Apuntes de Psicología
2002, Vol. 20, número 1, pp. 81-96.
ISSN 0213-3334

Colegio Oficial de Psicólogos
de Andalucía Occidental y
Universidad de Sevilla

Reflexiones críticas sobre la investigación en medición mediante tests en España

Antonio J. ROJAS TEJADA
Universidad de Almería

Resumen

En el presente artículo se presentan diferentes claves para entender por qué la investigación psicométrica en el campo de la medición mediante tests en España esté reflejando los últimos avances mundiales, y, sin embargo, está totalmente desvinculada de la práctica de la medición a gran escala. Para ello, se profundizará en la relación que mantienen la práctica de la medición mediante tests con la investigación en dicho campo. Partiendo de la idea de que las preocupaciones y cuestiones de investigación en Teoría de Test son similares en el España y en Estados Unidos, se hará una comparativa entre la práctica del sistema de medición mediante tests que se lleva a cabo en ambos países. En las conclusiones se comentan cuáles son los factores que están motivando la agenda de investigación española así como algunas ideas de futuro.

Palabras clave: tests, medición mediante tests, investigación en medición mediante tests, aplicación de la medición mediante tests.

Abstract

In the present paper, the relation between applied testing and testing research is analysed The aim is to point to different aspects that could explain why, although Spanish testing research reflects the latest trends in the field, this does not follow in applied contexts. Since test theory research is very similar in Spain and the USA, a comparative study of the practice of testing in both countries is carried out. In the conclusions, the reasons which conform the agenda of Spanish testing research community are commented, together with some ideas that could be useful for future developments.

Key words: tests, testing, testing research, applied testing.

Dirección del autor: Departamento de Ciencias Humanas y Sociales. Área de Metodología de las Ciencias del Comportamiento. Universidad de Almería. Carretera de Sacramento, s/n. La Cañada de San Urbano. 04120 Almería. *Correo electrónico:* arojas@ual.es

El presente trabajo se inició cuando el autor realizaba una estancia en el *Educational Testing Service* (ETS), de Princeton, en New Jersey, EE.UU. Sin las valiosas ideas y reflexiones sugeridas por Isaac I. Bejar (ETS), este trabajo no habría tomado cuerpo. María José Navas (UNED) leyó pacientemente un borrador y realizó unos comentarios muy acertados que se incorporaron al texto. Juan S. Fernández Prados (Universidad de Almería) también ha contribuido de forma importante a lo que hoy ve la luz.

La investigación psicométrica[1] española está en auge y entre las mejores del mundo, al menos en lo que a Teoría de Tests de refiere. Ronald K. Hambleton (una reconocida autoridad mundial en psicometría) declaró públicamente, en la conferencia de clausura del VI Congreso Nacional de Metodología, que España ocupaba el tercer puesto en el *ranking* mundial, tras Estados Unidos y Holanda (aunque, quizás, desbordando optimismo).

Los textos y artículos que se publican en España sobre Teoría de Tests reflejan prácticamente las mismas cuestiones e inquietudes de investigación que aparecen en los textos y artículos de las revistas con mayor reconocimiento sobre el tema (por ejemplo *Applied Psychological Measurement*, *Journal of Educational Measurement*, *Journal of Measurement in Education*, *Education and Psychological Measurement*, *Psycometrika*, etc.). Existe entre nosotros el mismo interés sobre la temática general de la Teoría, Modelos y Aplicaciones de la Respuesta a los Items (TRI) que en Estados Unidos. Temas más específicos como el estudio de la Dimensionalidad de los tests, los Bancos de Items (BI), el Funcionamiento Diferencial de los Items (DIF) o los Tests Adaptativos Informatizados (TAI), están siendo las líneas de interés de los investigadores españoles. Todo ello, obviamente, a menor escala.

Este interés también se extiende a los ya clásicos estudios sobre fiabilidad y validación que se realizan para multitud de pruebas que, generalmente, no llegan a estandarizarse, ya que en muchas ocasiones tienen una finalidad exclusivamente de investigación, pero donde también se incluyen las últimas novedades en estos campos (por ejemplo análisis factorial confirmatorio, modelos de ecuaciones estructurales, etc.).

Para corroborar esto, tan solo hay que revisar los últimos números de las revistas españolas que tienen secciones que incluyen artículos sobre investigaciones psicométricas, por ejemplo *Psicológica* o *Psicothema*.

En 1999, el volumen 20 de *Psicológica* (compuesta por tres números), publicó ocho artículos en la sección de *Metodología*. Entre ellos dos estaban dedicados a estudios sobre modelos de Teoría de Respuesta a los Items (TRI), dos a DIF y uno a los TAIs. En el mismo año, *Psicothema* (con cuatro números), publicó en su sección de *Sotfware, Instrumentación y Metodología* un total de 12 artículos. De ellos, uno estaba dedicado al estudio de modelos de TRI, dos se dedicaban a la dimensionalidad, uno al DIF y otros tres a estudios de fiabilidad o validez de tests. A esta lista habría que añadir otras publicaciones recientes que reflejan los nuevos desarrollos psicométricos (por ejemplo Muñiz, 1996) o incluso temas mucho más específicos como los TAIs, donde destacan el monográfico de la *Revista Electrónica de Investigación y Evaluación Educativa* (RELIEVE, 1998), el texto de Olea, Ponsoda y

1 En general, podemos decir que el papel que juega la psicometría en el contexto psicológico y educativo es proporcionar soluciones a los problemas de medición. La psicometría está encargada de la elaboración de teorías y modelos formales así como de establecer las pautas para construir y aplicar métodos e instrumentos para la medición de variables psicológicas (por ejemplo Olea, Ponsoda y Prieto, 1999). Dentro del amplio campo de la Psicometría, la Teoría de los Tests se ha ocupado de formalizar mediante modelos matemáticos las relaciones entre las respuestas de las personas a los items de un test y el grado en que esas personas poseen la variable que se mide con dichos items.

Prieto (1999), y el monográfico que publicó *Psicológica* en el año 2000. También, cada vez más, se ven nombres de españoles publicando en las revistas de mayor impacto americanas, como, por ejemplo, el trabajo de Ferrando (1999) en *Applied Psychological Measurement* o el de Lorenzo (2000) en *Psicometrika.*

Además de las publicaciones, estos temas punteros se están dejando ver no solo en programas de tercer ciclo (por ejemplo universidades como la Autónoma y la Complutense de Madrid, o las de Oviedo, Barcelona, Valencia, Granada, UNED, etc. tienen cursos sobre TRI, DIF, BI, TAI, etc.), sino en los curricula académicos de formación de Licenciados en Psicología, donde la asignatura troncal de Psicometría incluye estos mismos temas. Y todo lo expuesto es sin ánimo de ser exhaustivos. Como vemos, la idea de Hambleton no estaba vacía de contenido.

Esto podría hacernos pensar que la práctica de la medición mediante tests que se hace actualmente en España es paralela al avanzado desarrollo de la investigación psicométrica, de forma que serían las demandas de medición requeridas por la práctica las que crearían las necesidades de dicha investigación, al tiempo que la práctica de la medición iría reflejando los avances de la investigación. Tal y como ocurre, por ejemplo, en Estados Unidos.

Este artículo pretende presentar diferentes claves para entender cuáles han sido los factores que hacen que la investigación psicométrica en el campo de la medición mediante tests en España esté reflejando los últimos avances mundiales, con lo que se profundizará en la relación que mantienen la práctica de la medición mediante tests con la investigación en dicho campo. Para ello, y partiendo de la idea de que las preocupaciones y cuestiones de investigación en Teoría de Test son similares en el España y en Estados Unidos, se hará una comparativa entre la práctica del sistema de medición mediante tests que se lleva a cabo en ambos países, donde se pretende analizar qué está motivando la agenda de investigación en Teoría de Tests tanto en Estados Unidos como en España.

Visión general de la medición mediante tests en Estados Unidos

La investigación en Teoría de Tests en Estados Unidos ha estado ligada a la medición mediante tests estandarizados. Los tests entandarizados imponen ciertos controles en las condiciones de medidas, que son iguales para todos los examinados (por ejemplo forma de aplicación, normas de puntuación), donde se garantiza que la puntuación asignada a un sujeto es independiente de la persona que asigna dicha puntuación (por ejemplo, Crocker y Algina, 1986). Una de las consecuencias de ello es que los tests estandarizados son (y necesitan serlo) mucho más fiables que, por ejemplo, los tests elaborados por profesores (por ejemplo, Hopkins, 1998).

Los distintos ámbitos de aplicación de tests

En Estados Unidos se administran muchos millones de tests estandarizados al año. Varios son los campos donde se utilizan los tests, pero, en general, se circunscriben al ámbito educativo, al ámbito profesional y a los diferentes ámbitos de la aplicación de la psicología (por ejemplo, evaluación clínica, selección de personal, etc.). Pero, sin duda, la medición mediante tests en Estados Unidos ha estado fuertemente vinculada a la medición educativa.

Tests en el ámbito educativo

En el *ámbito educativo* es en el que mayor volumen de tests estandarizados se producen y se utilizan en Estados Unidos. A su vez, podríamos clasificar los tests estandarizados educativos en tres grandes áreas: tests de rendimiento (*achievement tests*), tests de admisiones (*admisions tests*) y tests de clasificaciones (*placement tests*). Aunque también podríamos incluir aquí las encuestas educativas (*educational surveys*).

Los resultados que proporcionan los *tests de rendimiento* (*achievement tests*) son usados por las instituciones, generalmente en educación primaria y secundaria, para tomar decisiones acerca del futuro educativo de los estudiantes. Aunque este tipo de tests estandarizados son un fenómeno de los años 70, han sido adoptados por cientos de instituciones educativas y docenas de estados en Estados Unidos (Jaeger, 1989).

Un segundo, y quizás el mayor de los grupos entre los tests estandarizados educativos, hace referencia a los *tests de admisión* (*admissions tests*). Estos tests son usados con el fin de que las instituciones de educación superior (universidades, escuelas de postgrado e institutos de investigación) decidan si admitir o no a los estudiantes que solicitan su ingreso en ellas. Hasta 1926, la mayoría de estas instituciones usaban para las admisiones bien las relaciones familiares, o bien tests no estandarizados y desarrollados localmente para evaluar las destrezas de los aspirantes. En respuesta a la necesidad de procedimientos más eficientes y estandarizados, el *College Entrance Examination Board* introdujo, en 1926, la batería de pruebas conocida como el *Scholastic Aptitude Test* o SAT (por ejemplo Wainer, 1990; Whitney, 1989). También en 1959 se fundó el *American College Testing Program*, organización que sigue realizando el test ACT (*American College Test*), utilizado amplia y conjuntamente con el SAT para las admisiones en Estados Unidos. También, para la admisión en Centros de Postgrado, se elaboraron pruebas con el ánimo de ser extendidas a todas las instituciones. El ejemplo más importante de éstos es el GRE (*Graduate Record Examination*). Pero, generalmente, los programas de postgrado desarrollaron tests estandarizados específicos, por ejemplo, para estudios de empresariales se creó el GMAT (*Graduate Management Admisions Test*), para estudios de derecho el LSAT (*Law School Admission Test*) y para estudios de medicina el MCAT (*Medical College Admision Test*).

Los *tests de clasificación u ordenación* (*placement tests*) han sido utilizados durante muchos años para poder asignar a los nuevos estudiantes a los cursos que se ajusten a su experiencia y sus destrezas educativas. Generalmente, han sido aplicados por las instituciones educativas después de la admisión. Aunque se han creado tests que han tenido gran repercusión, como el *AP* (*Advanced Placement*) desarrollado por el *Educational Testing Service* (ETS) o el sistema *ACCUPLACER*, un servicio que proporciona el *College Board*, que consiste en un programa de este tipo de tests, con la novedad de ser tests adaptativos informatizados y administrados por internet (College Board, 2000), los tests de clasificación u ordenación han sido desarrollados localmente y dependientes de la instituciones educativas interesadas en ellos (Whitney, 1989).

Y, por último, habría que mencionar las *encuestas educativas* (*educational*

surveys), que también son tests estandarizados que pretenden obtener información de cómo se está llevando a cabo el proceso enseñanza-aprendizaje en el sistema de educación público, con el fin de mejorar los programas instruccionales o la formación de los profesores. Un ejemplo de este tipo de encuestas sería el *National Assessment of Educational Progress* (NAEP). En estas encuestas se mide la ejecución de los estudiantes en diferentes cursos (a los 9, 13 y 17 años) y en diferentes áreas (lectura, escritura, matemáticas y ciencias) (por ejemplo, McRury, Nagy y Traub, 1991).

La mayoría de los tests educativos estandarizados son administrados a gran escala con propósitos institucionales (por ejemplo, Bennett, 1997). Como podemos imaginar, el ámbito educativo supone el grueso de la medición mediante tests en Estados Unidos.

Tests en el ámbito profesional

Un segundo campo lo ocupan los tests que se utilizan para el *ámbito profesional*. En este sentido cabe mencionar dos grandes tipos de pruebas: los tests de licencias y los tests de certificaciones.

Los *tests de licencias profesionales* (*professional licensing tests*) son exigidos por el Gobierno Federal o por los estatales para poder ejercer una determinada profesión. Un ejemplo de este tipo de test es el *NCLEX (National Council Licensure EXamination for Registered Nurses*), que es necesario para colegiarse como enfermero/a en cualquier estado de Estados Unidos. Con este test se pretende comprobar el grado de competencia en la ejecución de tareas clínicas requeridas para el trabajo de enfermero/a (ETS, 1994).

Los *tests de certificaciones* (*certification tests*) proporcionan el reconocimiento de que, tras un determinado proceso de instrucción, se ha alcanzado un estándar de excelencia. Este tipo de tests está, generalmente, vinculado a cursos de formación técnicos. Un ejemplo de este tipo de tests es el que expide el *American Board of Medical Specialities* (ABMS), para sus 24 especialidades. El resultado de la medida obtenida mediante estas pruebas nos indica si una persona es considerada o no experta en una determinada especialidad, respecto a unos estándares de ejecución previamente decididos por un Consejo, que en caso de los médicos son el *American Board of Medical Specialties* (ABMS) y la *American Medical Association* (AMA).

A caballo entre el ámbito educativo y profesional están los *tests de preparación para el trabajo* (*work keys tests*). Fundamentalmente sirven para conocer qué habilidades se deben perfeccionar para conseguir buenos trabajos. Estos tests son muy útiles en empresas, en programas de entrenamiento y en sistemas educativos para conocer la preparación que tienen los evaluados para desarrollar trabajos o estudios específicos. La finalidad es asegurar que los aspirantes lleguen al mundo laboral o académico con la preparación necesaria para desarrollar adecuadamente cualquier tarea. Las preguntas que se utilizan se asemejan a los problemas que se encuentran en la vida diaria en el contexto del trabajo o en el académico. Entre estos tests destacan los elaborados por la compañía ACT (*American College Testing*), que evalúan diferentes habilidades como: matemáticas aplicadas, comprensión oral y escrita, escritura, uso de información, trabajo en equipo, observación, uso de tecnología aplicada, etc.

Para finalizar, los tests estandarizados también son utilizados en los diferentes ámbitos de la aplicación de la psicología (por ejemplo, diagnóstico clínico, selección de personal, etc.). En este sentido, hay que mencionar que, si bien existe una gran cantidad de tests utilizados con diferentes fines en la psicología aplicada (tests como el *MMPI-2* o el *Strong Interest Inventory* son ampliamente utilizados en contextos clínicos o de orientación profesional, respectivamente), no podemos considerar que este tipo de uso suponga algo comparable a la medición mediante tests a gran escala que se hace con los tests educativos.

Las organizaciones que elaboran tests

De todo lo señalado en el apartado anterior se puede deducir que la medición mediante tests en Estados Unidos supone también un gran negocio. Existen grandes compañías que se dedican a proporcionar los servicios de medición para dar cobertura a los ámbitos de aplicación citados. Entre las más grandes podemos citar ETS (*Educational Testing Service*) o ACT (*American College Testing*), aunque la *Association of Test Publishers* (*ATP*) proporciona una lista de 53 organizaciones dedicadas a ello solo en Estados Unidos (ATP, 1998).

A título informativo, y para hacernos una idea del volumen de negocio que se mueve en Estados Unidos referente a este tema, podemos dar algunas cifras: el ETS, creado en 1947, es la institución no lucrativa más grande en Estados Unidos (2.100 empleados) dedicada a medición educativa. En el curso 1998-99, ETS aplicó aproximadamente un total de 11.213.000 de tests, contando solo los más usuales (ETS, 1999). ACT administró unos 6.000.000 de tests en el mismo curso. ACT fue fundada en 1959 y actualmente sus empleados son 1200 (ACT, 2000).

Las cifras sobre el número de aplicaciones de los tests citados en el apartado anterior son las siguientes: el ACT *Assessment* lo realizaron cerca de un millón de estudiantes en 1997 (ACT, 1997), el SAT tuvo 2.468.600 aplicaciones durante el curso 1998-1999, el AP 1.153.100, el GRE y del GMAT, en el mismo curso, se administraron 291.000 y 210.000 respectivamente, y el NAEP lo cumplimentaron 680.000 estudiantes (ETS, 1999). Como vemos son cifras nada despreciables.

La investigación en medición mediante tests

Que duda cabe que la amplia preocupación por mejorar la calidad de las medidas obtenidas mediante tests ha sido (y será) la gran impulsora de los cambios que se han visto en la investigación en medición mediante tests, pero podemos decir, sin temor a equivocarnos, que estos cambios casi siempre han estado motivados por, al menos, dos factores más: legales y económicos. En Estados Unidos, la medición mediante tests está en continua revisión. Por un lado, por la presión social que impone ciertos controles y exige ciertas garantías a estas mediciones, que acaban, generalmente, generando legislación al respecto (Linn, 1989). Y, por otro lado, por la necesidad que tienen estas empresas de aprovechar eficientemente sus recursos. Ilustremos lo dicho con algunos ejemplos.

Debido a que los tests de rendimiento en Estados Unidos son usados para tomar decisiones educativas muy importantes, como dar un diploma de educación de secundaria, repetir curso o ser asignado para

clases de recuperación, no es sorprendente pensar que existen muchos casos en que los programas de medición mediante tests de rendimiento hayan sido llevados a los tribunales de justicia (ver, por ejemplo, *Educational Measurement: Issues and Practice, Vol.2(4)*). Tal es así, que se ha legislado acerca de lo que los tests de rendimiento intentan medir. Así, se ha constituido como principio que los profesionales o instituciones que desarrollen programas de medición mediante tests de rendimiento deben demostrar que los estudiantes han tenido la oportunidad de aprender los contenidos y materiales sobre los que van a ser evaluados (Jaeger, 1989). Esta necesidad ha provocado la realización de numerosísimos estudios sobre la validación de contenido. Estudios que han debido realizar los agentes que se dedican a elaborar este tipo de tests. De hecho, la mayoría de los programas de tests de admisiones ofrecen un conjunto de estudios de validación, donde se exponen los resultados de los mismos, y donde, incluso, se ofrece ayuda para interpretar la validez del uso de las admisiones realizadas con sus tests (Jaeger, 1989). Muchas de las investigaciones de la sección de estudios de validación de la revista *Educational and Psychological Measurement* son un claro ejemplo de ellos.

Otro ejemplo tiene que ver con el sesgo de los tests. El origen multicultural de la sociedad estadounidense (por ejemplo, Villegas, 1992) ha propiciado que se haga un especial hincapié en cómo cada colectivo responde a los tests. Por ejemplo, el *National Center for Fair and Open Testing* (*FairTest*) es una organización para la defensa de un uso justo de los tests, que trabaja para evitar el abuso, el mal uso y los defectos en los tests estandarizados (por ejemplo, sesgos raciales, de clase, culturales o de género), y para asegurar que la evaluación de los estudiantes y trabajadores sea justa, abierta y adecuada (FairTest, 2000), o el *Consortium for Equity in Standards and Testing* (CEST, 1998) con objetivos similares. Organizaciones como las citadas han provocado que las instituciones que elaboran tests, hayan incrementado sus esfuerzos por remover estereotipos y sesgos de contenido, incluyendo en las etapas de construcción del test estudios de funcionamiento diferencial de los ítems, donde se detecten y corrijan estos defectos. Y, sin duda, la atención prestada al análisis de ítems, por posibles funcionamientos diferenciales (sesgos), ha desembocado en muchos avances metodológicos sobre el tema (por ejemplo, Holland y Wainer, 1993).

También se ha invertido mucho en investigación psicométrica con el fin de descubrir cada vez más modelos y métodos que reduzcan los costes de la medición. Así no es de extrañar que actualmente se estén investigando nuevas formas de medición basadas en el ordenador. Por ejemplo, podemos hablar de las ventajas económicas (por ejemplo, Bennett, 1998), además de las técnicas de precisión y eficiencia en la medida (por ejemplo, Weiss y Schleisman, 1999), que suponen el uso los tests adaptativos informatizados (por ejemplo, Meijer y Nering, 1999; Van der Linden, 1999): estandarización de todos los pasos en el proceso de medición, tests más cortos, corrección automática e inmediata de respuestas, generación automática de ítems, evaluación mediante internet, etc. O, citando otro ejemplo, los múltiples estudios sobre los métodos de detección de Funcionamiento Diferencial de los Items, que se dieron hace unos años (y que aún continúan), para llegar a tener procedimientos

eficaces que se apliquen con facilidad y garantías en el proceso de construcción de tests.

Estos aspectos de investigación son tan importantes que las grandes instituciones que se dedican a la elaboración de tests tienen divisiones dedicadas exclusivamente a investigación (por ejemplo, ACT *Research,* ETS *Research*), donde se trabaja solo y exclusivamente para solucionar los problemas que surgen en la medición con los tests que desarrollan ellas mismas. Estos centros se han convertido, de hecho, en una de las principales fuentes generadoras de informes técnicos de investigación y publicaciones especializadas sobre medición.

Estos son ejemplos de cómo la aplicación de tests estandarizados ha necesitado de investigación psicométrica para garantizar un uso eficiente, adecuado y justo de los tests. Podemos decir que la demanda de aplicación ha creado una basta línea de investigación en Teoría de Tests. Y, a su vez, que los avances de investigación tienen un basto campo de aplicación.

Todo lo anterior puede proporcionarnos diferentes claves para entender la gruesa agenda de investigación y la cantidad de publicaciones que se llevan a cabo sobre Teoría de Tests en Estados Unidos, y cómo estas investigaciones han estado muy ligadas al contexto educativo.

La medición mediante tests en España

Para hablar sobre medición mediante tests en España habría que mencionar, en primer lugar, que no existe ningún ámbito de aplicación donde se utilicen tests estandarizados a gran escala. En España solo se aplican tests estandarizados en los diferentes *ámbitos de la aplicación de la psicología.* Los objetivos fundamentales de diagnóstico clínico, de selección de personal y de orientación educativa, etc. que se realizan en estas áreas no requieren grandes cantidades de tests, o al menos, no existe una medición mediante tests a gran escala tal y como hemos visto en el caso de Estados Unidos.

Pero esto no quiere decir que en el España no haya ámbitos donde los tests estandarizados pudieran y debieran utilizarse, dada las ventajas que proporcionan. De hecho, en España se hace medición (y evaluación) a gran escala pero con tests y exámenes elaborados por profesores y para cada ocasión, con todas las críticas e inconvenientes que ello supone desde el punto de vista psicométrico.

Los distintos ámbitos de aplicación de tests

Los ámbitos donde se aplican mediciones a gran escala son, fundamentalmente: *pruebas de acceso a la universidad* (selectividad), *pruebas de acceso a empleos públicos* (por ejemplo profesores de enseñanzas primarias y secundarias, en el ámbito de la enseñanza; militares profesionales, en el ámbito profesional), *pruebas de acceso a la formación sanitaria especializada* (por ejemplo pruebas MIR para médicos, FIR para farmacéuticos, BIR para biólogos, PIR para psicólogos, etc.), etc.

Los *exámenes de acceso a la universidad* (selectividad) son pruebas de rendimiento elaboradas para medir los conocimientos necesarios que debe de tener un alumno para acceder a la educación universitaria. Su construcción corre a cargo de una comisión integrada por profesores universitarios (uno por cada materia) que actúan como coordinadores con los profeso-

res de enseñanza secundaria. Esta comisión marca las directrices que integran los temarios del curso. Coordinadores y profesores de enseñanza secundaria elaboran la estructura del examen (por ejemplo formato de los ítems, tiempo requerido, normas de corrección, etc.). La administración y corrección del examen lo efectúan los distintos tribunales seleccionados por distritos universitarios.

Estos exámenes son pruebas no estandarizadas. Los ítems que componen el test no han estado sometidos a ningún control psicométrico (o al menos no consta ningún informe de ello). En este sentido, nada garantiza que una persona que tome un examen en el territorio español (suponiendo que el examen sea el mismo para todos), obtenga la misma puntuación si ese examen es corregido en dos puntos diferentes de la misma geografía. No existen ni estándares de ejecución, ni estudios de fiabilidad, ni estudios de validación al respecto.

No obstante, estos exámenes tienen gran trascendencia en la vida de los estudiantes, ya que a partir de la nota obtenida se decide su futura carrera universitaria. La puntuación media ponderada de la selectividad (40% la calificación global de la prueba) y la nota media del expediente académico en la enseñanza secundaria, que supondrá el 60%, determinará la futura admisión de cada estudiante a la Universidad (MEC, 1999). Cada carrera universitaria, en cada universidad, seleccionará para su acceso a las personas con mejores notas, dependiendo de su oferta y demanda. Estas pruebas de selectividad están siendo actualmente revisadas, y parece que en un futuro próximo serán eliminadas en favor de pruebas de acceso que cada universidad determinará.

Las *pruebas de acceso a empleos públicos* son otro ejemplo del uso masivo de medición mediante tests no estandarizados. Con las pruebas de acceso a empleos públicos en el *ámbito de la enseñanza* (por ejemplo profesores de enseñanzas primarias y secundarias), ocurre algo parecido a lo que ocurre con la selectividad. Aunque puede haber leves variaciones entre comunidades autónomas, en general, en la parte de concurso, estas pruebas (una prueba escrita de contenido de carácter teórico y otra de carácter práctico y una prueba oral) son sobre los conocimientos específicos de los candidatos necesarios para impartir docencia, su aptitud pedagógica y su dominio de las técnicas necesarias para el ejercicio docente (MEC, 1999). En este caso también existen distintos tribunales provinciales que califican estas pruebas por separado. La selección se hace entre los aspirantes que han superado las tres pruebas de la oposición, sumando a la media aritmética de las tres, la puntuación que hayan obtenido en el concurso de méritos, juzgados mediante unos baremos por el mismo tribunal.

Sobre las pruebas de acceso para otros sectores de la administración pública (por ejemplo, militares profesionales) aun hay menos información. Los aspirantes deben superar una primera prueba de evaluación personalizada, donde se evalúa la capacidad '*para obtener el rendimiento académico y profesional militar requerido, mediante pruebas que miden factores intelectuales y aptitudinales. La aplicación de estas pruebas será realizada por el personal del Cuerpo Militar de Sanidad (Psicología)*' (BOE, 2000, 5816). Para hacernos una idea de las cifras, para el año 2000 estos exámenes se realizarán para cubrir 17.500 plazas.

Y así podríamos seguir para otros puestos de la administración pública: jueces, fiscales, fuerzas de seguridad del estado, etc.

Las *pruebas de acceso a la Formación Sanitaria Especializada* (por ejemplo MIR, FIR, BIR, PIR, etc.) son de ámbito nacional, y están compuestas por 250 ítems con cinco alternativas de respuestas, cuyos contenidos versan sobre las áreas de enseñanza comprendidas en las Licenciaturas respectivas. La puntuación final es una media ponderada de las puntuaciones al test (75%) y la valoración de los méritos académicos (25%) de los aspirantes. La adjudicación de las plazas se efectúa siguiendo el orden de mayor a menor puntuación total individual de cada aspirante (MSC, 2000). Estas pruebas no han estado sometidas a ningún control psicométrico (o al menos no consta ningún informe sobre ello).

Además, a esta lista habría que añadir *las encuestas educativas*, que recientemente ha llevado a cabo el Instituto Nacional de Calidad y Evaluación (INCE), con la finalidad de proporcionar información relevante a las administraciones educativas, a los órganos de participación institucional, a los agentes implicados en el proceso educativo (familias, alumnos, profesores y otras entidades), así como a los ciudadanos en general, sobre el grado de calidad que el sistema educativo ha alcanzado en un determinado momento de su desarrollo (INCE, 1998). Así, podemos encontrar trabajos que pretenden esta finalidad con distintas especialidades como, por ejemplo, el de Gil y Alabau (1997) para educación física o el de Pérez, García-Gallo y Gil (1995) para la lengua inglesa, o incluso algunos intentos de este tipo de evaluación a nivel internacional. En estos estudios suelen existir datos sobre el modelo de medida empleado para la calibración de los ítems y las mediciones de los sujetos, pero en ningún caso se informa sobre la fiabilidad de las puntuaciones o sobre estudios de validación.

Un ejemplo más del uso de medición mediante tests en España lo constituyen los tests para obtener el *permiso de conducción*. Son tests con ítems de elección múltiple, donde existe un estándar de ejecución, por debajo del cual no se puede pasar a la prueba práctica (por ejemplo, tener como máximo cinco fallos en 40 ítems). Estos tests son de aplicación nacional, si bien cada Dirección Provincial de Tráfico compone sus propios exámenes. Tampoco existen informes sobre las propiedades psicométricas de los mismos.

Es difícil demostrar que no existen trabajos sobre la calidad de las medidas a gran escala que se hace en España, pero podemos decir, al menos, que si esa información existe, no es divulgada. La información sobre temas como la elaboración de estándares de puntuación, fiabilidad o validación no acompaña a estas mediciones, lo cual nos hace ver la mínima importancia que se le otorga a ello. En cambio, esta información sí aparece en las publicaciones especializadas cuando se utilizan tests con fines de investigación.

En general, podemos afirmar que la estandarización de los tests no ha sido una preocupación ni de los profesionales encargados de hacerlos, ni de la sociedad donde se aplican, ya que no existe demanda alguna. Así pues, aunque es cierto que para la realización de la medición mediante tests a gran escala que se hace en España existen equipos de profesionales cualificados, también lo es la inexistencia de información psicométrica sobre los mismos.

Las organizaciones que elaboran tests

Como vemos, el panorama de medición mediante tests no se asemeja mucho a lo que ocurre en Estados Unidos. Si bien existe medición mediante tests a gran escala, las pruebas que se utilizan no están estandarizadas y son elaboradas por tribunales o por grupos de expertos para cada ocasión y para cada territorio. En este sentido, los tests estandarizados en España no suponen un gran negocio. Y fiel reflejo de ello es la carencia de organizaciones dedicadas a la elaboración y administración de tests.

La mayor empresa dedicada a la venta y distribución (y escasamente a la elaboración) de tests y pruebas de evaluación psicológica en España es TEA Ediciones. Para hacernos una idea, esta empresa cuenta con una plantilla de 40 personas (TEA, 1999). También podemos citar otras compañías tales como el grupo ALBOR-COHS o EOS, pero sus secciones sobre tests son más pequeñas.

La investigación en la medición mediante tests

Lo curioso de todo el asunto es que dada la gran cantidad de medición mediante tests que se lleva a cabo en España, no existan movimientos ni reivindicaciones a favor de una evaluación justa y estandarizada, que evite los problemas de subjetividad que implican la evaluación de preguntas abiertas por tribunales muy diferentes. Con este panorama se podría decir que, en España, la medición mediante tests estandarizados está por venir.

Sin embargo, no se puede decir esto mismo de la investigación sobre Teoría de Tests, ya que la investigación que se hace en este campo no se corresponde con la aplicación que se hace en el Estado Español. Cabría pensar, por tanto, que el panorama de la investigación psicométrica en España es bien reducido. Pero, como hemos comentado anteriormente, nos encontramos con la paradoja de que la investigación psicométrica en España es una de las más productivas y avanzadas a escala mundial.

Ante este panorama, habría que plantearse qué factores pueden dar cuenta de que la investigación en Teoría de Tests en España, que se fundamenta en aplicaciones de tests estandarizados a gran escala (y más aún si nos referimos a la Teoría de Respuesta a los Items que requiere para su aplicación grandes cantidades de datos), esté tan avanzada, cuando en nuestro país no existen ámbitos donde aplicar y mejorar dicha investigación.

La primera consideración a tener en cuenta es que la investigación psicométrica en España esta reducida a la investigación que se hace en las Universidades. En general, la investigación psicométrica que se hace en la universidad española está influenciada por múltiples factores, entre las que cabe indicar la presión por publicar en temas que se consideran punteros, con el fin de consolidar curriculum aceptable para poder promocionar. Así, los investigadores universitarios beben de las principales publicaciones a nivel internacional. Es decir, leen las revistas sobre medición psicológica y educativa que proceden de Estados Unidos. En este sentido, hacen suyos los problemas de medición que se dan en Estados Unidos, e investigan cuestiones que en España no tienen ninguna repercusión práctica.

Podemos retomar los ejemplos del Funcionamiento Diferencial de los Items (DIF) y de los Tests Adaptativos Informatizados (TAI). En España se están publi-

cando trabajos de investigación sobre DIF (por ejemplo, Elosúa y López, 1999; Padilla, González y Pérez, 1998; Prieto, Barbero y San Luís, 1999) o sobre TAIs (por ejemplo Olea, Ponsoda y Prieto, 1999; Ponsoda, Olea y Revuelta, 1994; Ponsoda, Wise, Olea y Revuelta, 1997; o el monográfico que le dedica *Psicológica* en el año 2000) similares en calidad y alcance que los que se realizan en Estados Unidos. Pero con una diferencia fundamental, en España no existe ningún ámbito donde aplicarlo. O, mejor expresado, si existen ámbitos donde aplicarlos pero no se ve la necesidad social de ello.

Si como hemos comentado, en nuestro país ni siquiera se ha planteado nunca la necesidad de realizar un test estandarizado para evaluar los contenidos de la educación secundaria, si a las pruebas de selectividad nunca se le han realizado ni siquiera un estudio sobre fiabilidad, si los estudios de validación para las pruebas de acceso a empleos públicos o a la formación sanitaria especializada no existen. Qué sentido tiene que la investigación universitaria en Teoría de Tests esté llevando a cabo rigurosísimos estudios sobre que métodos de detección de Funcionamiento Diferencial de los Items tienen mejores resultados, si no existe ningún movimiento social que vele por los intereses de una medición justa y equitativa. O qué sentido tiene establecer nuevos desarrollos de Tests Informatizados cuando todavía no nos han llegado los problemas económicos que supone la medición mediante tests estandarizados a gran escala.

Conclusiones

Como hemos comentado, si bien los problemas de investigación en España y en Estados Unidos son idénticos, la práctica de la medición mediante tests difiere bastante entre uno y otro. Los motivos de ello pueden ser múltiples, pero el más obvio es que la investigación en este campo se lleva, casi exclusivamente, en las universidades. Y las universidades marcan sus agendas de investigación en función de las revistas de mayor impacto científico, es decir, en materia de medición mediante tests, de las revistas norteamericanas. Los investigadores españoles se ven abocados a publicar sobre los mismos temas y con la aspiración de hacerlo en las mismas revistas.

Parece obvio que el futuro de la medición pasa por que los adelantos que se llevan a cabo en la investigación mediante tests en España salten de las universidades a los programas de medición a gran escala.

En este salto, como en todo proceso de institucionalización, habría que atender a tres factores que caminan independientemente: inquietud científica, demanda social y la institucionalización en sí misma. Para que un hecho determinado llegue a institucionalizarse debe haberse comprobado su utilidad social, donde los grupos sociales deben demandarlo como un derecho y una solución a sus problemas, siempre teniendo en cuenta la eficacia demostrada en la investigación científica.

En Estados Unidos es obvio que la transferencia de la investigación psicométrica está llegando a la aplicación práctica de la medición mediante tests, donde los grupos sociales juegan un gran papel en las demandas de sistemas de medidas cada vez más efectivos, adecuados y justos. En ese país la medición mediante tests está totalmente institucionalizada.

En España, sin embargo, la investigación no logra transferir sus conocimientos al ámbito de la aplicación, primero, porque

no existe ningún tipo de demanda social respecto a la medición a gran escala, y, segundo, porque la investigación no surge de la necesidad de solucionar los problemas de aplicación que ocurren en este país. A esto, se le podrían unir cuestiones como la escasa conciencia de los administradores públicos (políticos) acerca de la necesidad de profesionalización psicométrica, la todavía escasa conexión entre universidad y empresa, e incluso la escasa concienciación y formación, que en materia psicométrica, tienen los colegiados profesionales de psicología (por ejemplo, Prieto y Muñiz, 2001; Muñiz y cols, 1999; Muñiz y Fernández Hermida, 2000).

Parece que el futuro de la investigación psicométrica en España pasa por estar cada vez más presente en los ámbitos de investigación mundial. Esto va a suponer que cada vez habrá, dentro del ámbito de la medición mediante tests, una menor conexión entre la práctica (atrasada) y la investigación (demasiado adelantada). A su vez, otros países que sí tienen institucionalizada la medición mediante tests se favorecerán de la investigación que se hace en nuestro país.

En este sentido, el papel del investigador en Teoría de Tests también debe ser la de divulgador y promotor para que sus investigaciones se vean reflejadas en las aplicaciones que se realizan. Quizás los propios investigadores deban formar un grupo de presión social (y alentar a otros) con el fin de exigir unas garantías de calidad mínimas que debe acompañar a toda medición mediante tests. A esto debería unirse una mayor actividad social de las facultades de psicología, fomentando la transferencia de la investigación psicométrica a las distintas instituciones sociales que hacen uso de los tests, mediante la creación de unidades o institutos de medición aplicada. Y, sin duda, en este proceso de reivindicación debe tomar parte el Colegio Oficial de Psicólogos, máximo órgano de expresión de la profesión. De esta forma se podría presionar a los responsables políticos que tienen competencia en los sistemas masivos de evaluación.

Esto, o ver cómo se sigue incrementando esa desconexión entre investigación y *praxis*, o, ¿por qué no?, cambiar nuestros programas de investigación para ajustarlos a lo que se demanda.

Este artículo tiene el ánimo de iniciar una reflexión sobre la investigación psicométrica en España. En este sentido, y como consecuencia de todo lo expuesto, quizás deberíamos replantear el *qué*, el *porqué* y el *para qué* de la investigación en Teoría de Tests, aunando los esfuerzos de los profesionales dedicados a la medición mediante tests o que hacen uso de ella.

Referencias

ACT (2000). *About ACT. History of ACT.* [on line]. Obtenido el 31/01/2000 en URL: http://www.act.org/aboutACT/Acthist.html. American College Testing.

ACT (1997). *Trend of Increase in ACT College-Entrance Scores Continues.* [on line]. Obtenido el 08/02/2000 en URL: http://www.act.org/news/releases/1997/08-13-97.html

ATP (1998). *Member Directory.* [on line]. Obtenido el 31/01/2000 en URL: http://www.testpublishers.org/memdir.htm. American Tests Publishers.

Bennett, R.E. (1997). *Speculations on the Future of Large-Scale Educational Assessment.* Reunión del National Research Council's Board on Testing

and Assessment. Orlando (Florida), Febrero.

Bennett, R.E. (1998). *Reinventing Assessment: speculations on the Future of Large-Scale Educational Assessment.* Princeton, NJ: Educational Testing Service.

BOE (2000). RESOLUCIÓN 452/38019/2000, de 2 de febrero, de la Subsecretaría, por la que se convocan plazas para acceso a militar profesional de Tropa y Marinería. *Boletín Oficial del Estado, Núm. 33. 8/02/2000,* 5813-5843.

CEST (1998) *The Consortium for Equity in Standards and Testing.* [on line]. Obtenido 10/02/2000 en URL: http://www.csteep.bc.edu/ctest

College Board (2000). *ACCUPLACER™.* [on line]. Obtenido el 10/02/2000 en URL: http://www.collegeboard.org/accuplacer/html/accupla1.html

Crocker, L.M. y Algina, J. (1986). *Introduction to Classical and Modern Tests Theory.* Nueva York, NY: Holt, Rinehart and Winston.

Elosúa, P. y López, A. (1999). Differential Item Functioning and bias in the adaptation of two verbal tests. *Psicológica, 20 (1),* 23-40.

ETS (1994). *National Council Licensure Examination: Registered Nurse (NCLEX:RN).* Test TC 018025. Princeton, N.J.: Educational Testing Service.

ETS (1999). *What is ETS?* [on line]. Obtenido el 31/01/2000 en URL: http://www.ets.org/aboutets/visitors.html. Princeton, N.J.: Educational Testing Service.

FairTest (2000). *FairTest Goals and Principles.* [on line]. Obtenido el 31/01/2000 en URL: http://www.fairtest.org/ftgoals.htm.

Ferrando, P.J. (1999). Likert scaling using continuous, censored, and graded response models: effects on criterion-related validity. *Applied Psychological Measurement, 23 (2),* 161-175.

Gil, G. y Alabau, I. (1997). *Evaluación comparada de la enseñanza y el aprendizaje de la lengua inglesa.* Madrid: INCE.

Hopkins, K.D. (1998). *Educational and Psychological Measurement and Evaluation.* Needhamheight, MA: Allyn y Bacon.

Holland, P.W. y Wainer, H. (1993). *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaulm Associates Pub.

INCE (1998). ¿Qué es el INCE? [on line]. Obtenido el 31/01/2000 en URL: http://www.ince.mec.es/pres/que.htm#pres1.

Jaeger, R.M. (1989). Certification of Student Competence. En R.L. Linn (Ed.), *Educational Measurement. (3rd ed.).* Nueva York, NY: Macmillan Pub.

Linn, R.L. (1989). Current Perspectives and Future Directions. En R.L. Linn (Ed.), *Educational Measurement.* (3rd ed.). Nueva York, NY: Macmillan Pub.

Lorenzo, U. (2000). The weighted oblimin rotation. *Psychometrika, 65 (3),* 301-318.

McRury, K., Nagy, P. y Traub, R.E. (1991). Reflections on Large-Scale Assessment of Student Achievement. En R.K. Hambleton y J.N. Zaal (Eds.), *Advances in Educational and Psychological Testing.* Norwell, MS: Kluwer Academic Pub.

MEC (1999). *Pruebas de acceso a la Universidad.* [on line]. Obtenido el 31/01/2000 en URL: http://www.mec.es/inf/comoinfo/d-3-1-1.htm

MSC (2000). *Acceso a Plazas de Formación.* [on line]. Obtenido el 31/01/2000 en URL: http://www.msc.es/formacion/info_general/f_plazas.htm

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

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,* 151-157.

Olea, J., Ponsoda, V. y Prieto, G. (1999). *Tests informatizados.* Madrid: Pirámide.

Olea, J., Ponsoda, V., Revuelta, J., Hontangas, P. y Suero, M. (1999). Investigación en tests adaptativos informatizados. En J. Olea, V. Ponsoda y G. Prieto (Eds.), *Tests informatizados.* Madrid: Pirámide.

Padilla, J.L., González, A. y Pérez, C. (1998). Instructional differences and differential item functioning: Agreement between the Mantel-Haenszel and the logistic regression. *Psicológica, 19(3),* 201-215

Pérez, M.J., García-Gallo, J. y Gil, G.(1995). *Evaluación de la Educación física en la educación primaria.* Madrid: INCE.

Ponsoda, V. Olea, J. y Revuelta, J. (1994). ADTEST: A computer-adaptative test based on the maximun information principle. *Educational and Psychological Measurement, 54,* 680- 686.

Ponsoda, V., Wise, S.L., Olea, J. y Revuelta, J. (1997). An investigation of self-adapted testing in a Spanish high school population. *Educational and Psychological Measurement, 57(2),* 210-221.

Prieto, P., Barbero, M.I. y San Luís, C. (1999). Detección del funcionamiento diferencial de los ítems en una prueba de ciencias. *Psicothema, 11 (3),* 691-697.

Prieto, P. y Muñiz, J. (2001). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 66-71.

RELIEVE (1998). *Revista Electrónica de Investigación y Evaluación Educativa, 4 (2)* [on line]. Obtenido el 07/02/2000 en URL: http://www2.uca.es/RELIEVE/INDICEV4N2.HTM

TEA (1999). *Presentación.* [on line]. Obtenido el 31/01/2000 en URL: http://www.teaediciones.es

Van der Linden, W.J. (1999). Computerized Educational Testing. En G.N. Masters y J.P. Keeves (Eds.), *Advances in Measurement in Education Research and Assessment.* Kidlington, Oxford: Pergamon.

Villegas, A.M. (1992). The Competence needed by Beginning Teachers in a Multicultural Society. *Annual Meeting of the American Association of Colleges of Teacher Education.* San Antonio, TX. Febrero.

Wainer, H. (1990). Introduction and History. En H. Wainer (Ed.), *Computerized Adaptive Tesing: A Primer.* Hillsdale, NJ: Lawrence Erlbaulm Associates Pub.

Weiss, D.J. y Schleisman, J.L. (1999). Adaptative Testing. En G.N. Masters y J.P. Keeves (Eds.), *Advances in Measurement in Education Research and Assessment.* Kidlington, Oxford: Pergamon.

Whitney, D.R. (1989). Educational Admisisons and Placement. En R.L. Linn (Ed.). *Educational Measurement. (3rd Ed.).* Nueva York: MacMillan Pub.