL AÑO 2000 Y EN EL AÑO 2010

Ítems	Resultados 2000 (Media)	Resultados 2010 (Media)
1	2,41	2,43
4	2,38	2,73
6	2,57	2,59
8	4,23	4,12
9	4,34	4,39
11	4,33	4,10
12	4,29	4,01
14	2,42	2,10
16	1,85	1,93
18	1,57	1,77
19	4,15	4,13
21	3,56	3,76
22	4,41	4,46
23	4,37	4,41
25-1	3,60	3,51
25-2	2,63	2,64
25-5	3,07	3,10
25-6	2,91	2,92
25-7	3,28	3,21
25-8	2,99	2,97

cométricos seguirán siendo herramientas fundamentales, dada su objetividad y economía de medios y tiempo (Phelps, 2005, 2008). Según la apreciación de un especialista como el profesor de la Universidad de Massachusetts Ronald K. Hambleton (Hambleton, 2004, 2006), seis grandes áreas atraerán la atención de investigadores y profesionales en los próximos años. La *primera* es el uso internacional de los tests, debido a la globalización creciente y a las facilidades de comunicación, lo cual plantea todo un conjunto de problemas de adaptación de los tests de unos países a otros (Byrne et al., 2009; Hambleton et al., 2005). La *segunda* es el uso de nuevos modelos psicométricos y tecnologías para generar y analizar los tests. Cabe mencionar aquí toda la nueva psicometría derivada de los modelos de Teoría de Respuesta a los Ítems (TRI), los cuales vienen a solucionar algunos problemas que no encontraban buena solución dentro del marco clásico, pero como siempre ocurre a la vez que se solucionan unos problemas surgen otros nuevos que no estaban previstos. La *tercera* es la aparición de nuevos formatos de ítems derivados de los grandes avances informáticos y multimedia. De las modestas matrices en blanco y negro pasamos hoy a pantallas interactivas, con animación y sonido, capaces de reaccionar a las respuestas de las personas evaluadas (Irvine y Kyllonen, 2002; Shermis y Burstein, 2003; Sireci y Zenisky, 2006; Zenisky y Sireci, 2002). Ahora bien, no se trata de innovar por innovar, antes de sustituir los viejos por los nuevos formatos hay que demostrar empíricamente que mejoran lo anterior, las propiedades psicométricas como la fiabilidad y la validez no son negociables. La *cuarta* área que reclamará gran atención es todo lo relacionado con los tests informatizados y sus relaciones con Internet. Mención especial merecen en este campo los Tests Adaptativos Informatizados (TAI) que permiten ajustar la prueba a las características de la persona evaluada, sin por ello perder objetividad o comparabilidad entre las personas, lo cual abre perspectivas muy prometedoras en la evaluación (Olea, Ponsoda y Prieto, 1999). La evaluación a distancia o tele-evaluación es otra línea que se abre camino con rapidez, lo cual plantea serios problemas de seguridad de los datos y de las personas, pues hay que comprobar que la persona que se está evaluando es la que realmente dice ser, sobre todo en contextos de selección de personal o de pruebas con importantes repercusiones para la vida futura de la persona evaluada. En este campo se están dando grandes avances básicos y aplicados (Bartram y Hambleton, 2006; Leeson, 2006; Mills et al., 2002; Parshall et al., 2002). En *quinto* lugar cabe señalar un campo que puede parecer periférico pero que está cobrando gran importancia, se trata de los sistemas a utilizar para dar los resultados a los usuarios y partes legítimamente implicadas. Es fundamental que estos comprendan sin equívocos los resultados de las evaluaciones, y no es obvio cuál es la mejor manera de hacerlo, sobre todo si se tienen que enviar para su interpretación y explicación del profesional, como ocurre en numerosas situaciones de selección de personal, o en la evaluación educativa (Goodman y Hambleton, 2004). Obviamente esto tiene menor influencia en contextos clínicos. Finalmente es muy probable que en futuro haya una gran demanda de formación por parte de distintos profesionales relacionados con la evaluación, no necesariamente psicólogos, aunque también, tales como profesores, médicos, enfermeros, etc. No se trata de que estos profesionales puedan utilizar e interpretar los tests propiamente psicológicos, sino que demanden información para poder comprender y participar en los procesos evaluativos y de certificación que se desarrollan en su ámbito laboral. Estas son algunas líneas de futuro sobre las que muy probablemente girarán las actividades evaluadoras en un futuro no muy lejano, no se trata de hacer una relación exhaustiva ni mucho menos, sino indicar algunas pistas para orientarse en el mundo cambiante de la evaluación psicológica.

AGRADECIMIENTOS

Los autores desean expresar su más sincero agradecimiento a los miembros de la Comisión de Tests del COP por su ayuda y colaboración en la realización del trabajo: Rocío Fernández Ballesteros, Miguel Martínez, Eduardo Montes, Jaime Pereña y Javier Rubio. Muchas gracias también al Colegio Oficial de Psicólogos, sin cuya ayuda el trabajo no hubiese sido posible. La ayuda de Ángela Campillo, Eduardo Fonseca y Elsa Peña ha sido fundamental en el procesamiento de los datos, muchas gracias. Finalmente agradecer sinceramente la colaboración de todos los profesionales que respondieron a la encuesta, nada se habría hecho sin ellos. La financiación para la realización del trabajo ha sido aportada por el COP y por el Ministerio de Ciencia e Innovación (Ref. nº PSI2008-03934).

ANEXO

ENCUESTA UTILIZADA PARA RECOGER LAS OPINIONES DE LOS PSICÓLOGOS SOBRE LA PRÁCTICA DE LOS TESTS

DATOS GENERALES

Edad: Sexo: ❑ Hombre ❑ Mujer

Año en el que obtuvo la Licenciatura en Psicología:

Años que lleva colegiado:

Especialidad Profesional: ❑ Clínica-Salud ❑ Educativa ❑ Trabajo ❑ Otras (Indicar)

Trabaja actualmente como Psicólogo ❑ sí ❑ no

Trabaja en el sector: ❑ Público ❑ Privado ❑ En paro

Número de años que lleva en el trabajo actual

INSTRUCCIONES

Las cuestiones que aparecen a continuación están formuladas para responder en una escala de 1 a 5. Si está *en desacuerdo total* con la frase señale el 1, si está *totalmente de acuerdo* señale el 5, Utilice los números 2, 3 y 4 para las situaciones intermedias. La encuesta es totalmente anónima.

CUESTIONARIO

1. La formación recibida en la carrera de Psicología es suficiente para la utilización correcta de la mayoría de los tests
2. La formación recibida en cursos y Másters es suficiente para el uso correcto de la mayoría de los tests
3. La *Federación Europea de Asociaciones de Psicólogos* (EFPA) debería de establecer un sistema para acreditar la competencia de los usuarios de tests
4. Los profesionales disponen de suficiente información (revisiones independientes, investigaciones, documentación...) sobre la calidad de los tests editados en nuestro país
5. En mi campo profesional los tests computerizados están reemplazando progresivamente a los tests de papel y lápiz
6. Mis conocimientos actuales en relación con los tests los recibí fundamentalmente durante la carrera de Psicología
7. La aplicación de los tests por Internet tiene muchas ventajas en comparación con la aplicación clásica de papel y lápiz
8. El uso de los tests psicológicos debería de restringirse a psicólogos cualificados
9. Aunque los no psicólogos podrían aplicar y puntuar los tests, la interpretación e información sobre los resultados deberían estar restringidos a los psicólogos
10. Los informes generados automáticamente por ordenador no tienen ninguna validez
11. Los estándares y directrices que definen las cualidades técnicas mínimas de un test deberían de ser obligatorios [por ejemplo los estándares de *la Federación Europea de Asociaciones de Psicólogos* (EFPA), o los de la *Asociación de Psicología Americana* (APA)].
12. Se necesita legislación para controlar los abusos más serios con los tests
13. La aplicación de los tests por Internet pone en desventaja a algunas personas evaluadas
14. Todo aquel que sea capaz de demostrar su competencia en el uso de los tests (sea psicólogo o no) debería de ser autorizado para utilizarlos
15. Si se utiliza adecuadamente, Internet puede mejorar mucho la calidad de la aplicación de los tests
16. Los controles sobre los tests deberían de ser mínimos, pues inhiben el desarrollo de nuevas ideas y nuevos procedimientos de evaluación
17. La aplicación de tests por Internet no permite proteger la privacidad de los usuarios
18. Habría que permitir a los editores que vendiesen cualquier test que ellos consideren adecuado
19. El *Colegio Oficial de Psicólogos* debería de ejercer un papel más activo para regular y mejorar el uso que se hace de los tests
20. La aplicación de tests por Internet abre posibilidades de fraude
21. En el desempeño de mi profesión utilizo tests habitualmente
22. Los tests constituyen una excelente fuente de información si se combinan con otros datos psicológicos
23. Utilizados correctamente, los tests son de gran ayuda para el psicólogo
24. Teniendo en cuenta todos los aspectos, creo que en la última década el uso de los tests ha mejorado en mi país
25. Estime la frecuencia de los siguientes problemas de uso de los tests en su entorno profesional (1: muy poco frecuente; 5: muy habitual)
    1. Hacer fotocopias de materiales sujetos a *copyright*
    2. Hacer evaluaciones utilizando tests inadecuados
    3. No estar al día
    4. No contrastar las interpretaciones con otros
    5. No tener en cuenta los errores de medida de las puntuaciones
    6. No restringir la aplicación de los tests a personal cualificado
    7. No tener en cuenta las condiciones locales (país, región) que pueden afectar a la validez
    8. Hacer interpretaciones que van más allá de los límites del test

26. Cite los tres tests que utiliza con más frecuencia en el ejercicio de su profesión:
    1. ……
    2. ……
    3. ……

Observaciones: Comente cualquier otro aspecto que considere oportuno (si es necesario puede adjuntar más hojas)

REFERENCIAS

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

Bartram, D. (1996). Test qualifications and test use in the UK: The competence approach. *European Journal of Psychological Assessment, 12*, 62-71.

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

Bartram, D. y Coyne, I. (1998). Variations in national patterns of testing and test use: The ITC/EFPPA international survey. *European Journal of Psychological Assessment, 14*, 249-260.

Bartram, D. y Hambleton, R. K. (Eds.) (2006). *Computer-based testing and the Internet*. Chichester: John Wiley and Sons.

Bartram, D. y Roe, R. A. (2005). Definition and assessment of competences in the context of the European diploma in psychology. *European Psychologist, 10,* 93-102.

Bennett, R. E. (1999). Using new technology to improve assessment. *Educational Measurement: Issues and practice. 18(3)*, 5-12.

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. En D. Bartram and R. K. Hambleton (Eds.), *Computer-based testing and the Internet*. Chichester: John Wiley and Sons. (pp. 201-217).

Binet, A. y Simon, T. H. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique, 11*, 191-244.

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

Brennan, R. L. (Ed.) (2006). *Educational measurement*. Westport, CT: ACE/Praeger.

Byrne, B. M., Leong, F. T., Hambleton, R. K., Oakland, T., van de Vijver, F. J., y Cheung, F. M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3(2)*, 94-105.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Downing, S. M. y Haladyna, T. M. (Eds.) (2006). *Handbook of test development.* Hillsdale, NJ.: LEA.

Drasgow, F., Luecht, R. M. y Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement*. Westport, CT: ACE/Praeger. (págs. 471-515).

European Federation of Professional Psychologists' Associations (2005). *Meta-Code of ethics*. Brussels: Author (www.efpa.eu).

Evers, A. (2001a). Improving test quality in the Netherlands: Results of 18 years of tests ratings. *International Journal of Testing, 1*, 137-153.

Evers. A. (2001b). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155-182.

Fernández-Ballesteros, R., De Bruyn, E., Godoy, A., Hornke, L., Ter Laak, J., y Vizcarro, C. et al. (2001). Guidelines for the assessment process (GAP): A proposal for discussion. *European Journal of Psychological Assessment, 17*, 187-200.

Ferrando, P. J. y Anguiano, C. (2010). El análisis factorial como técnica de investigación en Psicología. *Papeles del Psicólogo, 31(1), 18-33.*

Goodman, D.P. y Hambleton, R.K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*, 145-220.

Hambleton, R. K. (2004). Theory, methods, and practices in testing for the 21st century. *Psicothema, 16*, 696-701.

Hambleton, R. K. (2006). *Testing practices in the 21st century*. Key Note Address, University of Oviedo, Spain, March 8th.

Hambleton, R. K., Merenda, P. F., y Spielberger, C. D. (Eds.) (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Londres: LEA.

Irvine, S. y Kyllonen, P. (Eds.) (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum.

Joint Committee on Testing Practices. (2002). *Ethical principles of psychologists and code of conduct.* Washington DC: Joint Committee on Testing Practices.

Koocher, G. y Kith-Spiegel, P. (2007). *Ethics in psychology*. Nueva York: Oxford University Press.

Leach, M. y Oakland, T. (2007). Ethics standards impacting test development and use: A review of 31 ethics codes impacting practices in 35 countries. *International Journal of Testing, 7*, 71-88.

Leeson, H. V. (2006). The mode effect: A literature review of human and technological issues in computerized testing. *International Journal of Testing, 6*, 1-24.

Lindsay, G., Koene, C., Ovreeide. H., y Lang, F. (2008). *Ethics for European psychologists*. Gottingen and Cambridge, MA: Hogrefe.

Lunt, I. (2005). The implications of the "Bologna process" for the development of a European qualification in psychology. *European Psychologist, 10,* 86-92.

Mills, C.N., Potenza, M.T., Fremer, J.J., y Ward, W.C. (Eds.) (2002). *Computer-based testing: Building the foundation for future assessments*. Hillsdale, NJ: LEA.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S., y Most, R. B. (1995). Assessment of test user qualifications. *American Psychologist, 5*, 1, 14-23.

Muñiz, J. (1997). Aspectos éticos y deontológicos de la evaluación psicológica. En A. Cordero (Ed.), *Evaluación psicológica en el año 2000* (pp. 307-345). Madrid: TEA Ediciones.

Muñiz, J. y Bartram, D. (2007). Improving international tests and testing. *European Psychologist*, 12(3), 206-219.

Muñiz, J., Bartram, D., Evers, A., Boben, D., Matesic, K., Glabeke, K., Fernández-Hermida, J.R., y Zaal, J. (2001). Testing practices in European countries. *European Journal of Psychological Assessment, 17(3)*, 201-211.

Muñiz, J., y Fernández-Hermida, J.R. (2000). La utilización de los tests en España. *Papeles del Psicólogo, 76*, 41-49.

Muñiz, J., Prieto, G., Almeida, L., y Bartram, D. (1999). Test use in Spain, Portugal and Latin American countries. *European Journal of Psychological Assessment, 15*(2), 151-157.

Olea, J., Ponsoda, V., y Prieto, G. (1999). *Tests informatizados: fundamentos y aplicaciones*. Madrid: Pirámide.

Papeles del Psicólogo (2009). *Número monográfico sobre Ética Profesional y Deontología*. Vol. 30, 182-254.

Parshall, C.G., Spray, J.A., Kalohn, J.C., y Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer.

Peiró, J.M. (2003) La enseñanza de la Psicología en Europa. Un proyecto de Titulación Europea. *Papeles del Psicólogo, 86*, 25-33

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

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

Prieto, G. y Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77*, 65-71.

Shermis, M. D. y Burstein, J. C. (Eds.) (2003). *Automated essay scoring.* London: LEA.

Simner, M. L. (1996). Recommendations by the Canadian Psychological Association for improving the North American safeguards that help protect the public against test misuse. *European Journal of Psychological Assessment, 12*, 72-82.

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

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

Zenisky, A.L. y Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337-362.

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 122-130
http://www.cop.es/papeles

POSIBILIDADES Y RELEVANCIA DE LA OBSERVACIÓN SISTEMÁTICA POR EL PROFESIONAL DE LA PSICOLOGÍA*

POSSIBILITIES AND RELEVANCE OF SYSTEMATIC OBSERVATION BY THE PSYCHOLOGY PROFESSIONAL*

M. Teresa Anguera
Universidad de Barcelona

La metodología observacional en contextos naturales o habituales es un procedimiento científico que permite estudiar la ocurrencia de comportamientos perceptibles, de forma que se registren y cuantifiquen adecuadamente, lo cual implicará poder analizar relaciones de secuencialidad, asociación y covariación. En numerosas situaciones la metodología observacional es la mejor estrategia, o incluso la única posible; existen numerosos ejemplos en la evaluación de programas de baja intervención, interacciones entre iguales, entre niños y adultos, estudio de la interacción social en diferentes edades, discusiones en una pareja, o en el lugar de trabajo, repertorio conductual del bebé, posturas corporales en tareas específicas, comunicación kinésica no verbal (de profesores, deportistas, actores, etc.), análisis del movimiento en múltiples actividades, ocupación de espacios, o análisis de pautas de socialización y desocialización. Como se señala en el texto, la observación en contextos naturales supone desarrollar un procedimiento que resalta la ocurrencia de conductas cotidianas, y el análisis de las relaciones entre ellas. Estas relaciones se pueden identificar objetivamente a partir del proceso de análisis de datos idóneo en función del respectivo diseño observacional, combinando las perspectivas cualitativa y cuantitativa.
***Palabras clave**: Diseños observacionales, Registro, Codificación, Formatos de campo, Sistemas de categorías.*

Observational methods applied to natural or habitual contexts are scientific procedures that reveal the occurrence of perceptible behaviours, allowing them to be formally recorded and quantified. They also allow the analysis of the relations between these behaviours, such as sequentiality, association, and covariation. In many situations observational methods are the best strategy, or even the only strategy possible: examples are the assessment of low level intervention programs, interactions between peers, between children and adults, social interactions at different ages, disputes between couples or in the workplace, the behavioural repertoire of the baby, body posture for specific tasks, kinetic non-verbal communication (of teachers, sportsmen and women, actors, etc.), analysis of movement in multiple activities, occupation of a particular space, or the analysis of norms of socialization and desocialization.
As we stated in the text, observation in natural contexts involves developing a procedure that highlights the occurrence of everyday behaviours, and allows an analysis of the relations between them. These relations can be identified objectively as a result of the analysis of data linked to the corresponding observational design, combining the qualitative and quantitative perspective.
***Key words**: Observational designs, Recording, Coding, field formats, Category systems.*

¿POR QUÉ EL PSICÓLOGO NECESITA CONOCER Y UTILIZAR LA METODOLOGÍA OBSERVACIONAL?

Son diversos los personajes eminentes que han pronunciado frases célebres en las cuales se lee que el conocimiento se inicia con la observación: "La ciencia es el simple sentido común llevado al máximo: observación cuidadosa y rigor ante las falacias lógicas" (Huxley); "La observación fortuïta de este hecho despertó en mi una idea" (Bernard); "Si observas, conoces; si conoces, quieres, y si quieres, proteges" (Sabater Pi).

La observación tiene un inmenso potencial en el estudio del comportamiento humano. Nos permite estudiar las acciones y conductas perceptibles que tienen lugar de forma espontánea o habitual en el propio contexto, así como analizar los diversos procesos que tienen lugar en el ser humano y en los grupos y colectivos de los cuales forma parte. El psicólogo, en función de su especialización, deberá profesionalmente diagnosticar e intervenir en campos muy diversos, como, a modo de ejemplo, en programas de prosocialidad en una escuela infantil, programas de mantenimiento en actividad física de tercera edad, programas de apoyo social en barrios o comunidades en las que se han ubicado familias procedentes de cualquier nacionalidad, programas de educación para la salud en guarderías o en residencias geriátricas,

Correspondencia: M. Teresa Anguera. *Dept. Metodología de las Ciencias del Comportamiento. Instituto de Investigación del Cerebro, Cognición y Conducta (IR3C). Facultad de Psicología. Universidad de Barcelona. Campus Mundet. Pº Vall d'Hebron, 171. 08035 Barcelona. España. E-mailtanguera@ub.edu*

* Este trabajo forma parte de la investigación *Avances tecnológicos y metodológicos en la automatización de estudios observacionales en deporte*, que ha sido subvencionado por la Dirección General de Investigación, Ministerio de Ciencia e Innovación (PSI2008-01179), durante el trienio 2008-2011.

programas de asistencia a familias maltratadoras o negligentes en las pautas de crianza de sus hijos, programas preventivos del SIDA en adolescentes, programas de apoyo a familiares de jóvenes fallecidos por accidente, programas de relajación en deportistas, programas de socialización en centros penitenciarios, o en centros de acogida de menores, etc. En la amplia diversidad de situaciones que se le presentan, además, se hallará con la encrucijada de los diversos contextos, y de las externalidades (personales, sociales, políticas, etc.) que procedan del período en el cual se vive y de las cuestiones coyuntales que se planteen.

En un experimento se manipula el comportamiento, para lo cual se aplican técnicas diversas (se dan consignas, se establecen grupos mediante aleatorización, ...), pero en la observación no, sino que se estudia el comportamiento tal cual ocurre, sin más preocupación que el cumplimiento de los requisitos éticos para que esta observación sea posible (y a ello nos referiremos posteriormente) y el seguimiento del proceso para la objetivación de la parcela de realidad que nos interesa. Y en un estudio que siga la metodología selectiva siempre hay una elicitación de la respuesta, entendiendo por tal la solicitud o demanda de información a quién es nuestro objeto de estudio, diagnóstico o tratamiento, sea oralmente mediante entrevista, o haciendo uso de protocolos en el caso de cuestionarios, o utilizando el amplio espectro de tests psicológicos existentes.

Pero la realidad es muy diferente cuando el profesional de la psicología necesita conocer y estudiar el comportamiento tal cual se produce, de forma natural o espontánea, en cualquier contexto (familia, escuela, oficina, lugar de ocio, etc.), sea en un momento dado, o en el marco de un proceso específico.

Como metodología de estudio, uso profesional e investigación, su desarrollo ha sido imparable en las tres últimas décadas, tanto a nivel mundial como europeo y español. Desde una situación inicial enormemente borrosa, en la cual se adolecía de la necesaria sistematización y objetividad que caracteriza el método científico, hasta el momento actual, en el cual su estatus científico está perfectamente consolidado, su rigurosidad se halla garantizada, y los resultados obtenidos en muy diversas aplicaciones respaldan su credibilidad.

Como indica algún autor en la actualidad, nos hallamos en el borde de una 'revolución observacional' (Dawkins, 2007, pp. 148) de forma totalmente justificada, aunándose la fortaleza metodológica del estudio del comportamiento en contextos naturales con el perfeccionamiento incesante de nuevos recursos tecnológicos que actúan complementariamente.

METODOLOGÍA OBSERVACIONAL Y COTIDIANEIDAD

La metodología observacional es sumamente flexible y adaptable a los comportamientos y a los contextos. Ahora bien, como todo método, supone seguir un proceso de forma disciplinada y rigurosa. Es la cara y cruz de la moneda.

Por una parte, siempre implicará transcurrir por las cuatro grandes etapas de delimitación del problema, recogida de datos, análisis de datos e interpretación de resultados, las cuales pueden desglosarse de forma notable; pero, por otra, la riqueza de información que se obtiene es altamente valorable por captar directamente la parcela de realidad que nos interesa en su transcurrir cotidiano, sin tener que preguntar o pedir información o datos concretos (como sí ocurre, por el contrario, en la entrevista, cuestionario, tests psicológicos, ...), y sin tener que someter a los individuos y/o grupos (pacientes, clientes, usuarios, ...) a una situación experimental o cuasiexperimental en la cual se formulen determinadas consignas y se lleve a cabo un control de las variables implicadas.

La cotidianeidad que siempre hallamos como referente en el estudio observacional del comportamiento humano constituye el 'filón' de información al cual el psicólogo deberá acudir, y del cual tendrá que extraer adecuadamente los datos que precise, gestionarlos en función de sus objetivos, y analizarlos para la necesaria obtención de los resultados.

La actividad cotidiana supone un avance continuado en el tiempo en donde se suceden diversas conductas, homogéneas o dispares, es un recorrido por el curso vital de cada uno, es un proceso dinámico sumamente complejo del que en muchas ocasiones no somos conscientes de cuánto alberga (Anguera, 1999). El análisis de la cotidianeidad implica una contemplación de conductas diversas desde distintos niveles que se sitúan en una estructura piramidal. Si nos situamos en la cúspide de la pirámide, mediante el análisis de la cotidianeidad el psicólogo avanza en el conocimiento de la trayectoria vital de cualquier individuo. Al descender en la pirámide desglosamos la cotidianeidad en diferentes planos (familia, profesión, relaciones sociales, ocio, ...) y la contemplamos desde diferentes ámbitos transversales (salud, afecto, tensión, satisfacción, conflictos, ...).

¿CÓMO SE INICIA LA APLICACIÓN DE LA METODOLOGÍA OBSERVACIONAL?

Antes mencionamos las cuatro grandes etapas del método científico, y que, lógicamente, también lo son de la metodología observacional: delimitación del problema, recogida de datos, análisis de datos e interpretación de resultados

La primera decisión a adoptar, la de carácter sustantivo, consistirá en la delimitación temática de la actividad cotidiana (comportamiento perceptible del día-a-día) que nos interese estudiar, y deberá contemplarse en interacción con el entorno. Es decir, al objetivo posible de observación le afectarán tres únicas restricciones: Su carácter perceptible, la espontaneidad del comportamiento, y la naturalidad o habitualidad del contexto.

Cumplidas todas ellas, se puede ya delimitar el dominio temático que nos planteamos someter a observación. Como consecuencia, nos preguntamos ¿qué conductas podemos estudiar desde la metodología observacional? Nos interesan todas las que tienen un carácter perceptible, y, por consiguiente, las que captamos a través de nuestros órganos sensoriales (esencialmente vista y oído), aunque es obvio que no cubren el contenido semántico de la cotidianeidad, pero sí el de la cotidianeidad que percibimos. Ésta se halla conformada por innumerables conductas de contenido sumamente diversificado y de amplitud igualmente diferenciada, haciendo gala del carácter relativo de molaridad y molecularidad (por ejemplo, en la realización de actividad física, este recorrido de lo más molar a lo más molecular se materializaría desde la ejecución de una tabla de gimnasia, a la realización de saltos, carreras, flexiones, giros, …, y de éstos al análisis detallado del movimiento en cada uno de dichos saltos o flexiones), así como de su ubicación en algún lugar del rico espectro que contempla infinitas combinaciones entre ellas.

La segunda acotación a tener en cuenta es metodológica, por lo que nos planteamos para la observación la siempre difícil segmentación en unidades de conducta, conectada indudablemente a una segunda decisión acerca de la vertiente predominante en la complementariedad entre lo cualitativo y lo cuantitativo. Estas dificultades anunciadas nos llevan a formular serios interrogantes para los que no sabemos si existe respuesta, aunque lo intentaremos, así como a revisar posicionamientos tradicionalmente heterodoxos que posibiliten, desde la metodología, combinar en un feliz anclaje una amplia flexibilidad propia del análisis de lo cotidiano con el rigor del método científico.

COMPLEMENTARIEDAD DE PERSPECTIVAS CUALITATIVA Y CUANTITATIVA EN EL USO PROFESIONAL DE LA METODOLOGÍA OBSERVACIONAL

Hace ya varias décadas que surgió una polémica entre los planteamientos cualitativo y cuantitativo, que se atizó con el radicalismo con que unos y otros defendían sus respectivas posturas, y que Cook y Reichardt (1979) sintetizaron perfectamente. La metodología observacional no ha sido en absoluto ajena a esta confrontación inicial (Anguera, 1979, 2004; Anguera e Izquierdo, 2006), que después ha dado paso a una posición de complementariedad.

La polémica planteada podríamos considerarla como poliédrica, dado que se han desplegado diversos planos en el ruedo de la confrontación, y todos ellos tienen relevancia en este conflicto epistemológico-paradigmático-metodológico, que está teniendo una importante trascendencia (Bryman, 1994), aunque aquí no vamos a referirnos a este debate, que se contempla en otro artículo de este número monográfico (López, Blanco, Scandroglio, y Rasskin, en prensa). En este trabajo presentamos de forma sucinta el estado de la cuestión, cada vez más proclive a la complementariedad, y desde un planteamiento procedimental propio de la metodología observacional

Esta complejidad conceptual genera a los profesionales un buen número de interrogantes, indecisiones y dudas a nivel metodológico. La disciplina que impone el procedimiento, sin embargo, no debe estar reñida con la preservación de espontaneidad, o, al menos de la habitualidad con que contemplamos la producción de innumerables conductas, a modo de moléculas -formada cada una por átomos- que interactúan entre sí de forma variada y forman agrupaciones de mayor o menor envergadura. Indudablemente, la perspectiva desde la cual nos ubiquemos conceptualmente –siempre factible, pero siempre discutible- constituirá el referente que en cada caso asuma la responsabilidad primaria y vertebradora del planteamiento efectuado.

Priorización de la perspectiva cualitativa en la etapa de recogida de datos

La observación científica de la conducta interactiva, una vez definido específicamente el objeto de estudio (¿qué conductas nos interesa observar?, ¿de cuál o cuáles individuos?, ¿en qué contexto/s?, etc.), se inicia con el registro. ¿Y qué es registrar? Consiste simplemente en efectuar un volcado de la realidad sobre algún soporte determinado, y utilizando algún sistema de códigos. Este

apresamiento de la realidad sólo puede llevarse a cabo desde una vertiente procedimental de carácter cualitativo (Anguera, 2004).

En su acepción más extendida y aceptada, "las metodologías cualitativas se refieren a procedimientos de investigación que dan lugar a datos descriptivos (...)" (Bogdan y Taylor, 1975, p. 4). Esta afirmación, sin embargo, comporta implícitamente un trasfondo que se configuró en la década de los setenta, y que en la actualidad se halla en fase de depuración –no exenta de una sofisticación probablemente exagerada- que permite pensar claramente en su complementariedad con una metodología cuantitativa, a la que incluso puede llegar a superar en algunos casos en grado de formalización.

Hasta hace unos años, se trataba de una opción metodológica claramente marginal y con escaso poder de convocatoria. La situación en la actualidad parece comenzar a cambiar, aunque el paradigma vigente siga siendo el empírico positivo. Con frecuencia, la investigación cualitativa se describe como holística, preocupándose por los seres humanos y su ambiente en toda su complejidad, y encaja perfectamente en la fase de registro de un estudio observacional, siendo posible un despliegue taxonómico de modalidades de registro.

A modo de mera ilustración, podemos pensar en su gran adaptabilidad a lo que supondría el estudio de diversos comportamientos en todos los ámbitos que ofrece la vida cotidiana, como el familiar, profesional, de relaciones sociales, o de implementación de programas de intervención (Valles, 1997; Anguera, 1999; Rabadán y Ato, 2003; Sánchez-Algarra y Anguera, in press).

Si, como ejemplo, pensáramos aplicarlo tal cual al estudio observacional de la conducta interactiva en una situación de actividad cotidiana, son innumerables las discusiones y polémicas que pueden desprenderse de estas palabras, y de forma especial la detección y plasmación de incidentes clave en el registro mediante términos descriptivos, así como el situarlos en una cierta relación con el más amplio contexto social. ¿Cómo se logra por el psicólogo sin caer en una mera praxis acientífica y exenta de rigor? ¿Es que la metodología cualitativa debe quedar proscrita a un mero estudio exploratorio? ¿Se trata de una etiqueta con connotaciones de única verdad para algunos y peyorativas para otros? ¿Cómo debe el profesional resolver esta cuestión?

En el fondo se trata de un problema de operativización, que permitirá seleccionar la información considerada relevante, y como consecuencia recoger los datos de una u otra forma -en la actualidad cada vez gana más terreno la opción de grabar el episodio, digitalizarlo, y, como se ha dicho anteriormente, llevar a cabo una codificación informatizada-. Éste es el núcleo del problema, y la cuestión esencial en torno a la cual se conforman las actitudes a favor o en contra, y, por tanto, dando lugar a la vertebración de una metodología cualitativa o cuantitativa. En la primera fase del proceso que implica la metodología observacional se impone la metodología cualitativa, dadas sus amplísimas posibilidades en la obtención de los datos.

La estrategia que inspira la metodología cualitativa implica un intercambio dinámico entre la teoría, los conceptos y los datos con retroinformación e incidencia constante de los datos recogidos. En muchas ocasiones, además, el marco teórico, si existe, se halla sumamente debilitado (por la falta de comprobación empírica de sus postulados, sin que por realizar dicha afirmación se nos pueda acusar de reduccionismo), por lo que actúa de manera puramente referencial, a modo de metateoría.

Las situaciones problema no plantean un necesario cumplimiento de requisitos, a menos que en su formulación quede explícita la operativización que conlleve a iniciar y proseguir el proceso de investigación mediante una metodología cuantitativa; si nos planteamos un estudio relativo a tiempos de reacción ante determinado estímulo, como en psicología del tráfico, es indudable que no resulta pertinente la metodología cualitativa, pero en cambio es indiscutible en una investigación sobre conducta interactiva en el proceso de aplicar pautas de crianza de los hijos, o de irrupción de sujetos extraños en conducta comunicativa, o en el análisis de redes de apoyo social en tercera edad.

La matización que acabamos de realizar tiene posteriormente una enorme trascedencia. La inicial decisión sobre la selección de determinada información entresacada del entramado que constituye el problema va a conformar una trayectoria de partida correspondiente a la metodología cualitativa, aunque en un momento posterior, y en virtud de la complementariedad que defendemos, se quiebre para dar paso a la posición alternativa.

Es posible que en fases posteriores predomine el carácter cuantitativo de las operaciones a realizar, pero a nuestro juicio es secundario, a pesar de que tenga su importancia. La naturaleza del dato de partida la vamos a considerar constitutiva para la caracterización de la metodología cualitativa, aunque no todos los autores están de acuerdo con esta consideración.

Registro y codificación como segunda etapa del proceso

En el apartado anterior nos hemos referido a la existencia de una serie de modalidades de registro. La elevada lista de dichas modalidades de registro se reconduce en la actualidad al uso de programas informáticos, de los cuales existe un amplio listado. El impresionante desarrollo de los avances tecnológicos en los últimos años han dejado atrás una larga tradición de registros 'de papel y lápiz', que además conlleva importantes beneficios. Por una parte, evita errores derivados de un visionado analógico, en el cual se tenía que efectuar manualmente toda una serie de operaciones que, como mínimo, predisponían a un alto riesgo de cometer imprecisiones; en segundo lugar, aumenta la agilidad del proceso, así como la posibilidad de considerar unidades temporales cada vez más cortas, como el *frame* (1/25 segundo); en tercer lugar, se posibilita la transformación de ficheros de registro, permitiendo una intercambiabilidad altamente funcional y versátil de acuerdo con la estructura sintáctica de los respectivos programas informáticos; y, finalmente, en cuarto lugar, la información, en forma de bases de datos, queda disponible para someterla, ya en la tercera etapa del proceso, a un control de calidad de los datos y a un proceso cuantitativo de análisis, de forma que actúe un cierto grado de automatización tecnológica en el proceso, mediatizada únicamente por las decisiones adoptadas por el investigador en función de los condicionantes específicos de cada estudio.

El estudio observacional de la conducta interactiva, por considerar un ejemplo, se apoya en la suposición previa de actuación de posibles niveles de respuesta (canales interactivos), como podrían ser el intercambio de miradas, la distancia interpersonal, las vocalizaciones, el intercambio de mensajes verbales, etc. Además, van a tener que considerarse, por una parte, las co-ocurrencias o sincronías temporales producidas por acciones especificas de cada uno de los canales interactivos (sea *frame* a *frame*, o bien en intervalos temporales previamente establecidos), y, por otra, la sucesión de dichas co-ocurrencias a lo largo de un determinado período de tiempo o sesión. En consecuencia, se van a requerir programas informáticos que posibiliten la obtención de grandes matrices de códigos, de forma que cada fila consista en la relación de códigos correspondientes a las conductas o acciones co-ocurrentes en un instante dado, mientras que la sucesión de filas de la matriz corresponda al desarrollo diacrónico de la sesión considerada.

Son muchos los programas informáticos que se adaptan a estos requerimiento. A modo de ilustración, y además de programas de carácter general, como el EXCEL y el ACCESS, nos referimos a algunos en los que se ha comprobado la adecuación de sus prestaciones. Entre otros, destacamos THE OBSERVER (1993), SDIS-GSEQ (Bakeman y Quera, 1996), THÈMECODER (Pattern Vision, 2001), MATCH VISION STUDIO (Perea, Alday y Castellano, 2004), etc.

Como caso particular, limitado únicamente al canal interactivo verbal, podemos referirnos también a aquellas situaciones de conducta interactiva transcrita, y dispuesta, por tanto, en forma documental. El material de carácter textual presenta unas singularidades a tener en cuenta en esta segunda etapa del proceso, que habitualmente queda reconducida a un análisis de contenido (Krippendorf, 1980; Muskens, 1985; Roberts, 2000; Hogenraad, McKenzie y Péladeau, 2003), en cuyo desarrollo específico no entramos en este artículo. Por supuesto, también existen programas informáticos específicos, como AQUAD6, ATLAS.ti, MAXqda2, NUDIST, NVivo, etc., y queremos resaltar que precisamente el uso de estos programas ha favorecido el desarrollo de un tratamiento únicamente cualitativo, alcanzándose estructuras relacionales (familias, redes, etc.) que gozan de una cierta estabilidad, al menos aparente, y siempre a partir de la toma de decisiones del investigador.

El plano en que se sitúa el registro de la conducta es pobre e insuficiente si pretendemos una elaboración posterior -y también la cuantificación- de la plasmación de la conducta espontánea en episodios interactivos mediante la observación sistemática. Y de ahí la necesidad, mediante la codificación, de construir y utilizar un sistema de símbolos -que pueden ser de muy diversos órdenes- que permita la obtención de las medidas requeridas en cada caso.

La sistematización completa del comportamiento se logra mediante un sistema de códigos (icónicos, literales, numéricos, mixtos, cromáticos, etc.) que pueden adoptar una estructura de cadena, modular, en cascada, etc. Por supuesto, se puede llevar a cabo una simple codificación binaria (presencia/ausencia, que se podría codificar, respectivamente, como 1/0), o de un único tipo de elementos –por ejemplo, conducta interactiva verbal-, pero habitualmente interesará, como se ha indicado anteriormente, una codificación simultánea de varios aspectos concurrentes, por lo que es posible elaborar una sintaxis completa de cualquier situación de observación, que al-

canza un grado máximo de sistematización, sin requerir de ningún término descriptivo. En este caso conviene elaborar un manual de codificación. Indudablemente, esta transformación debe validarse en la medida en que sea factible la decodificación, con lo que se obtendría el correspondiente registro descriptivo en su forma inicial no sistematizada; precisamente en aquellos casos en los cuales no funcione esta operación (por obtenerse un registro descriptivo distorsionado o mutilado como consecuencia de la decodificación) podemos diagnosticar la naturaleza de los errores cometidos durante la codificación.

El manual de codificación se compone de dos partes bien diferenciadas. En la primera, se incluirán todos los términos (conductas) utilizados en el registro sistematizado con la inclusión del código correspondiente que las representa, y sin que haya ninguna limitación en cuanto al tipo de código. Y en la segunda parte del manual de codificación deben incluirse las reglas sintácticas que regulan el uso de los códigos, designando específicamente la sintaxis de la concurrencia de códigos y de la secuencia de dichas concurrencias (Anguera e Izquierdo, 2006).

Obviamente, dada la amplia y vasta gama de conductas que se generan en un episodio de conducta, se justifica perfectamente la construcción de un instrumento de observación *ad hoc*. En el estudio de la mayor parte de conductas, dada la práctica imposibilidad que supondría categorizar los comportamientos perceptibles correspondientes a cada uno de los canales (dado que implicaría cumplir los requisitos de exhaustividad y mútua exclusividad), el único instrumento posible de observación es el formato de campo (*field format*), caracterizado por la no forzosa necesidad de contar con marco teórico, y su carácter abierto (por tanto, deliberadamente no exhaustivo), multidimensional, de código múltiple, y autorregulable (Izquierdo y Anguera, 2001; Anguera, 2003; Anguera e Izquierdo, 2006).

En la Figura 1 se muestra esquemáticamente el papel de un formato de campo (de seis criterios o dimensiones) y un ejemplo de registro mediante una serie de configuraciones (filas de la matriz de registro formadas por los códigos correspondientes a las conductas co-ocurrentes), que reúnen las características de sincronía entre los códigos registrados (uno de cada dimensión, como máximo), y, por otra, la sucesión de configuraciones (filas) se ordena secuencialmente a lo largo del paso del tiempo.

Esta segunda etapa tiene un papel absolutamente relevante, dado que actúa como engarce entre la vertiente con predominancia cualitativa y la caracterizada por una predominancia cuantitativa. Su importante virtualidad consiste en favorecer grandemente la integración -más que la complementariedad- entre las perspectivas cualitativa y cuantitativa, y ello se consigue sin forzar ningún planteamiento epistemológico ni metodológico.

Priorización de la perspectiva cuantitativa en la tercera etapa del proceso

El proceso que sigue la metodología observacional, que en una primera fase ha requerido un especial cuidado para justificar el encaje de la metodología cualitativa, y donde la gran dificultad estribaba en la obtención del dato, una vez éste se ha obtenido gracias al proceso de codificación de la segunda etapa, en una tercera fase deberá llevase a cabo su control de calidad para la detección de posibles errores y su subsanación, para someterse, una vez superado dicho control, a los análisis adecuados en función del diseño observacional adecuado.

La figura del diseño observacional es sumamente relevante, dado que actúa como esqueleto y soporte metodológico de cualquier estudio en que se siga la metodología observacional. Nuestra propuesta, desarrollada en trabajos anteriores (Anguera, Blanco y Losada, 2001; Blanco, Losada y Anguera, 2003), parte del cruce de tres dimensiones generadoras de dichos diseños, y que, en la representación gráfica, son: Diámetro vertical, relativo al carácter idiográfico o nomotético del estudio; diámetro horizontal, relativo al carácter puntual o de seguimiento temporal; y circunferencias concéntricas, relativo a la uni o multidimensionalidad del estudio. En la

FIGURA 1
ESQUEMA DE CONSTRUCCIÓN Y FUNCIONAMIENTO DEL FORMATO DE CAMPO (ANGUERA E IZQUIERDO, 2006)

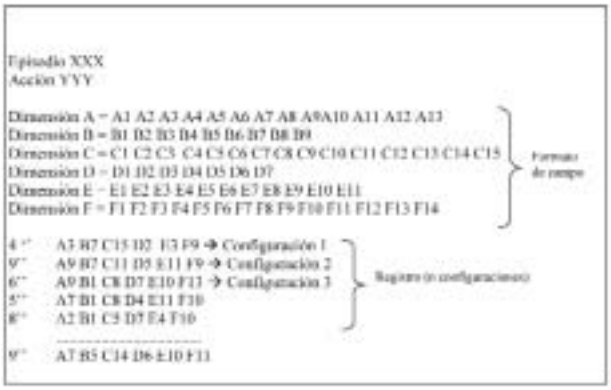


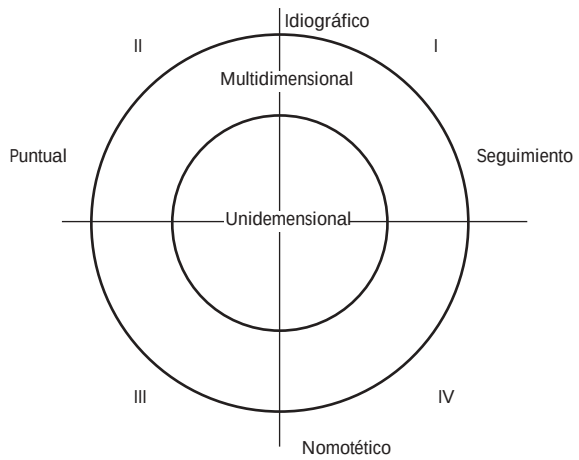
Figura 2 se muestra esquemáticamente, indicándose los ocho diseños observacionales resultantes:

Puntual / Idiográfico / Unidimensional
Puntual / Nomotético / Unidimensional
Seguimiento / Idiográfico / Unidimensional
Seguimiento / Nomotético / Unidimensional
Puntual / Idiográfico / Multidimensional
Puntual / Nomotético / Multidimensional
Seguimiento / Idiográfico / Multidimensional
Seguimiento / Nomotético / Multidimensional

Tradicionalmente se ha afirmado que los seguidores de la metodología cuantitativa tienden a traducir sus observaciones en cifras, y estos valores numéricos proceden de conteo o recuento, medida, o de constatación de la secuencia u orden, permitiendo descubrir, verificar o identificar relaciones simétricas o no entre conceptos que derivan de un esquema teórico elaborado de acuerdo con los criterios que rigen cada una de las situaciones de cotidianeidad que interese estudiar. Desde los planteamientos de la metodología cuantitativa, para llevar a cabo el contraste de la hipótesis será preciso cumplir el requisito de representatividad y aleatorización, lo cual comportará a su vez unas adecuadas técnicas de muestreo, a la vez que pueden proponerse sofisticadas técnicas de análisis (Anguera, 2004).

Si revisamos las revistas científicas en Psicología, en muchos países es justa la crítica de una endémica debilidad metodológica de la gran mayoría de los estudios observacionales de conductas en contextos naturales que son puestos en práctica por parte de instituciones tanto públicas como privadas. No obstante, en los países en los que existe una mayor tradición se aprecian, cada vez de forma más generalizada, importantes avances consistentes en el uso de recursos metodológicos sofisticados que permiten un rigor mucho más elevado, y que, si bien no todos proceden de estudios realizados en contextos naturales, sí serían análisis adecuados en muchos de ellos, siempre que se dispusiera de los datos adecuados.

FIGURA 2
DISEÑOS OBSERVACIONALES (ANGUERA, BLANCO Y LOSADA, 2001; BLANCO, LOSADA Y ANGUERA, 2003)



A modo de ilustración podemos señalar en este sentido, por su especial relevancia, la aplicación del análisis secuencial, sea en su forma clásica de análisis secuencial de retardos (Bakeman y Gottman, 1987, 1997), o bien de detección de T-Patterns (Magnusson, 1996, 2000; Anguera, 2005), así como el análisis de coordenadas polares (Sackett, 1980), basado igualmente en el análisis secuencial, y de muchos otros. El análisis secuencial, en cualquiera de los dos planteamientos, nos permitirá detectar la existencia de patrones de conducta que no son directamente perceptibles, y que tan útiles serán para el profesional de la psicología en procesos de diagnóstico y de intervención. Asimismo, el análisis de coordenadas polares nos permite obtener un mapa completo de relaciones entre conductas, pudiéndose objetivar en qué medida cada una de ellas repercute en otras, y si esta repercusión es activadora o inhibitoria.

La cuestión básica a la que nos tenemos de referir es que, en función del diseño planteado y de la naturaleza de los datos, procederá una u otra técnica analítica. En consecuencia, según cuál sea el cuadrante y el diseño en que se ubique un determinado estudio, resultarán idóneas unas u oras técnicas cuantitativas de análisis de datos (Anguera, Blanco y Losada, 2001; Blanco, Losada y Anguera, 2003). En cualquier caso, si la metodología cualitativa nos ayudó en la obtención del dato, la cuantitativa nos suministra los recursos analíticos para su tratamiento más conveniente.

INTEGRACIÓN DE LO CUALITATIVO Y LO CUANTITATIVO EN METODOLOGÍA OBSERVACIONAL

Es cierto que ya está cuajando una tradición consistente en que se produce en el desarrollo de la metodología observacional una combinación entre perspectivas metodológicas cualitativa y cuantitativa, sin entrar en la discusión acerca de si el paradigma cuantitativo se basa en el positivismo y si el paradigma cualitativo se base en el interpretativismo y constructivismo. Ambos flancos se han desarrollado independientemente, y muchas veces más preocupados por criticar la posición antagónica que por

mejorar la propia. Cada uno de ellos ha tenido amplio eco en revistas científicas acordes con su respectivo posicionamiento, e incluso se han acuñado términos y expresiones con lecturas antagónicas según el planteamiento desde el cual se usan.

Nuestra propuesta en este artículo se sitúa en una posición claramente complementaria entre las metodologías cualitativa y cuantitativa, y pensando en el profesional de la psicología. La lógica del proceso en la metodología observacional permite secuenciar las perspectivas, iniciando el estudio con una predominancia de la cualitativa, para después someterse a un determinado tipo de registro, mediante el importante apoyo del formato de campo, y a una codificación -preferentemente informatizada- generadora de una matriz de datos intercambiables formalmente, para finalmente invertirse el criterio y continuar con predominancia de la perspectiva cuantitativa (Anguera, 2004; Anguera y Izquierdo, 2006).

En numerosos estudios se ha comprobado su eficacia (Arias y Anguera, 2004, 2005; Jonsson, Anguera, Blanco-Villaseñor, Losada, Hernández-Mendo, Ardá, Camerino y Castellano, 2006), y queremos resaltar que el marco metodológico que lo permite de forma óptima es el de la metodología observacional, debido precisamente a las especificidades que la caracterizan.

COMPETENCIA DEL OBSERVADOR

La competencia observacional tiene una larga vida pero una corta historia. En el transcurso del último medio siglo ha sido esporádicamente estudiada, y tradicionalmente los psicólogos se han referido a ella de forma confusa, debido, probablemente, a que se han vinculado habilidades observacionales a cualidades admirables, a estrategias efectivas de aprendizaje, a una actualización del propio observador a partir del entrenamiento, y a la equiparación de la competencia observacional con el éxito (Anguera, Blanco, Losada y Sánchez-Algarra, 1999).

A pesar de los escasos estudios realizados sobre competencia observacional, han sido suficientes para considerar que 'el observador no nace, sino que se hace', y que su proceso de formación debe cuidarse con minuciosidad.

REFERENCIAS

Anguera, M.T. (1979). Observational Typology. *Quality & Quantity. European-American Journal of Methodology, 13* (6), 449-484.

Anguera, M.T. (1999). *Hacia una evaluación de la actividad cotidiana y su contexto: ¿Presente o futuro para la metodología?* Conferencia de ingreso en la Real Acadèmia de Doctors, Barcelona. [Reprinted in A. Bazán Ramírez y A. Arce Ferrer (Eds.) (2001), *Métodos de evaluación y medición del comportamiento en Psicología* (pp. 11-86). México: Instituto Tecnológico de Sonora y Universidad Autónoma de Yucatán].

Anguera, M.T. (2003). Observational Methods (General). In R. Fernández-Ballesteros (Ed.), *Encyclopedia of Psychological Assessment*, Vol. 2 (pp. 632-637). London: Sage.

Anguera, M.T. (2004). Posición de la metodología observacional en el debate entre las opciones metodológicas cualitativa y cuantitativa. ¿Enfrentamiento, complementariedad, integración? *Psicologia em Revista* (Belo Horizonte, Brasil), 10 (15), 13-27.

Anguera, M.T. (2005). Microanalysis of T-patterns. Analysis of symmetry/assymetry in social interaction. In L. Anolli, S. Duncan, M. Magnusson y G. Riva (Eds.), *The hidden structure of social interaction. From Genomics to Culture Patterns* (pp. 51-70). Amsterdam: IOS Press.

Anguera, M.T., Blanco, A. y Losada, J.L. (2001). Diseños observacionales, cuestión clave en el proceso de la metodología observacional. *Metodología de las Ciencias del Comportamiento, 3* (2), 135-160.

Anguera, M.T., Blanco, A., Losada, J.L. y Sánchez-Algarra, P. (1999). Análisis de la competencia en la selección de observadores. *Metodología de las Ciencias del Comportamiento, 1* (1), 95-114.

Anguera, M.T. e Izquierdo, C. (2006). Methodological approaches in human communication. From complexity of situation to data analysis. In G. Riva, M.T. Anguera, B.K. Wiederhold y F. Mantovani (Coord.), *From Communication to Presence. Cognition, Emotions and Culture towards the Ultimate Communicative Experience* (pp. 203-222). Amsterdam: IOS Press.

Arias, E. y Anguera, M.T. (2004). Detección de patrones de conducta comunicativa en un grupo terapéutico de adolescentes. *Acción Psicológica, 3* (3), 199-206.

Arias, E. y Anguera, M.T. (2005). Análisis de la comunicación en un grupo terapéutico de adolescentes: estudio diacrónico. *Revista de Psicopatología y Salud Mental del Niño y del Adolescente, M1*, 25-36.

Bakeman, R. y Gottman, J.M. (1986). *Observing interaction. Introduction to sequential analysis*. Cambridge: Cambridge University Press.

Bakeman, R. y Gottman, J.M. (1987). Applying observational methods: A systematic view. In J.D. Osofsky (Ed.) *Handbook of infant development* (pp. 818-853). New York: Wiley.

Bakeman, R. y Quera, V. (1996). *Análisis de la interacción. Análisis secuencial con SDIS-GSEQ*. Madrid, Ra-Ma [http://www.ub.es/comporta/sg/sg_e_programs.htm].

Blanco, A., Losada, J.L. y Anguera, M.T. (2003). Analytic techniques in observational designs in environment-behavior relation. *Medio Ambiente y Comportamiento Humano, 4* (2), 111-126.

Bogdan, R. y Taylor, S.J. (1975). *Introduction to qualitative research methods*. New York: Wiley & Sons.

Bryman, A. (1994). The debate about quantitative and qualitative research: A question of method or epistemology? *British Journal of Sociology, 35*, 75-92.

Cook, T.D. y Reichardt (Eds.) (1979). *Métodos cualitativos y cuantitativos en investigación evaluativa*. Madrid: Morata.

Dawkins, M.S. (2007). *Observing animal behaviour. Design and analysis of quantitative data.* Oxford: Oxford University Press.

Hogenraad, R., McKenzie, D.P. y Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality & Quantity. International Journal of Methodology, 37*, 221-238.

Izquierdo, C. y Anguera, M.T. (2001). The rol of the morphokinetic notational system in the observation of movement. En Ch. Cavé, I. Guaïtella et S. Santi (Eds.), *Oralité et Gestualité. Interactions et comportements multimodaux dans la communication* (pp. 385-389). Paris: L'Harmattan.

Jonsson, G.K., Anguera, M.T., Blanco-Villaseñor, A., Losada, J.L., Hernández-Mendo, A., Ardá, T., Camerino, O. y Castellano, J. (2006). Hidden patterns of play interaction in soccer using SOF-CODER. *Behavior Research Methods, 38* (3), 372-381.

Krippendorf, K. (1980). *Content analysis: An introduction to its methodology*. Bevery Hills, CA.: Sage.

López, J.S., Blanco, F., Scandroglio, B. y Rasskin, I. (en prensa). Una aproximación a las prácticas cualitativas en Psicología desde una perspectiva integradora. *Papeles del Psicólogo*.

Magnusson, M.S. (1996). Hidden real-time patterns in intra- and inter-individual behavior. *European Journal of Psychological Assessment, 12* (2), 112-123.

Magnusson, M.S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers, 32* (1), 93-110.

Muskens, G. (1985). Mathematical analysis of content. *Quality & Quantity. International Journal of Methodology, 19*, 99-103.

PatternVision (2001). *ThèmeCoder* [software], Retrieved January 15, 2002 [http://www.patternvision.com].

Perea, A., Alday, L. y Castellano, J. (2004). *Software para la observación deportiva Match Vision Studio*. III Congreso Vasco del Deporte. Socialización y Deporte / Kirolaren III Euskal Biltzarra. Sozializazioa era Virola. Vitoria.

Rabadán, R. y Ato, M. (2003). *Técnicas cualitativas para investigación de mercados*. Madrid: Pirámide.

Roberts, C.W. (2000). A conceptual framework for quantitative text analysis. *Quality & Quantity. International Journal of Methodology, 34*, 259-274.

Sackett, G.P. (1980). Lag sequential analysis as a data reduction technique in social interaction research. In D.B. Sawin, R.C. Hawkins, L.O. Walker y J.H. Penticuff (Eds.). *Exceptional infant. Psychosocial risks in infant-environment transactions* (pp. 300-340). New York: Brunner/Mazel.

Sánchez-Algarra, P. y Anguera, M.T. (in press). Qualitative/quantitative integration in the inductive observational study of interactive behaviour: Impact of recording and coding predominating perspectives. *Quality & Quantity. International Journal of Methodology, 43*.

The Observer [Software] (1993). Sterling, VA, Noldus Information Technology [www.noldus.nl]

Valles, M.S. (1997). *Técnicas cualitativas de investigación social. Reflexión metodológica y práctica profesional*. Madrid: Síntesis.

SUGERENCIAS PARA EL PROFESIONAL DE LA PSICOLOGÍA

1. Lectura del siguiente trabajo:

Anguera, M.T., Blanco, A. y Losada, J.L. (2001). Diseños observacionales, cuestión clave en el proceso de la metodología observacional. *Metodología de las Ciencias del Comportamiento, 3* (2), 135-160.

2. Se recomienda especialmente el uso de los siguientes programas informáticos de acceso libre:
   Kinovea [http://www.kinovea.org/en/]
   SDIS-GSEQ [http://www.ub.edu/gcai/gseq/]

Papeles del Psicólogo, 2010. Vol. 31(1), pp. 131-142
http://www.cop.es/papeles

UNA APROXIMACIÓN A LAS PRÁCTICAS CUALITATIVAS EN PSICOLOGÍA DESDE UNA PERSPECTIVA INTEGRADORA

AN APPROACH TO QUALITATIVE PRACTICES IN PSYCHOLOGY FROM AN INTEGRATIVE PERSPECTIVE

Jorge S. López, Florentino Blanco, Bárbara Scandroglio e Irina Rasskin Gutman
Universidad Autónoma de Madrid

Este trabajo pretende ofrecer una visión de conjunto de las prácticas cualitativas más frecuentes en Psicología, subrayando su compatibilidad con las prácticas de tipo cuantitativo y sus garantías metodológicas. Mostraremos, además, la lógica general de una estrategia de investigación cualitativa y revisaremos sucintamente las técnicas más habituales de recogida de información cualitativa. Por último, repasaremos algunas de las estrategias de análisis tradicionalmente vinculadas a la investigación cualitativa y cerraremos con un brevísimo comentario sobre algunas de las herramientas informáticas de asistencia al análisis cualitativo que creemos más útiles para el psicólogo.
Palabras clave: *Investigación cualitativa, Métodos cualitativos, Epistemología*

The aim of this work is to offer an overall view of the most frequently used qualitative practices in psychology, emphasizing their compatibility with quantitative practices and their methodological guarantees. Furthermore, we will show the general logic of a qualitative research practice and a brief examination of the most common techniques for gathering data and analyzing qualitative information. Finally, we will review some of the analytical strategies traditionally linked to qualitative research and we will end with a very brief remark about some assistance software tools for qualitative analysis that we find useful for psychologists.
Key words: *Qualitative research, Qualitative methods, Epistemology*

El objetivo de este trabajo no es tanto dar cuenta de los últimos avances que han tenido lugar en el dominio de los denominados "métodos cualitativos" cuanto, más bien, intentar que el psicólogo profesional tome conciencia de la posible utilidad de este tipo de aproximaciones estratégicas en su quehacer diario. Para ello necesitaríamos convencerle de que no se trata de estrategias metodológicas opuestas a las estrategias cuantitativas. Defenderemos que se trata de prácticas compatibles e, incluso, complementarias. Además, deberíamos hacerle ver que las prácticas cualitativas no son pre-científicas, subjetivas, irracionales o poco rigurosas. Y no lo son ni en el caso de la Psicología ni en el caso de otras disciplinas que hacen uso de este tipo de estrategias metodológicas con menos complejos. Por ejemplo, el Análisis Orgánico Cualitativo es una práctica fundamental en Química Orgánica que permite determinar la familia química a la que éste pertenece y orientar, si fuese necesario, análisis posteriores. En otro orden de cosas, por ejemplo, para entender las prioridades sanitarias de una comunidad de desplazados, y garantizar eventualmente la eficacia de un plan de intervención socio-sanitaria, un equipo de medicina comunitaria necesita estudiar cualitativamente las representaciones sobre la salud y la enfermedad que la comunidad maneja. Para ello estaría obligado a observar, participar en la vida de la comunidad y reconstruir rigurosamente la forma de vida en la que semejantes concepciones cobran sentido. Aunque toda esta información pudiera merecer un tratamiento cuantitativo ulterior, sin esta aproximación cualitativa inicial el equipo médico correría el riesgo de proyectar injustificadamente sobre la comunidad sus propias necesidades, errar el tiro y perder eficacia.

En cualquier caso, el lector debe reconocer que partimos con una cierta desventaja respecto al resto de trabajos de este volumen, que cuentan con garantías sobrevenidas. En cierta medida porque la Psicología ha hipotecado buena parte de su autoestima disciplinar a la posibilidad de percibirse a sí misma como una ciencia y, en concreto, como una ciencia positiva, cuyos conocimientos derivan de procesos de elaboración formal de fenómenos observables y cuantificables. No hay seguramente día más glorioso para el futuro psicólogo que aquel en el que, por fin, puede salir por la austera puerta de su facultad con el maletín del WAIS en su mano derecha, intuyendo el poder simbólico que le confiere la posibilidad de estampar un número en la casilla del CI.

Esta obsesión por garantizar, aunque sólo sea aparentemente, nuestro estatuto de ciencia nos ha llevado a exagerar en cierto modo nuestras estrategias de defensa frente a la irracionalidad, el subjetivismo, la palabrería o

Correspondencia: Jorge S. López. *Departamento de Psicología Social y Metodología. Facultad de Psicología Universidad Autónoma de Madrid. C/Iván Pavlov, 6. 28049 Madrid. España.*
E-mail: jorge.lopez@uam.es

la superstición. En nuestra opinión, esta tendencia ha convertido a la Psicología en una ciencia acomplejada y, por lo tanto, *normativamente hipertrofiada* (Blanco, 2002; Blanco y Montero, 2009) y algo *metodólatra* (Montero, 2006). Los psicólogos (especialmente los académicos, todo hay que decirlo) hemos creado una cultura normativa a todas luces excesiva, tanto que a veces nuestras normas (por ejemplo, las que regulan la escritura científica) sirven curiosamente para regular la conducta de otros colectivos científicos (ver Madigan, Johnson y Linton, 1995). Pocas disciplinas han invertido tantos recursos en consolidarse metodológicamente como la Psicología. La inversión ha llegado incluso a traducirse ocasionalmente en un área de conocimiento y se ha convertido en un elemento central de las historias oficiales de la Psicología: la Psicología es, por ejemplo, una de las pocas disciplinas que identifica su origen histórico, no con un hallazgo empírico o teórico, sino, curiosamente, con la creación de un laboratorio de psicología experimental (Jiménez, et. al., 2001).

El desprestigio relativo de las prácticas cualitativas en Psicología tiene mucho que ver con este "exceso de celo" metodológico, que lleva, en muchas ocasiones, a un uso gratuito, ornamental o estrictamente retórico de los números, como si su sola presencia en un informe de investigación o, en general, en un argumento, fuese una garantía de rigor y objetividad. Afortunadamente cada vez se aprecian más reacciones críticas, y de distinto signo, desde el propio dominio de la metodología ante esta banalización progresiva de su sentido histórico (ver, por ejemplo, Delgado, 2006; León, 2006).

En nuestra opinión es imprescindible que la Psicología empiece a cancelar esta absurda hipoteca histórica, haciéndose metodológicamente más flexible y, por lo tanto, más capaz de valorar la relevancia de las cuestiones a resolver, participando críticamente, incluso, en la definición de nuevas agendas de problemas. Nuestra obsesión por las garantías metodológicas, nuestro afán de neutralidad y objetividad, contribuye a mantenernos casi siempre ajenos a los debates públicos en los que se deciden estas nuevas agendas.

ALGUNAS IDEAS SOBRE EL SENTIDO DE LA DISTINCIÓN ENTRE PRÁCTICAS CUANTITATIVAS Y CUALITATIVAS

La distinción, y a menudo incluso oposición, entre lo cuantitativo y lo cualitativo es, como podemos ya intuir, solidaria con una cierta forma de ver el mundo, si se nos permite la expresión. Se nos entenderá mejor si el lector nos concede que una cierta visión del mundo implica (1) una idea sobre **lo que el mundo, esencialmente, es** (materia, energía, hechos, fenómenos, ideas, relaciones numéricas, construcciones sociales), (2) una idea sobre **cómo puede ser conocido** (empirismo, racionalismo, fenomenismo, positivismo, fenomenología, constructivismo), (3) una idea, o un conjunto de ideas, sobre **el modo de garantizar nuestro conocimiento del mundo** y (4) un conjunto de **valores** que orienten nuestra tarea. Es decir, aunque habría otras formas más sofisticadas de representarse el asunto, una cierta idea del mundo podría implicar, respectivamente: (1) una **ontología**, (2) una **epistemología**, (3) una **metodología** y (4) una **axiología**.

En concreto la distinción en el plano metodológico entre prácticas cuantitativas y cualitativas se suele hacer corresponder, tradicionalmente, y de forma algo maniquea, con sendas distinciones paralelas en los planos ontológico y epistemológico. En el plano metodológico, este dualismo, ontológico y epistemológico, se refleja, por lo tanto, sobre la distinción que da lugar a este trabajo, es decir, la distinción entre métodos cuantitativos y métodos cualitativos: los primeros se encargarían de establecer las garantías necesarias para **explicar** los fenómenos que tienen frecuencia, duración y/o intensidad, mientras que los segundos propondrían los criterios necesarios para **comprender** las acciones humanas y sus productos. Este dualismo sobrepasa siempre la esfera de los fenómenos científicos de los que parece depender para convertirse en un modo de proyectar en esta esfera valores estéticos, éticos, ideológicos y políticos, es decir, axiologías, lo que explica el tono tenso, incluso agrio, que a menudo presentan los debates *cuali vs. cuanti*.

Evidentemente, y aunque estas dos posiciones extremas se proyectan (y se han venido proyectando desde tiempos inmemoriales) en todos los dominios culturales, es justamente en el dominio de la Psicología donde los debates se han puesto más crudos. A menudo estos debates traducen, además, intereses espurios (poder, dinero, narcisismo intelectual), ajenos, en principio, a la vocación de autonomía de la racionalidad científica (Blanco, 2002). Nada más lamentable y aburrido, por frecuentado, que un dogmático acusando de dogmatismo a otro dogmático.

El esquema que acabamos de proponer no es demasiado original, pero nos permite intuir por dónde debería discurrir la posibilidad de eludir las partes más oscuras y mezquinas del debate, para convertirlo en un diálogo racional y fructífero. Por supuesto, la primera consecuencia que deberíamos extraer de este primer análisis es

que la radicalización interesada de las posiciones en litigio impide ver que entre los extremos (reflejos *vs.* acciones intencionales, por ejemplo) no hay un vacío absoluto sino una nube de situaciones o acontecimientos (reacciones circulares, respuestas condicionadas, acciones normativas, rasgos de personalidad), que exigen soluciones metodológicas estratégicas articuladas en torno a procedimientos (lo que algunos llamarían "técnicas") de distinta naturaleza, que deben ser ajustados a la lógica de los problemas. En definitiva, y como proponíamos en el primer párrafo de este trabajo las prácticas cualitativas y cuantitativas deben ser consideradas como recursos estratégicos con propósitos distintos que muy a menudo pueden ser conjugados en el seno de un mismo proceso de investigación o de intervención. Veamos por qué.

SOBRE LA ADECUACIÓN DE LA METODOLOGÍA CUALITATIVA A LOS CRITERIOS DE CALIDAD EN LA INVESTIGACIÓN PSICOLÓGICA: REFUTANDO ALGUNOS TÓPICOS

¿Es "subjetiva" la investigación cualitativa?

Al menos en dos sentidos se suele decir de las prácticas cualitativas que son "subjetivas". Por un lado, se dice que las prácticas cualitativas son subjetivas porque su objeto de estudio es, de una u otra forma, la subjetividad. Cabe decir a este respecto que si entendemos lo subjetivo (lo relativo a un sujeto) como una cualidad de los estados y procesos mentales, entonces también la psicología cognitiva experimental debería ser tildada de subjetivista.

Por otro lado, se suele acusar a las prácticas cualitativas de ser subjetivas en tanto que se supone que el conocimiento que proponen está anclado a la perspectiva "subjetiva" del investigador. Evidentemente todo proceso de investigación es, en este sentido general, subjetivo, pero también cabe decir que todo proceso de investigación aspira a transcender el punto de vista del observador y a producir conocimiento compartido o intersubjetivo. Es cierto que algunas prácticas cualitativas enfatizan el valor, a veces insustituible del investigador, pero también lo es que aspiran a que, bajo la perspectiva del investigador cualificado, el fenómeno quede acotado de tal forma que pueda ser compartido, revisado y criticado por cualquier otro investigador cualificado. Para un investigador "cuantitativo" no cualificado la diferencia entre un bebé autista y uno sordo puede pasar tan inadvertida como para un investigador "cualitativo" no cualificado lo sería la diferencia entre las indumentarias respectivas de un *skin head* pacífico y uno violento, con consecuencias igualmente nefastas en ambos casos.

¿Posee la investigación cualitativa la sistematicidad y la transparencia necesarias para generar conocimiento válido y fiable?

Tanto los métodos cualitativos como los cuantitativos son vulnerables a la asistematicidad y a la falta de transparencia de sus practicantes. Los diseños de investigación incluidos en los abordajes cualitativos, más abiertos y menos prefijados que los desarrollados desde abordajes cuantitativos, han proporcionado ciertamente cobijo a prácticas metodológicamente dudosas (Antaki, et. al., 2003). Sin embargo, diferentes autores, como Elliott, et. al. (1999), Miles y Huberman (1994) o Stiles (1993) han desarrollado un conjunto de estrategias consistente y sistematizado que permiten garantizar las cuestiones de control de calidad desde las prácticas cualitativas. El criterio de fiabilidad aparece reformulado en este ámbito a través del concepto de *dependencia-auditabilidad* y se garantiza estableciendo a través de todo el proceso de investigación procedimientos recursivos, explícitos y transparentes que permitan contrastar la consistencia de los resultados y su interpretación a través de diferentes investigadores, sujetos, contextos y momentos temporales. Por su parte, el criterio de validez interna se articula a través del concepto de *credibilidad/autenticidad* y comprende procedimientos destinados a garantizar la riqueza y la significación de la información recogida, su coherencia teórica y su contrastabilidad; a su vez, la validez externa se traduce en el concepto de *transferibilidad/adecuación* y se garantiza explicitando los criterios de generalización de los resultados y contrastando las predicciones en otros contextos y situaciones (Madill, et. al., 2000; Hammersley, 2007).

¿Permite la investigación cualitativa contrastar hipótesis y producir conocimiento generalizable?

En términos generales, el *sistema lógico* que fundamenta y guía la contrastación de hipótesis es el mismo en las prácticas cualitativas y cuantitativas, y se proyecta en la siguiente secuencia de acciones:

(1) condensación de la información,
(2) formulación de hipótesis,
(3) falsación a partir de la información muestral y
(4) examen de la posibilidad de generalización del resultado muestral a la población.

Las prácticas cualitativas permiten estructurar la información a través de sistemas conceptuales de codificación y categorización, plantear hipótesis o, al menos, conjeturas, formuladas a través de afirmaciones verbales, y someter dichas afirmaciones a procesos de falsación de carácter abierto y recursivo (Miller y Fredericks,

1987; Miles y Huberman, 1994). La generalización de los resultados a un marco poblacional definido se hace viable a partir del uso de criterios como la saturación o el contraste de los paralelismos teóricos y empíricos con otros contextos/fenómenos. El diseño de la muestra puede variar en función de las necesidades derivadas de los resultados y en función de los objetivos (por ejemplo, variación máxima, estratificación, tipicidad, intensidad u homogeneidad).

ALGUNOS ARGUMENTOS ADICIONALES PARA INTEGRAR LA METODOLOGÍA CUALITATIVA EN LA PRAXIS DE LA INVESTIGACIÓN

(1) En primer lugar, y dada su flexibilidad, las prácticas cualitativas son una *excelente herramienta para abordar de forma sistemática la exploración de fenómenos desconocidos y novedosos,* ofreciendo a su vez una adecuada aproximación a aquéllos que tienen lugar en contextos naturales.

(2) Permite igualmente *elaborar y difundir descripciones extensivas de gran riqueza,* que resultan de extrema utilidad para ofrecer un conocimiento directo de dichos fenómenos y representan una fuente para la generación de explicaciones e hipótesis tentativas.

(3) Posibilita una *aproximación sistemática a la perspectiva de los sujetos y a los significados que estos otorgan a sus acciones*, pudiendo servir a su vez como complemento a otro tipo de abordajes y orientando las explicaciones meramente especulativas sobre los resultados obtenidos mediante indicadores externos.

TABLA 1
ALGUNOS POSIBLES OBJETIVOS DE UNA INVESTIGACIÓN CUALITATIVA

- ✔ Sistematizar y analizar la información que ya existe sobre un fenómeno a partir del examen de fuentes secundarias (textos, imágenes, material audiovisual).
- ✔ Descubrir y analizar aspectos novedosos o inadvertidos de un fenómeno conocido.
- ✔ Describir los antecedentes, condiciones, características y consecuencias de un fenómeno novedoso.
- ✔ Valorar la posibilidad de aplicar una teoría ya existente a un fenómeno.
- ✔ Explorar cómo se construyen socialmente las percepciones y los discursos en relación con un tema.
- ✔ Analizar los patrones de interacción que desarrollan diferentes personas o grupos.
- ✔ Analizar los patrones culturales y su interpretación por parte de los miembros de una comunidad o grupo.
- ✔ Evaluar las percepciones que mantienen los destinatarios sobre un programa/acción de intervención.
- ✔ Producir ideas de forma colectiva, generar consenso y/o implicación para acciones/programas de intervención social.

(4) Ofrece la posibilidad de *alcanzar una perspectiva de los procesos,* ofreciendo herramientas para recoger, de forma prospectiva o retrospectiva, información sobre el modo en que se han desarrollado determinados fenómenos a lo largo de un período concreto.

(5) Permite *abordar fenómenos caracterizados por dinámicas interactivas de elevada complejidad,* que son difícilmente aprehensibles y sistematizables mediante indicadores prefijados, ofreciendo herramientas para la *detección de patrones* que pueden repetirse a lo largo de diferentes contextos situacionales o temporales.

(6) Permite el *análisis y seguimiento de los casos discordantes* a los que difícilmente se accede desde la perspectiva nomotética.

(7) Representa, finalmente, una excelente herramienta para posibilitar y sistematizar la *participación de los sujetos implicados* en los fenómenos objeto de estudio en la construcción conjunta del conocimiento sobre ellos.

LA LÓGICA GENERAL DE LAS PRÁCTICAS CUALITATIVAS

El desarrollo de una investigación de carácter cualitativo comienza, como cualquier otro proceso de indagación racional, por la elección del área de interés y la delimitación del objeto de estudio y continúa con la formulación de los objetivos y las preguntas a las cuáles se desea responder. En este sentido, existe un amplio abanico de preguntas a las que puede responder una investigación cualitativa. Por una parte, la metodología cualitativa es apropiada para dar respuesta a **preguntas de carácter abierto o exploratorio**, que son típicas del primer acercamiento a un fenómeno. Por otra, también puede dar respuesta a preguntas mucho más concretas, trabajando mediante la formulación de hipótesis que pueden ser sometidas a contrastación. Las preguntas y objetivos de la investigación definen inicialmente el tipo de diseño, la muestra, las técnicas de recogida de información y el tipo de análisis.

La definición del **diseño** comparte algunas dimensiones con los diseños de carácter cuantitativo. Así, una investigación cualitativa puede ser **transversal** (recoger información en un solo momento temporal) o **longitudinal** (recoger información en diferentes momentos temporales), **deductiva** (partir de una teoría y contrastarla a través de la información recogida) o **inductiva** (partir de la información recogida y construir una teoría a

partir de ella). Sin embargo, es importante destacar que la investigación cualitativa funciona habitualmente con diseños más abiertos que la investigación cuantitativa, ya que el investigador puede trabajar modificando y re-orientando las hipótesis, la muestra, las técnicas y/o los contenidos de la investigación en función de los resultados que obtenga con su trabajo. Es común igualmente en los sistemas de trabajo cualitativos alternar los procesos inductivos y los deductivos, modificando y enriqueciendo las formulaciones teóricas tentativas o las hipótesis iniciales a partir de los resultados que se obtienen y sometiendo éstas a contrastación con nuevos análisis o nueva información. Especial mención merecen los llamados diseños o métodos colaborativos, como la **Investigación-Acción-Participativa**, que integran a los sujetos en la definición de los objetivos y los procedimientos de la investigación e incorporan la transformación de la realidad social como elemento sustantivo e inherente al propio proceso investigador (véase López-Cabanas y Chacón, 1999).

La selección de la **muestra,** esto es, de los sujetos a quienes serán aplicadas las técnicas de recogida de información, o de los casos a los que se aplicarán las técnicas de análisis, tiene importantes consecuencias sobre los resultados de la investigación. Si el objetivo es obtener resultados descriptivos, debe tenerse en cuenta que las técnicas cualitativas no pretenden dimensionar numéricamente los fenómenos, sino ofrecer descripciones ricas, extensas o dinámicas de sus propiedades. Por ejemplo, una aproximación cualitativa nos permitiría describir diferentes modos de construcción de un discurso compartido por parte de integrantes de grupos violentos, o caracterizar las diferentes formas de acoso laboral que existen en un determinado sector profesional. También permiten acotar un fenómeno específico o un caso crítico, con el objeto de analizar los procesos que en él se dan, contextualizarlos, orientar la intervención o proporcionar datos que avalen el desarrollo de futuras investigaciones (por ejemplo, analizar por qué una población concreta presenta una tasa de suicidio juvenil elevada o explorar las razones por las cuales se ha generado un conflicto interno en una empresa). De este modo, la investigación de orientación cualitativa se beneficiará en menor medida de la selección de casos al azar y más del análisis de casos que puedan proporcionar información rica y extensa.

Cuando un estudio cualitativo pretende obtener resultados dirigidos a la *exploración* y *contrastación de hipótesis*, las técnicas cualitativas se enfrentan a dificultades similares a las que afrontan las técnicas cuantitativas a la hora de establecer en qué medida pueden generalizarse sus resultados. En este caso, es necesario seleccionar y hacer explícitos los criterios que permiten establecer equivalencias entre la muestra y la población, delimitando el marco y la validez de la generalización de los resultados. Los resultados derivados de una muestra de carácter cualitativo podrán generalizarse en la medida en que pueda defenderse que los procesos que los fundamentan en la muestra son equivalentes a los que se dan en la población a los que se desean aplicar. Para apoyar esta equivalencia, además de recurrir a los referentes teóricos y a los restantes estudios (si es que existen), algunas orientaciones cualitativas (por ejemplo el Análisis de Teoría Fundamentada o Método de Comparación Constante) utilizan el criterio denominado *saturación.* Este criterio se alcanza cuando la adición de nuevos sujetos a la muestra no modifica sustantivamente los resultados obtenidos previamente. Además, puede ser necesario modificar el rumbo inicial de la investigación, restringiendo el objeto de estudio y/o su contexto o bien ampliando la muestra. En la figura 1 mostramos la lógica que guía el uso del criterio de saturación.

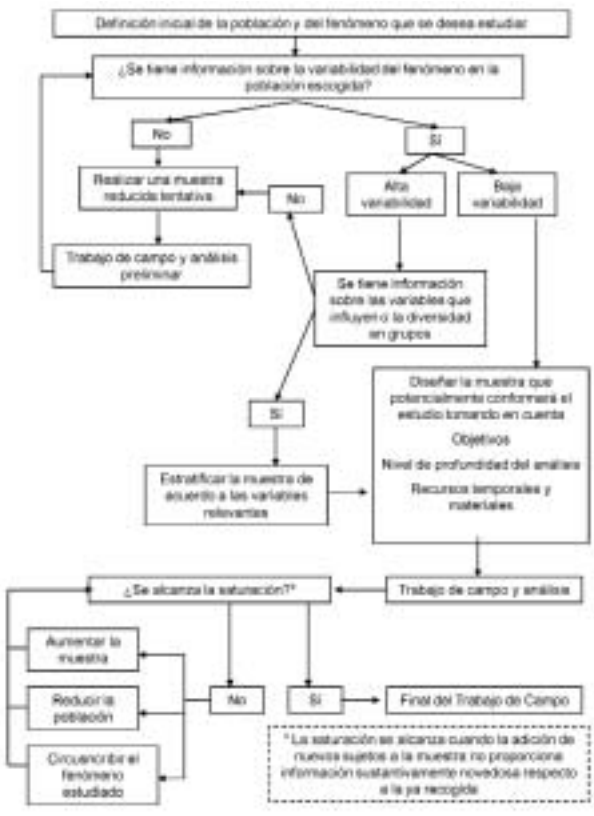
Aunque existen diferentes **técnicas de recogida de información** en este ámbito, el análisis cualitativo puede aplicarse sobre una gran variedad de sustratos de información, incluso aquellos generados desde enfoques eminentemente cuantitativos. Ello es así porque el tipo de análisis que se ponga en juego depende más de la "mirada" del investigador que de las características propias de la información. En cualquier caso, entre las técnicas más utilizadas en la aproximaciones cualitativas en Psicología destacan el análisis de fuentes secundarias (cualquier formato de documentos de texto, imágenes y/o audio procedentes de fuentes distintas al propio investigador/a), la observación, la entrevista, la historia de vida y una gran variedad de técnicas grupales (grupo de discusión, entrevista de grupo, grupo nominal, entre otras muchas). Además, las nuevas tecnologías de la información están dando lugar a un amplio conjunto de posibilidades que amplían y modifican las técnicas más tradicionales. En la Tabla 2, condensamos algunas de las técnicas más usuales, señalando su denominación, describiéndolas brevemente e indicando a partir de qué perspectivas teóricas suele ser más habitual su análisis.

ESTRATEGIAS DE ANÁLISIS

En el caso de la metodología cualitativa, y en lo tocante al análisis de la información, es más adecuado hablar

de "estrategias" que de "técnicas", ya que los procedimientos tienen un carácter más abierto y flexible que el de las aproximaciones cuantitativas (Gordo y Serrano, 2008). Dentro del marco de la investigación cualitativa existen numerosas propuestas que establecen procedimientos generales para el desarrollo del proceso de análisis (véase p.ej. la excelentes sistematizaciones de Miles y Huberman, 1994, Ryan y Bernard, 2000 o González-Rey, 2000). Sin embargo, existen corrientes que han estructurado sistemas de trabajo más específicos a los que se les ha otorgado un nombre concreto que se utiliza de forma más o menos consensuada por la comunidad científica. En la Tabla 3, recogemos un resumen de estos últimos, seleccionando entre las numerosas alternativas existentes en la literatura aquéllas que se hallan más próximas al ámbito de aplicación psicológico. En la Tabla 4 hemos incluido algunos ejemplos de investigación que se podrían llevar a cabo desde cada práctica cualitativa propuesta.

FIGURA 1
PASOS PARA LA DEFINICIÓN DE UNA MUESTRA DESDE LA PERSPECTIVA CUALITATIVA CON EL CRITERIO DE SATURACIÓN (FUENTE: LÓPEZ Y SCANDROGLIO, 2007)



DÓNDE AMPLIAR INFORMACIÓN SOBRE METODOLOGÍA CUALITATIVA

La publicación de textos en el ámbito de la metodología cualitativa ha experimentado un importante crecimiento en los últimos años. Sin perjuicio de que puedan existir otras opciones igualmente válidas, nos permitimos orientar al lector sobre algunas referencias accesibles que pueden ser de utilidad.

Una aproximación sintética al tema, desarrollada con algo más de extensión que el presente trabajo, puede encontrarse en López y Scandroglio (2007). Para una visión introductoria más amplia, son útiles los textos de González-Rey (2000), Gordo y Serrano (2008) y Vallés (2000). El texto de Galindo (1998) ofrece una revisión en profundidad de diferentes técnicas de investigación cualitativa y la obra de Gutiérrez y Delgado (1994) brinda una mayor profundización en los aspectos epistemológicos, con especial atención a la perspectiva de los sistemas complejos. En el texto editado recientemente por Gordo y Serrano (2008) se encuentran ejemplos de la mayoría de las prácticas de recogida y análisis de datos que hemos propuesto. Como ejemplos recientes de investigaciones de carácter cualitativo, publicadas en nuestro contexto y en el ámbito psicológico, pueden ser útiles los trabajos de López y cols. (2008), Scandroglio (2009), Martín (2005), Blanco y Sánchez-Criado (2006), Rasskin, (2007) o Gómez-Soriano y Vianna (2005)

HERRAMIENTAS INFORMÁTICAS DE APOYO A LAS PRÁCTICAS CUALITATIVAS

El desarrollo de la informática ha introducido cambios importantísimos en las prácticas cualitativas, afectando tanto a la recogida y tratamiento inicial de los datos, como a los procedimientos de análisis (véase Lewins y Silver, 2006 para una panorámica). El desarrollo y mejora de los soportes digitales de registro y almacenamiento de información audiovisual (videocámaras, grabadores de audio, scanners, memorias portátiles, etc.), ha permitido en muchos casos que los investigadores cualitativos hayan visto "sus sueños hechos realidad". Al margen de las posibilidades logísticas (organización y edición superficial de la información) que incorporan de serie muchos de estos dispositivos, es importante recordar también la importancia del desarrollo de programas de ordenador de edición audiovisual que permiten filtrar y ordenar adecuadamente la información registrada sobre el terreno. Pero el desarrollo de la informática ha sido especialmente decisivo en la sofisticación y mejora de los **procedimientos de transcripción de material audiovisual** previos al análisis y, sobre todo, en el diseño de

TABLA 2
ALGUNAS TÉCNICAS DE RECOGIDA DE INFORMACIÓN

DENOMINACIÓN		DESCRIPCIÓN	PERSPECTIVA DE ANÁLISIS CUALITATIVO
Análisis de material documental		Recopilación y análisis de documentos escritos, visuales o audio-visuales.	Amplias posibilidades de análisis que recorren todo el espectro de estrategias de análisis cualitativo.
Observación	Participante	Recogida de información a partir de la percepción de un agente externo que se implica en el suceso observado e interacciona con los actores.	Especialmente vinculada a la perspectiva etnográfica, pero abarcable también desde otras perspectivas cualitativas.
	No participante	Recogida de información a partir de la percepción de un agente externo no implicado en el proceso observado.	Amplias posibilidades de análisis cualitativo condicionadas por el grado de sistematicidad y estructuración previa de la observación.
Entrevista		Obtención de información a partir de una interacción comunicativa dialógica entre el investigador y el sujeto.	Amplias posibilidades de análisis cualitativo dependiendo en gran medida del grado de estructuración de la interacción.
Historia de vida/ autobiografía asistida		Recogida de información a partir de documentos y/o de la interacción comunicativa sobre la forma en que una persona construye y da sentido a su vida.	Especialmente vinculada a la perspectiva etnográfica pero, utilizable también desde otras perspectivas teóricas (hermenéutica, genealogía)
Técnicas grupales	Grupo de discusión	Interacción moderada por el investigador entre un grupo pequeño de sujetos que no se conocen entre sí y que guardan una relativa homogeneidad en relación con el aspecto investigado.	Amplias posibilidades de análisis cualitativo destacando especialmente el análisis del discurso.
	Entrevista de grupo	Interacción comunicativa entre el investigador y un grupo pre-existente.	Amplias posibilidades de análisis cualitativo.
	Técnicas de análisis y toma de decisiones	Generación de percepciones o decisiones consensuadas por un grupo a través de pautas estructuradas de interacción guiadas por el investigador.	Amplias posibilidades de análisis cualitativo.
Técnicas de dramatización y role-playing		Escenificación de situaciones en las que los sujetos deben actuar desempeñando papeles o funciones determinadas.	Amplias posibilidades de análisis cualitativo
Auto-informes	Cuestionarios	Recogida de información a partir del registro por escrito de las respuestas que da un sujeto a un conjunto prefijado de preguntas.	Especialmente vinculada al Análisis de Contenido, dado que la propia técnica pretende generar una información reducida y condensada. Sin embargo, es susceptible de un amplio rango de posibilidades de análisis cualitativo.
	Auto-registros	Recogida de información escrita por parte del propio sujeto investigando sus conductas y/o los contextos en los que tienen lugar.	Vinculadas tanto al estudio de caso único como a un amplio conjunto de posibilidades de análisis cualitativo.
Pruebas subjetivas		Recogida de información a partir de la calificación o clasificación que hace un sujeto de conceptos, objetos o personas, siguiendo pautas de un amplio grado de flexibilidad.	Muy vinculadas al estudio de caso único, pero susceptibles de un amplio conjunto de posibilidades de análisis cualitativo.
Test proyectivos		Obtención de información sobre la personalidad y/o cogniciones de un sujeto a través de su respuesta no estructurada a un conjunto de estímulos de carácter ambiguo.	Muy vinculadas al estudio de caso único, pero mantienen actualmente un amplio rango de posibilidades de análisis cuantitativo y cualitativo.
Cultura material (enseres, herramientas, disposiciones normativas, productos artísticos)		Obtenemos información general y específica sobre formas de organización de la actividad culturalmente vinculantes.	Especialmente útil para los enfoques genealógicos, *neomaterialistas*, *Actor-Network Theory* y psicología histórico-cultural.

programas de asistencia al análisis cualitativo.

Transana (http://www.transana.org/) es probablemente la herramienta de transcripción de material audiovisual más utilizada en la actualidad en el ámbito de las ciencias sociales. El trabajo con esta herramienta permite elegir entre diversos tipos de transcripciones, desde los registros narrativos informales hasta las transcripciones jeffersonianas que codifican todas las propiedades relevantes del habla (entonación, fonética). Si trabajamos con registros audiovisuales, *Transana* permite, además,

TABLA 3
ALGUNAS PRÁCTICAS DE ANÁLISIS CUALITATIVO

Análisis de Contenido Clásico	**Referencias**: Bardin (1967), Piñuel (2002) **Objetivo:** Condensar la información manifiesta de un texto en material estructurado susceptible de ulterior análisis	**Procedimiento**: (1) Estructuración y selección de la información. (2) Establecimiento inicial de categorías de carácter exhaustivo y excluyente a partir de los presupuestos teóricos y el análisis preliminar del texto. (3) Puesta a prueba del sistema de categorías y re-formulación. (4) Codificación definitiva del texto. (5) Establecimiento, en su caso, de índices de acuerdo inter-jueces. (6) Realización, en su caso, de análisis ulteriores (análisis de contenido latente, contraste de hipótesis, análisis cuantitativo) a partir de la codificación realizada
Descripción etnográfica	**Referencias:** Velasco y Díaz de Rada (1997), Wolcott (1999) **Objetivo**: Establecer una descripción exhaustiva de un determinado fenómeno social y de los significados atribuidos por los propios actores.	**Procedimiento:** (1) Descripción detallada del grupo o individuo que comparte con otros una cultura. (2) Análisis de los temas y las perspectivas del grupo. (3) Interpretación de los significados de la interacción social. (4) Generación de un retrato cultural holístico del grupo cultural que incluye el punto de vista de los actores (*emic*) y las interpretaciones y visiones del investigador respecto a la vida social humana (*etic*).
Inducción Analítica	**Referencias:** Manning (1982), Taylor y Bogdan (1984) **Objetivo:** Generar una teoría sobre un fenómeno social contrastando inductivamente su validez.	**Procedimiento:** (1) Definición inicial del fenómeno a explicar. (2) Formulación de una explicación hipotética. (3) Examen de un caso, en función de la hipótesis, para determinar si la hipótesis se ajusta a los hechos. (4) Validación o reformulación de la hipótesis o re-definición del fenómeno. (5) Integración de la información procedente de nuevos casos. (6) Nueva validación para lograr un buen nivel de *certeza práctica*, o nueva reformulación de la hipótesis o redefinición del fenómeno. (7) Establecimiento de una *relación universal*. (8) Integración teórica que incluye la descripción del fenómeno y un conjunto de proposiciones explicativas del objeto de estudio.
Análisis de Teoría Fundamentada-Método de Comparación Constante	**Referencias:** Glaser y Strauss (1967), Trinidad, Carrero y Soriano (2006) **Objetivo:** Generar una teoría sobre un fenómeno social, derivándola del análisis de la información empírica disponible y sometiéndola a un proceso de contraste recursivo de carácter inductivo y deductivo.	**Procedimiento:** (1) Muestreo inicial guiado teóricamente. (2) Colecta y estructuración de la información. (3) Codificación abierta: generación de categorías a través de la comparación de las unidades informativas y el hallazgo de elementos comunes. (4) Saturación de categorías: definición formal de categorías mediante el establecimiento de sus propiedades (condiciones, interacciones, tácticas/estrategias, consecuencias) y dimensiones. (5) Muestreo teórico: selección de las categorías teóricamente relevantes. (6) Categorización axial: integración en ejes de relación de categorías y propiedades y formulación de hipótesis. (7) Delimitación de la teoría, en función de los criterios de parsimonia y alcance. (8) Validación de la teoría a través del retorno a los textos y, en su caso, nuevos casos.
Análisis Retórico y de la Argumentación	**Referencias:** Albaladejo (1991), Bauer y Gaskell (2000), Plantin (1998), Vega (2003). **Objetivo:** Determinar los recursos retóricos y argumentativos que emplean los individuos para alcanzar el objetivo de ser persuasivos.	**Procedimiento:** (1) Establecimiento del carácter general del discurso o el texto en base a sus funciones y al auditorio al que va dirigido. (2) Esquematización del discurso o el texto identificando sus constituyentes formales o las *partes orationis*: el *exordio*, la *narración*, la *argumentación* y la *exhortación*. (3) Análisis de cada una de las partes y la relación que mantienen entre sí, especificando sus figuras retóricas, argumentativas y tropos.

vincular en tiempo real las transcripciones del material lingüístico con las imágenes. Más allá de su utilidad en la transcripción, esta herramienta nos permite categorizar nuestras transcripciones y relacionarlas entre sí con la misma lógica de una base de datos, lo que, a ciertos efectos, coloca a *Transana* en un espacio de transición entre las herramientas de transcripción y los programas de análisis de asistencia al análisis cualitativo.

FIGURA 2
ASPECTO DE UNA PANTALLA DE TRABAJO EN *TRANSANA*

FIGURA 3
ASPECTO DE UNA RED CONCEPTUAL EN *ATLAS.TI*

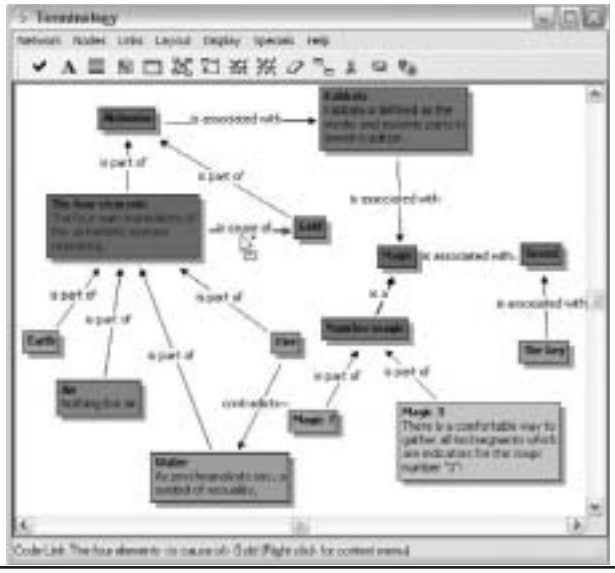


TABLA 3
ALGUNAS PRÁCTICAS DE ANÁLISIS CUALITATIVO (continuación)

Análisis de la Conversación (AC) **Análisis del Discurso (AD) y Análisis Crítico del Discurso (ACD)**	**Referencias:** AC: Drew (2003), Heritage (2004), Antaki y Díaz (2006) AD: Potter y Wetherell (1987), Willig (2003), Haidar (1998) ACD: Blommaert (2004), Wodak (2001). **Objetivo:** Determinar las prácticas sociales a través del lenguaje y/o de otros elementos simbólicos (por ejemplo, imágenes) que realizan las personas de un contexto o grupo social determinado.	**Procedimiento:** (1) Primera clasificación abierta, dictada por el objeto de estudio. (2) Búsqueda de variabilidad y consistencia a través de los repertorios interpretativos. (3) En el caso del AD, examen de las funciones del tipo de argumentación o construcción discursiva y análisis de la producción del discurso como una forma de solucionar problemas, identificando el problema y el modo en que ha sido resuelto (Potter y Wetherell, 1987), o como forma de plasmación de la relaciones de poder (Foucault, 2006). En esta misma línea, el ACD se centra en el estudio de prácticas discursivas a través de las cuales emerge la desigualdad social, integrando en sus análisis los aportes de la teoría social y el estudio del contexto sociopolítico y económico que posibilitan dichas prácticas. En el caso del AC, se analiza la estructura colaborativa que emerge de la conversación, identificando tanto los elementos que apuntalan la organización secuencial de tal conversación, como el manejo de turnos para tomar la palabra y las prácticas de apertura, sostenimiento y cierre de la conversación.
Análisis genealógico	**Referencias:** Foucault (1975/2005), Álvarez-Uría (2008) **Objetivo:** Problematizar y visibilizar las condiciones de posibilidad histórico-materiales de los fenómenos bajo estudio.	**Procedimiento:** (1) Visibilizar el problema (problematizar una práctica social). (2) Organizar períodos en la génesis de la práctica a partir de fuentes secundarias (historia política social e institucional de la práctica, documentos normativos). (3) Analizar la génesis del campo general en el que cobra sentido la práctica. (4 Estudiar la transformación del campo y de la práctica.
Análisis dramatúrgico	**Referencias:** Burke (1945/1984), Goffman (1959/1993) **Objetivo:** Estudiar el modo en que las acciones sociales (reales o ficcionales) se insertan en situaciones y contextos significativos.	**Procedimiento:** (1) Acotación del segmento dramatúrgico a estudiar. (2) Determinación de los casos (actor, acto, propósito, agencia y escenario). (3) Seguimiento de la dinámica temporal de los casos. (4) Determinación de ratios (relaciones diádicas) entre casos que dan lugar a la anomalía (alteración del curso normal de los acontecimientos) que provoca la pertinencia de la acción social o del relato.

Entre estos últimos, nos inclinamos por recomendar *Atlas.Ti* (http://www.atlasti.com/de/ productintro_es.html), un ambicioso programa que además de las funciones más habituales de codificación y análisis de material textual, facilita el análisis de registros sonoros, material de video y documentación gráfica. El corazón del programa es la Unidad Hermenéutica, un espacio virtual en el que podemos construir y reconstruir permanentemente las estructuras, mapas conceptuales o hipertextos, que vinculan los materiales con los que trabajamos en virtud de nuestras hipótesis.

Una mención especial merece, en nuestra opinión, *QDAMiner* (http://www.provalisresearch.com/QDAMiner/QDAMinerDesc.html), un programa de análisis de material lingüístico especialmente intuitivo, que se puede completar con dos herramientas vinculadas de asistencia al análisis estadístico de textos (*Wordstat*) y al análisis cuantitativo (*Simstat*) de variables cualitativas, que nos permiten ir más allá de la lógica habitual del análisis cualitativo y superar, el absurdo dualismo que criticábamos al comienzo de este trabajo.

TABLA 4
EJEMPLOS DE PROYECTOS DE INVESTIGACIÓN VINCULADOS A PRÁCTICAS CUALITATIVAS CONCRETAS

	Ejemplos de investigación
Análisis de contenido clásico	Analizar las respuestas a una entrevista realizada a un conjunto reducido de profesores de una institución educativa para conocer los problemas más relevantes que enfrentan en su tarea y poder construir un cuestionario de respuestas cerradas que se aplique a una muestra más amplia. Analizar los estereotipos de género que contienen los textos publicados en un periódico de prensa local para evidenciar la posible persistencia de representaciones sexistas y sus tipologías en un contexto concreto
Descripción etnográfica	Investigar las formas de relación y la estructura de una comunidad rural mediante observación participante y entrevistas en profundidad. Investigar la estética y los modos de consumo de los jóvenes que participan en el *botellón* mediante observación participante y entrevistas informales.
Inducción Analítica	Validar y, en su caso, realizar propuestas de modificación a la Teoría del Comportamiento Planificado sobre relaciones entre actitud y conducta a partir de entrevistas a personas que consumen drogas de síntesis. Validar y, en su caso, realizar propuestas de modificación de un modelo concreto de evolución de las fases de duelo a través de entrevistas, tanto a los profesionales que han seguido a familiares de víctimas de accidentes de tráfico como a los propios familiares.
Método de Comparación Constante / Análisis de Teoría Fundamentada	Construir una teoría para explicar por qué las familias niegan o conceden la donación de órganos de un familiar fallecido a partir de las entrevistas realizadas con personas que han participado en este proceso (familiares y coordinadores). Construir con la información obtenida a través de entrevistas un modelo teórico que relacione los diferentes factores que influyen en la calidad de vida de las personas que se ven en la necesidad de atender a un familiar con una enfermedad crónica incapacitante.
Análisis Retórico y de la Argumentación	Comparar los discursos pronunciados por un grupo político situado en el Gobierno y otro grupo político situado en la Oposición para justificar la existencia de corrupción entre su integrantes, a partir del análisis de material escrito y audiovisual publicado en un período concreto de tiempo. Analizar los argumentos que utiliza una corporación de telecomunicaciones para evadir la respuesta a determinadas demandas de sus clientes a través del análisis de material publicitario, de textos publicados en su web y del registro de llamadas de usuarios.
Etnometodología Análisis de la Conversación, Análisis del Discurso y Análisis Crítico del Discurso	Estudiar la forma en que los inmigrantes construyen su identidad personal y social en un contexto hostil, mediante la realización de entrevistas grupales a grupos formales e informales de de ciudadanos inmigrantes. Analizar las estrategias que utilizan los usuarios de redes de contacto en internet para establecer relaciones con menores, mediante el análisis de mensajes y conversaciones por chat archivadas. Estudiar la forma en la que los fumadores defienden la continuidad de su hábito a partir de la realización de grupos de discusión.
Análisis genealógico	Estudiar las relaciones entre las formas de organización autobiográfica del sufrimiento en desplazados por violencia socio-política y los procesos de constitución histórica de las instituciones políticas y las prácticas sociales implicadas en la gestión del desplazamiento. Estudiar las aportaciones de la confesión cristiana a la constitución histórica de la psicología clínica. Analizar los orígenes históricos del discurso de las personas con trastornos alimentarios.
Análisis dramatúrgico	Analizar las formas de dominación y ejercicio del poder entre los distintos elementos del *staff* clínico en un servicio hospitalario, como procedimiento para estudiar las causas de los conflictos laborales que se dan entre ellos. Estudiar las representaciones sociales que los españoles tienen de los psicólogos analizando dramatúrgicamente (actor, acto, propósito, agencia y escenario) su presencia en las series televisivas.

CONCLUSIONES

Hemos intentado mostrar que las herramientas de control de calidad en la investigación y la intervención psicológicas deben estar basadas en el examen de los niveles de sistematicidad, transparencia y explicitación de los procedimientos de consenso intersubjetivo que permiten aproximarse al objeto de estudio. Desde este planteamiento, aunque la denominada metodología cualitativa comprenda un conjunto de objetivos y procedimientos más abiertos y diversificados, posee en la actualidad todo un abanico de recursos que le permiten afrontar con las adecuadas garantías de calidad los procesos de descripción, contrastación y generalización. A su vez, dada su especial adecuación para abordar de forma flexible el estudio de los fenómenos psicológicos de elevada complejidad y variabilidad temporal representa un conjunto de alternativas de excepcional valor para superar algunos de los obstáculos presentes en la investigación psicológica actual. La marginación del abordaje cualitativo en el contexto académico resulta así desde nuestra visión un anacronismo y una limitación que reflejan más una situación de desconocimiento que un posicionamiento consciente sobre los fundamentos de la actividad científica.

REFERENCIAS

Albaladejo, T. (1991). *Retórica*. Madrid: Editorial Síntesis

Álvarez-Uría, F. (2008). El método genealógico: ejemplificación a partir del análisis sociológico de la institución manicomial. En Á. Gordo y A. Serrano (Ed.) *Estrategias y Prácticas Cualitativas de Investigación Social* (pp. 3-22). Madrid: Prentice-Hall.

Antaki, C., Billig, M., Edwards, D., Potter, J. (2003). Discourse Analysis Means Doing Analysis: A Critique Of Six Analytic Shortcomings. *Discourse Analysis Online*, *1(1)* [http://www.shu.ac.uk/daol/previous/v1/n1/index.htm]

Antaki, Ch. y Díaz, F. (2006). El análisis de la conversación y el estudio de la interacción social. En L. Iñíguez (aut.) *Análisis del discurso: Manual para las ciencias sociales* (pp. 129-142). Barcelona: Editorial UOC.

Bardin, L. (1967). *El análisis de contenido.* Madrid: Akal.

Bauer, M.W. y Gaskell, G. (2002). *Qualitative Researching. With text, image and sound*. London: Sage Publications.

Blanco, F. (2002) *El Cultivo de la Mente.* Madrid: Machado.

Blanco, F. and Sánchez-Criado, T. (2006). Speaking of Anorexia: a brief meditation on the notion of *mediation*. En Montero, I. (Ed.). *Current Research Trends in Private Speech. Proceedings of the First International Symposium on Self-Regulatory Functions of Language* (pp. 207-217). Madrid: Servicio de Publicaciones de la Universidad Autónoma de Madrid.

Blanco, F. y Montero, I. (2009). El sentido histórico del metodologismo en psicología: retórica antiretórica e hipertrofia normativa. Comunicación presentada en el X Congreso de Metodología de las Ciencias Sociales y de la Salud. Málaga, Septiembre de 2009.

Blommaert, J. (2004). *Discurse: a critical introduction.* New York: Cambridge University Press.

Burke, K. (1945/1984). *Grammar of motives*. New York, NY: Prentice Hall.

Delgado, J. (2006). Publicar sobre crisis y dogmas provoca encuentros. Y desencuentros. *Anuario de Psicología*, 37(*1-2*), 99-120.

Drew, P. (2003). Conversation analysis. En J.A. Smith: *Qualitative Psychology: A practical guide to research methods* (pp. 132-158). Londres: Sage.

Elliott, R., Fischer, C. T. y Rennie, D. L. (1999). Evolving guidelines for publication of qualitative research studies in psychology and related fields. *British Journal of Clinical Psychology*, 38, 215-229.

Foucault, M. de (1966/2006). *Las palabras y las cosas* (32ª edición en español). México: Siglo XXI.

Foucault, M. de (1975/2005). *Vigilar y castigar. Nacimiento de la prisión*. México: Siglo XXI.

Galindo, J. (Comp.) (2008). *Técnicas de investigación en sociedad, cultura y comunicación*. México: Pearson. Addison Wesley Longman.

Glaser, B.G. y Strauss, A. (1967). *The Discovery of Grounded Theory: Stretegies for Qualitative Research*. Chicago: Aldine.

Goffman, E. (1959/1993). *La presentación de la persona en la vida cotidiana*. Madrid: Amorrortu.

Gómez-Soriano, R. y Vianna, B. (2005). Eslabones encontrados: los grandes simios y el imaginario occidental. En Sánchez-Criado, T. y Blanco, F. (2005) (ed.) *AIBR. Revista de Antropología Iberoamericana. Ed. Electrónica, Número Especial* (Noviembre-Diciembre). *Cultura, Tecnociencia y Conocimiento: El reto constructivista de los Estudios de la Ciencia.*

González Rey, F.L. (2000). *Investigación cualitativa en Psicología.* México: Thomson Editores.

Gordo, A y Serrano, A (Coord.) (2008). *Estrategias y prácticas cualitativas de investigación de investigación social*. Madrid: Pearson-Prentice Hall.

Gutiérrez, J. y Delgado, J.M. (1994). *Métodos y técnicas*

cualitativas de investigación social. Madrid: Síntesis.

Haidar, J. (1998). Análisis del discurso. En: En: J. Galindo (Comp.): *Técnicas de investigación en Sociedad, cultura y comunicación* (pp. 117-164). México: Pearson.Addison Wesley Longman.

Hammersley, M. (2007). The issue of quality in qualitative research. *International Journal of Research y Method in Education,* 30(3), 287-305.

Heritage, J. (2004). Conversation analisis and institucional talk: analysing data. En D. Silverman, *Qualitative research. Theory, method and practice* (pp. 222-245). Londres: Sage (2ª Edición).

Jiménez, B., Blanco, F., Castro, J. y Morgade, M. (2001) La función de los mitos fundacionales en la promoción de una identidad disciplinar para la psicología. *Revista de Historia de la Psicología*, 22, (*3-4*), 297-310.

León, O. (2006) El monstruo de la razón produce sueños. *Revista de Historia de la Psicología*, 22, (*3-4*), 65-68.

Lewins, A. y Silver, C. (2006). *Choosing a CAQDAS Package*. CAQDAS Networking Proyect. http://caqdas.soc.surrey.ac.uk/

López, J.S. y Scandroglio, B. (2007). De la investigación a la intervención: la metodología cualitativa y su integración con la metodología cuantitativa (pp. 557-606). En A. Blanco y Rodríguez-Marín, J.: *Intervención psico-social.* Madrid: Prentice-Hall.

López, J.S., Martín, M.J., Martínez, J.M., Scandroglio, B. (2008). Family perception of organ donation process. Spanish Journal of Psychology, 11(1), 125-136.

López-Cabanas, M. y Chacón, F. (1999). Investigación-Acción Participativa. En M. López-Cabanas y F. Chacón: *Intervención psicosocial y servicios sociales* (pp. 163-182). Madrid: Síntesis.

Madigan, R., Johnson, S., y Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist, 50*(6), 428-436.

Madill, A., Jordan, A. y Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. *British Journal of Psychology,* 91, 1-20.

Manning, P. K. (1982). Analytic induction. En R.B. Smith y P.K. Manning (Eds.) *Handbook of Social Science methods: Qualitative methods*. Cambridge, MA: Ballinger.

Martín López, M.J. (2005). *Violencia juvenil exogrupal hacia la construcción de un modelo causal*. Madrid: Ministerio de Educación y Ciencia, Centro de Investigación y Documentación Educativa, D.L.

Miles, M.B. y Huberman, A.M. (1994). *Qualitative data analysis.* Thousand Oaks, CA.: Sage.

Miller, S.I. y Fredericks, M. (1987). The confirmation of hypotheses in qualitative research. *Methodika,* 1(1), pp. 25-40.

Montero, I. (2006) .Vino nuevo en odres viejos o la Metodología de un científico *deshonesto*. *Anuario de Psicología*, 37 (*1-2*), 75-80.

Piñuel, J.L. (2002). Epistemología, metodología y técnicas del análisis de contenido. *Estudios de Sociolingüística, 3,* 1, 1-42.

Plantin, C. (1998). *La argumentación*. Barcelona: Ariel.

Potter, J. y Wetherell, M. (1987). *Discourse and Social Psychology.* London: Sage.

Rasskin Gutman, I. (2007). Identidades en proceso de construcción: *¿Y tú cómo me ves?* En: Martin Rojo, L. y Mijares, L. (Ed.) *Voces del aula. Etnografías de la escuela multilingüe* (pp. 149-178). Madrid: CREADE (CIDE).

Ryan, G.W. y Bernard, R. H. (2000). Data management and Analysis. En: En N.K. Denzin y Y.S. Lincoln (Eds.), *The Handbook of qualitative reserach* (Segunda edición; pp. 769-802). Thousand Oaks, CA: Sage Publications.

Scandroglio, B. (2009). *Jóvenes, grupos y violencia. De las tribus urbanas a las bandas latinas.* Barcelona: Icaria.

Stiles, W. B. (1993). Quality control in qualitative research. *Clinical Psychology Review*, 13, 593-618.

Taylor, S. J. y Bodgan, R. (1984). *Introducción a los métodos cualitativos en investigación.* Barcelona. Paidós Básica.

Trinidad, A.; Carrero, V. y Soriano, R. M. (2006). *Teoría fundamentada "Grounded Theory". La construcción de la teoría a través del análisis interpretacional.* Cuadernos Metodológicos, 37. Madrid: Centro de Investigaciones Sociológicas.

Vallés, M.S. (2000). *Técnicas cualitativas de investigación social.* Madrid: Síntesis.

Vega, L. (2003). *Si de argumentar se trata.* Barcelona: Montesinos, D.L.

Velasco, H. y Díaz de Rada, A. (1997). *La lógica de la investigación etnográfica*, Madrid: Editorial Trotta.

Willig, C. (2003). Discourse analysis. En J.A. Smith: *Qualitative Psychology: A practical guide to research methods* pp. 159-183). Londres: Sage.

Wodak, R. (2001). *Methods of critical discourse analysis.* London: SAGE

Wolcott, H. (1999). *Ethnography: a way of seeing.* Walnut Creek, CA: Altamira Press.