La Medición en el Ámbito Educativo

Educational Measurement

María José Navas
Universidad Nacional de Educación a Distancia, España

Resumen. Se examina el papel que juega la medición en el trabajo del psicólogo educativo en el centro escolar y también dentro del campo general de la educación, pasando revista a los instrumentos que se utilizan en tres importantes ámbitos que ilustran las necesidades de información que tienen los distintos agentes educativos: los exámenes que ponen los profesores a sus alumnos, la prueba de acceso a la Universidad y las pruebas utilizadas en las evaluaciones del sistema educativo. Seguidamente se indican las claves para obtener buenas medidas educativas para finalizar realizando algunas llamadas de atención sobre puntos de interés relativos al trabajo del psicólogo educativo dentro del centro escolar y también a la medición en el campo general de la educación.
Palabras clave: medición, evaluación, psicólogo educativo, exámenes, prueba de acceso a la Universidad, evaluación del sistema educativo, calidad psicométrica.

Abstract. The role played by measurement in education as well as in the job carried out by an educational psychologist within school is reviewed. Three important measuring devices are examined, so as to show the kind of information that is needed in different areas: the exams teachers give regularly to students in the classroom, the college entrance test, and the tests administered in programs for national or international student assessments. Therewith, a discussion about how to achieve sound educational measures is presented. Finally, some points of interest are highlighted as far as the job of the educational psychologist within school is concerned and also in the more general field of education.
Keywords: measurement, educational psychologist, exams, college entrance test, assessment of educational progress, psychometric quality.

"La historia de la ciencia es la historia de la medida" (Cattell, 1890, p. 376). Esta escueta afirmación revela con claridad el decisivo papel que juega la medición en la ciencia, donde resulta clave para describir y explicar (y por ende en la predicción y control de) los fenómenos. La medición aporta precisión y rigor a las descripciones de los hechos, de las conductas de los sujetos y de las situaciones o contextos en las que éstas tienen lugar; por otro lado, la explicación, predicción y control de la realidad supone formular modelos o teorías, que no son más que una red de proposiciones referidas a un conjunto de constructos o variables que es necesario medir apropiadamente.

Una medición rigurosa de las variables constituye el paso previo e ineludible para cualquier uso posterior que se vaya a hacer de ellas. Los investigadores necesitan obtener medidas de las variables implicadas en sus hipótesis; los psicólogos educativos necesitan evaluar e intervenir y tanto la evaluación como la intervención se apoyan en la medición. Ésta proporciona información objetiva en la que basarse para tomar decisiones acerca de los sujetos, aumentando la eficiencia en la toma de decisiones por parte de psicólogos y educadores. De hecho, una de las razones que explica el gran éxito que tuvieron en EE.UU. los tests al despuntar el siglo XX es que representan la posibilidad de juzgar a las personas por sus competencias, habilidades o conocimientos, esto es, por méritos propios y no por su nivel socioeconómico, red social, apariencia o por el juicio subjetivo de profesores o supervisores.

Tras el éxito cosechado por los tests creados para reclutar personal para la contienda militar de la Primera Guerra Mundial, Cattell funda en 1922 la primera entidad dedicada a la producción masiva de tests (*Psychological Corporation*). En 1940 se habían comercializado ya 2600 baterías estandarizadas de rendimiento, evaluando desde la inteligencia y el retraso mental hasta la distancia social entre grupos, pasando por la personalidad, los intereses y actitudes de grupos y sujetos. La generalización de los tests en la sociedad es un hecho.

Ahora bien, en 1940 tiene también lugar la famosa reunión donde la Comisión para el Avance de la Ciencia de la Sociedad Británica invita amablemente a los psicólogos a renunciar a su empeño de medir variables subjetivas, ya que la medición fundamental o directa es inviable en psicología, al ser propiedades intensivas la mayoría de las variables psicológicas. Como certeramente señala Muñiz (1998a), "no resulta sencillo medir con rigor en sentido clásico, acorde con los axiomas de Hölder y las propuestas ortodoxas de

La correspondencia sobre este artículo debe enviarse al Departamento de Metodología de las Ciencias del Comportamiento. Facultad de Psicología UNED, Juan del Rosal, 10. 28040 Madrid E-mail: *mjnavas@psi.uned.es*

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Campbell. Sin embargo, los psicólogos miden sus variables y desarrollan sofisticados instrumentos a tal efecto, los más conocidos de los cuales para el público son los tests, si bien los especialistas utilizan una gama mucho más amplia en sus investigaciones e intervenciones" (p. 3).

Si bien la implantación y generalización de los tests en la sociedad desmentía rotundamente la conclusión de dicha comisión, la psicología estaba pagando el precio de su origen, en un marco claramente positivista y con la pretensión de emular el canon de ciencia entonces al uso, por lo que los primeros brotes de medición se desarrollaron dentro de la concepción conservadora de la medición (Savage y Ehrlich, 1990) en el mismo marco de la medición de las magnitudes físicas. Puede consultarse en Navas (2001) las distintas formas en las que la psicología consiguió sortear este aparente callejón sin salida.

Yela (1984) decía muy gráficamente que "medir no significa necesariamente aplicar una unidad extensa a objetos extensos, como el metro a un retal" (p. 7). La medición sirve para cuantificar y expresar –normalmente en forma numérica– las características de los estímulos y de las personas, de forma que se pueda utilizar los números –mucho más fáciles de manejar y, por tanto, más convenientes– como si se tratase de lo representado. Meliá (1990) define la medición en psicología como "la asignación de números según reglas a constructos teóricos mediante el uso de indicadores empíricos apropiados de los mismos" (p. 34). La teoría de la medición se ocupa del estudio de los modelos mediante los cuales se conocen las reglas para una correcta asignación de los números y la psicometría permite justificar y obtener medidas de las variables psicológicas y educativas.

La medición psicológica y educativa no solo es posible sino que es necesaria, imprescindible. En el siguiente apartado se va a examinar el papel que juega en el trabajo del psicólogo educativo en el centro escolar y también dentro del campo general de la educación, pasando revista a los instrumentos que se utilizan en tres importantes ámbitos que ilustran las necesidades de información que tienen los distintos agentes educativos: los exámenes que ponen los profesores a sus alumnos, la prueba de acceso a la Universidad y las pruebas utilizadas en las evaluaciones del sistema educativo.

Según Hopkins (1998), existe una estrecha correspondencia entre la madurez científica de una disciplina y el grado en que se pueden medir de forma objetiva y precisa las variables relevantes para la misma. En el siguiente apartado se indican las claves para obtener buenas medidas: instrumentos con una adecuada calidad métrica y profesionales con las competencias necesarias para seleccionar (o construir) el mejor instrumento en cada caso y utilizarlo de forma apropiada. Solo así se puede lograr medidas que proporcionen la información requerida y puedan ser utilizadas, si procede, para tomar las decisiones correspondientes. Asimismo, se reflexiona sobre las consecuencias de una mala práctica y el posible impacto adverso de la medición.

En el último apartado se realizan algunas llamadas de atención acerca de puntos de interés en los dos planos analizados en el primer apartado: el trabajo del psicólogo educativo en el centro escolar y la medición en el campo general de la educación.

El papel de la medición

En el trabajo del psicólogo educativo en el centro escolar

La medida también está en el origen de la Psicología Educativa que, en buena parte, es deudora de los trabajos de Binet realizados a principios del pasado siglo para responder al encargo del ministro de Educación de elaborar un método objetivo para ayudar en el diagnóstico de los niños con discapacidad intelectual que asistían a las escuelas públicas de París. El origen de esta profesión está muy vinculado a la aplicación de pruebas psicométricas estandarizadas que exigen el concurso de profesionales con la capacitación necesaria para ello. Como indica Méndez (2011), "la necesidad de evaluar y de hacerlo a través de pruebas estandarizadas ayudó a identificar una profesión emergente" (p. 40) con evaluaciones dirigidas en sus inicios sobre todo al ámbito de la educación especial.

La aplicación de pruebas estandarizadas sigue siendo una tarea muy relevante en el trabajo del psicólogo educativo y posiblemente constituya sus señas de identidad más visibles. Ahora bien, ni las pruebas estandarizadas son el único modo de obtener las medidas o información de interés ni el papel de la medición se agota ahí sino que está presente no solo en el principio sino también al final del trabajo del psicólogo educativo.

Aunque la definición del rol y funciones del psicólogo educativo en el centro escolar ha sido y sigue siendo objeto de no poca polémica y controversia (véase, por ejemplo, el número monográfico dedicado por esta misma revista a comienzos de 2011), se acepta generalmente como marco de su trabajo el proceso siguiente. A partir de un análisis de necesidades o de una situación problema, se obtiene una evaluación inicial o diagnóstico, que servirá para poner en marcha una determinada vía de acción o intervención, en base a programas (preventivos, de desarrollo, terapéuticos y de orientación académica y profesional) o terapéutica de enfoque clínico (Garaigordobil, 2009), en el propio centro o bien derivando la intervención a instituciones o centros especializados. El proceso siempre ha de terminar con una valoración de la intervención realizada que supone, a su vez, medir los efectos de los programas de intervención y tomar las decisiones correspon-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Tabla 1. Objetivos de la evaluación en el aula y de la evaluación a gran escala*

EN EL AULA	A GRAN ESCALA
a) Determinar si los estudiantes dominan un determinado concepto o habilidad.	a) Identificar puntos fuertes y débiles de los estudiantes.
b) Motivar a los estudiantes para que se involucren activamente en el proceso de enseñanza-aprendizaje.	b) Determinar si los estudiantes cumplen los objetivos educativos establecidos.
c) Conseguir que los estudiantes sean capaces de razonar y de aplicar los contenidos aprendidos.	c) Determinar cómo agrupar a los estudiantes.
d) Ayudar a desarrollar una actitud positiva hacia las asignaturas.	d) Identificar estudiantes con necesidades especiales.
e) Informar a los padres de lo que saben y son capaces de hacer sus hijos.	e) Comparar el rendimiento de determinados grupos de estudiantes con el promedio nacional.
f) Informar a los estudiantes de lo que saben y son capaces de hacer.	f) Evaluar la eficacia de un nuevo curriculum.
g) Informar a los estudiantes de lo que esperan de ellos.	g) Evaluar a profesores y directores de los centros.
h) Indicar a los estudiantes dónde han de centrarse para mejorar.	h) Proporcionar información para la acreditación de los centros.
i) Elaborar el boletín de notas.	i) Comparar distintos centros escolares.
j) Evaluar la eficacia de los métodos pedagógicos.	j) Distribuir recursos.

*Elaborada a partir de McMillan (2008, pp. 6-7).

dientes derivadas de dicha valoración. La medición proporciona la información necesaria para diseñar la mejor estrategia posible de intervención y la medición indica también hasta qué punto dicha estrategia ha conducido a los resultados apetecidos.

Donde existe poca controversia es en lo ingente de la tarea a acometer, ya que el trabajo del psicólogo educativo en el centro no está limitado a evaluar a estudiantes individuales (con necesidades tanto educativas como psicológicas) o grupos (de estudiantes, familias) sino también la propia organización escolar, examinando aspectos tan importantes como, por ejemplo, la satisfacción y el clima escolar, que tienen un profundo impacto en el rendimiento educativo.

Las técnicas más utilizadas para la evaluación inicial o diagnóstico son los tests psicológicos, la entrevista psicológica y la observación en sus diferentes modalidades, debiendo seleccionar en cada caso la(s) más adecuada(s) al objeto (individuos, grupos, instituciones) y al objetivo planteado en la evaluación. El resto de fases del proceso dependen críticamente de la información obtenida al inicio, por lo que resulta fundamental conocer y tener acceso a los mejores instrumentos de medida y aplicarlos con el rigor y en las condiciones establecidas por sus autores, para así disponer de medidas adecuadas que conduzcan a establecer un buen diagnóstico de la situación, base incuestionable de cualquier intervención posterior.

En el campo de la educación

Son muchas y poderosas las razones para evaluar, tanto para educandos como para los distintos agentes educativos. Los alumnos quieren saber sus notas; los padres quieren saber cómo lo están haciendo de bien sus hijos en el colegio; los profesores quieren saber cuáles son los conocimientos previos de sus alumnos antes de abordar un nuevo bloque de contenidos, cómo van asimilando éstos conforme los imparten y hasta qué punto lo han hecho una vez completada esa unidad temática; los directores quieren saber dónde se sitúan sus alumnos en relación a los de otros centros y a la media de su comunidad y de España, y las autoridades educativas quieren saber hasta qué punto la educación que reciben los estudiantes es una educación de calidad. En pocas palabras, se necesita la medición para tomar decisiones sobre los estudiantes, el proceso de enseñanza-aprendizaje, el currículo, los profesores y los centros escolares.

Simplificando enormemente se puede distinguir dos tipos de evaluaciones muy diferentes, que responden a objetivos distintos y utilizan procedimientos también distintos para medir u obtener la información necesaria: la evaluación que realizan los profesores dentro del aula con el objetivo de conocer pero, sobre todo, de mejorar el aprendizaje de sus alumnos y las evaluaciones a gran escala que tienen lugar periódicamente sobre el conjunto de la población (por ejemplo, las pruebas de acceso a la Universidad) o sobre una muestra de la misma (por ejemplo, el estudio PISA o las evaluaciones diagnósticas del sistema educativo), con el objetivo básico de acreditar o rendir cuentas acerca de los resultados educativos de los estudiantes. Las dos son esenciales en la actualidad. En la Tabla 1 se ofrece una sucinta caracterización de ambas y se detalla a continuación algunos aspectos de interés en cada una de ellas.

Exámenes en el aula

La evaluación es consustancial al trabajo docente. Gracias a ella, el profesor dispone de información acerca de lo que sabe realmente el alumno y puede organizar de manera acorde el trabajo para facilitar y después documentar o dar fe de su aprendizaje, informando a estudiantes, padres y a las instancias que corresponda

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Tabla 2. El proceso de toma de decisión del profesor antes, durante y después de la instrucción*

ANTES	DURANTE	DESPUÉS
¿Tienen los estudiantes los conocimientos y habilidades previas necesarias? ¿Qué puede interesar a los estudiantes? ¿Qué puede motivar a los estudiantes? ¿Cuánto tiempo debería dedicar a cada unidad temática? ¿Con qué método debería abordarla? *¿Cómo debería evaluar a los estudiantes?* ¿Qué tipo de aprendizaje en grupo convendría utilizar? ¿Cuáles son mis objetivos para el aprendizaje?	¿Me prestan atención los estudiantes? *¿Están entendiendo el material?* ¿A quiénes debería preguntar? *¿Qué debería preguntar?* ¿Cómo debería responder a las preguntas de los estudiantes? ¿Cuándo debería dejar de presentar material? *¿Qué estudiantes necesitan ayuda extra?* *¿A qué estudiantes podría dejar a su aire?*	*¿Cuánto han aprendido los estudiantes?* ¿Qué debería hacer a continuación? ¿Es necesario revisar algo que no se entendió? *¿Qué calificaciones debería dar?* *¿Qué debería decir a los estudiantes?* ¿Qué debería cambiar la próxima vez para enseñar el tema? *¿Las puntuaciones del examen reflejan realmente lo que saben y son capaces de hacer los estudiantes?* *¿Hay algo que los estudiantes no comprendan bien?*

*Elaborada a partir de la Figura 1.2 de McMillan (2008, p. 5)

sobre el progreso académico de éstos. McMillan (2008) concibe la enseñanza como un proceso de toma de decisiones que tienen lugar antes, durante y después del proceso de enseñanza-aprendizaje. La Tabla 2 indica con claridad hasta qué punto la evaluación (preguntas en cursiva) está ligada y al servicio de la instrucción, proporcionando una fuente de información básica que ayuda a tomar las decisiones correspondientes al profesor, junto a su experiencia, lógica e intuición.

La batería de procedimientos de medida a disposición del profesor es muy variada. Desde sencillos tests para obtener un rápido plano de situación antes de abordar una nueva unidad temática hasta un detallado examen que evalúe el nivel de conocimiento, comprensión y capacidad de aplicación de los contenidos estudiados a situaciones cotidianas, pudiendo utilizar para ello preguntas de elección (binaria, múltiple, de emparejamiento) o de respuesta abierta, en las que el alumno debe proporcionar y no seleccionar la respuesta a una tarea. Aquí el catálogo ofrece también múltiples posibilidades: desde preguntas que requieren una respuesta corta o completar una frase o tabla hasta la presentación oral sobre un tema, la redacción del informe acerca de alguna tarea realizada en el laboratorio (lo que se conoce con los términos de pruebas de desempeño o de evaluación de la actuación, *performance assessment* en inglés), escribir una carta al editor de un periódico local o reconocer y formular en términos matemáticos un problema planteado fuera del contexto escolar habitual (lo que se conoce con el término de pruebas de evaluación auténtica, *authentic assessment* en inglés). Este último tipo de tareas (véase Martínez-Arias, 2010 para una presentación en castellano) sirve además para resaltar la relevancia del aprendizaje, con el importante componente motivacional y de refuerzo que ello supone para el proceso de enseñanza-aprendizaje.

Como en cualquier otra tarea evaluativa o de investigación, es el objetivo que se plantea en cada momento el profesor la variable clave a la hora de decidir qué tipo de procedimiento escoger del catálogo para obtener las medidas de interés. En cualquier caso, a lo largo del curso el profesor puede ir recogiendo los trabajos de cada alumno en una carpeta (*portfolio assessment* en inglés), que ilustra su progreso y los resultados educativos logrados. Esta carpeta junto con la información procedente de la observación diaria del desarrollo del trabajo en el aula (o cualesquiera otros procedimientos menos formales o estructurados utilizados por el profesor) sirven para disponer de una estimación bastante aquilatada del grado de éxito de cada alumno en la consecución de los objetivos educativos.

La importancia de la evaluación en el proceso de enseñanza-aprendizaje queda patente en la Ley Orgánica de Educación 2/2006, al ser enunciada como la segunda función del profesorado (artículo 91) y al incluir los criterios de evaluación dentro de la definición misma del currículo junto con los objetivos, competencias básicas, contenidos y métodos pedagógicos (artículo 6), si bien la ley no da ningún tipo de indicación acerca de cómo ha de ser ésta, más allá de su carácter continuo en todas las etapas del sistema educativo (artículos 20, 28, 38 y 43).

Un examen a los planes de estudios de los nuevos grados en Maestro así como del máster de Formación del Profesorado pone de manifiesto que apenas hay ninguna asignatura cuyo objetivo sea proporcionar la preparación y competencia técnica necesaria para acometer tan importante tarea.

Sin regulación alguna acerca de cómo abordar las tareas de evaluación y con una escasísima formación técnica para ello, no resulta fácil conseguir las 7 competencias que establecieron para los profesores en 1990 tres importantes asociaciones profesionales estadounidenses (*American Federation of Teachers*, *National Council on Measurement in Education* y *National Education Association*). Según estas directrices, los profesores deben ser competentes a la hora de (1) elegir métodos de evaluación apropiados para tomar decisiones relativas a la instrucción, en función de su calidad técnica, utilidad, conveniencia e imparcialidad, (2) llevar a cabo cualquier tipo de evaluación, (3) apli-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

car, puntuar e interpretar tanto exámenes en el aula como pruebas estandarizadas, (4) utilizar los resultados de la evaluación para tomar decisiones sobre los estudiantes, la docencia, los programas de estudio y la mejora del centro escolar, (5) diseñar procedimientos válidos y justos para calificar a sus alumnos y (6) informar de los resultados de la evaluación a los estudiantes, padres, otros educadores y cualquier otra audiencia lega. La séptima directriz tiene que ver con la necesidad de adherirse a una práctica ética y legal. Ainsworth y Viegut (2006) incluyen un capítulo concebido para ayudar a los educadores a desarrollar estas competencias en evaluación, en un texto que constituye una guía muy útil y práctica para utilizar la evaluación en el proceso de enseñanza-aprendizaje.

Prueba de Acceso a la Universidad

La nota de la Selectividad constituye al menos el 40% de la calificación que finalmente abre o cierra la puerta a la enseñanza superior y este examen es probablemente la prueba de evaluación más masiva realizada en España (alrededor de 200.000 estudiantes cada año).

El sistema de acceso a la Universidad procede asumiendo el mismo valor a las calificaciones obtenidas en los exámenes de todas las universidades, supuesto perfectamente viable desde el punto de vista técnico pero del todo insostenible habida cuenta del modo de operar con estas pruebas.

Por un lado, los exámenes realizados en los distintos distritos universitarios son diferentes y seguramente no tienen el mismo nivel de dificultad. Por otro, los profesores encargados de la corrección de los ejercicios pueden no estar utilizando los mismos criterios a la hora de calificar. Esto significa que un alumno podría no acceder a la Universidad (o no alcanzar la nota que piden en la titulación de su elección en la Universidad donde quiere estudiar) con el examen realizado en el distrito universitario que le corresponde, pero sí con el examen –más fácil o más próximo a los contenidos que mejor conoce– que han puesto en otros lugares, o según lo estrictos o laxos que sean los correctores correspondientes. Dicho de otro modo, el que un alumno apruebe o no la Selectividad no solo depende de su capacidad para abordar con éxito los estudios universitarios sino también del examen y del tribunal calificador que le haya tocado.

Para complicar un poco más las cosas, apenas hay estudios acerca de la calidad métrica de estas pruebas. Los escasos estudios sobre el tema revelan que dentro de una misma materia la variedad en los exámenes es enorme: en la extensión abarcada del programa, en la estructura, contenido y nivel de exigencia de la prueba y en las especificaciones para la puntuación y criterios de corrección de las preguntas (Muñoz-Repiso y cols., 1997). Estos estudios también evidencian que las diferencias de criterio en los tribunales calificadores son responsables en buena parte de las diferencias en las calificaciones de los estudiantes (Martí Recober, 1998; Sanz, 1992). Sirva como dato ilustrativo que cuando un alumno presenta una reclamación en la Selectividad, se contempla la posibilidad de hasta 2 puntos de diferencia entre la calificación asignada por los dos correctores antes de efectuar una tercera corrección (hasta 3 puntos –la distancia entre un notable y un suspenso– en la legislación anterior); sin embargo, en ocasiones se dirime a nivel de decimales el poder entrar en la titulación de elección en una universidad determinada.

Con el nuevo Real Decreto 1640/2008 para regular las condiciones de acceso y admisión a las universidades públicas españolas, se desperdició una oportunidad histórica –a punto de entrar en el espacio europeo de educación superior y con dos ministerios con responsabilidad sobre el sistema educativo– al aprobar una reforma de la prueba de acceso a la Universidad que sigue sin garantizar la comparabilidad de las calificaciones obtenidas por alumnos que se han examinado en zonas geográficas diferentes. La parte positiva es que se propone una prueba algo más corta y flexible, con un ejercicio oral para el idioma extranjero y con la posibilidad de que la nota final se adecue mejor a las preferencias de elección del estudiante y a las exigencias específicas de formación de las distintas titulaciones. Asimismo, se prevé hacer público un informe anual de la prueba de acceso a la Universidad; algo ciertamente positivo, dado que apenas hay estudios que indiquen si la Selectividad desempeña adecuadamente la función prevista.

Buena parte de los problemas señalados para la prueba de la Selectividad son compartidos por las pruebas de acceso a la formación sanitaria especializada (FIR para farmacéuticos, MIR para médicos, PIR para psicólogos, etc.) y también con las pruebas de acceso a empleos públicos. En realidad, se trata de un problema generalizado en la mayoría de los ámbitos en los que se realiza en España una aplicación de pruebas a gran escala. Las pruebas suelen ser construidas localmente y para cada ocasión concreta por tribunales o grupos de expertos en el contenido correspondiente. Los ejercicios de los examinandos suelen ser evaluados por tribunales diferentes sin unos criterios de corrección suficientemente detallados, claros e inequívocos, dejando un espacio demasiado amplio a la subjetividad. Y las pruebas no se someten a ningún control métrico de calidad: no se evalúa su fiabilidad ni mucho menos su validez, así como tampoco se analiza si las preguntas incluidas proporcionan información relevante y funcionan del modo esperado.

Evaluaciones del sistema educativo

En los últimos años la cultura de la evaluación ha cobrado un fuerte impulso. De hecho, el título VI de la

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Ley Orgánica de Educación 2/2006 está dedicado íntegramente a la evaluación del sistema educativo, con objetivos muy ambiciosos y un ámbito que abarca desde los resultados de los alumnos hasta la evaluación de la inspección y de las propias Administraciones educativas (véase la Tabla 3).

Tabla 3. Objetivos y ámbitos de la evaluación en la Ley Orgánica de Educación 2/2006

OBJETIVOS	ÁMBITOS
a) Contribuir a mejorar la calidad y equidad de la educación. b) Orientar las políticas educativas. c) Aumentar la transparencia y eficacia del sistema educativo. d) Ofrecer información sobre el grado de cumplimiento de los objetivos de mejora establecidos por las Administraciones educativas. e) Proporcionar información sobre el grado de consecución de los objetivos educativos españoles y europeos, así como del cumplimiento de los compromisos educativos.	a) Procesos de aprendizaje y resultados de los alumnos. b) La actividad del profesorado. c) Los procesos educativos. d) La función directiva. e) El funcionamiento de los centros docentes. f) La inspección. g) Las Adminstraciones educativas.
(Artículo 140.1)	(Artículo 141)

La citada ley contempla la realización de dos evaluaciones de diagnóstico para conocer las competencias básicas de los estudiantes en momentos críticos de la enseñanza primaria (al finalizar 4º) y secundaria (al finalizar 2º de la ESO), con un carácter formativo y orientador para los centros e informativo para las familias y el conjunto de la comunidad educativa (artículos 21 y 29). Se ha llevado a cabo ya la primera evaluación en ambas etapas educativas, encuestando a cerca de 30.000 estudiantes con una participación de casi 900 centros y equipos directivos y con muestras de 1300 profesores en primaria y 4500 en secundaria, utilizando pruebas de lápiz y papel y archivos de audio para la comprensión oral como parte de la competencia en comunicación lingüística.

La participación en evaluaciones internacionales también se contempla en la ley (artículo 143.2) y España participa sistemáticamente en evaluaciones auspiciadas por organizaciones como la IEA y la OCDE (véase la Tabla 4). Además, cada vez es mayor la implicación de las Administraciones educativas autonómicas: en la edición 2009 del estudio PISA (en el que se inspiran en gran medida las dos evaluaciones de diagnóstico señaladas) han participado con muestra ampliada –para poder disponer así de datos representativos– 14 comunidades y también las ciudades autónomas de Ceuta y Melilla (ninguna, 3 y 10 comunidades en las ediciones anteriores del estudio PISA, respectivamente).

En algunas comunidades autónomas las Administraciones educativas han promovido además la realización de evaluaciones regionales de competencias básicas en otros momentos del periplo educativo de los estudiantes. Así, en Madrid se realiza anualmente la prueba de Conocimientos y Destrezas Indispensables en Lengua y Matemáticas, en 6º de Primaria desde el año 2005 y en 3º de la ESO desde 2008. En Cataluña se realiza desde 2009 la prueba de evaluación de competencias básicas (lingüística y matemática) en 6º de Primaria.

Además de las evaluaciones de diagnóstico, la Ley Orgánica de Educación 2/2006 contempla la elaboración del Sistema Nacional de Indicadores de la Educación para conocer el sistema educativo y orientar la toma de decisiones (artículo 143.3). Se trata de un conjunto de 38 indicadores agrupados en 5 bloques que se publican cada dos años. En 2006 se definieron 15 indicadores prioritarios y para éstos los datos se actualizan anualmente. A finales de los años ochenta, la OCDE puso en marcha un proyecto para ofrecer indicadores cuantitativos que permitieran la comparación de los sistemas educativos de los países miembros (proyecto INES) y conocer así la eficacia y evolución de dichos sistemas. Los primeros indicadores se publicaron en 1992 con el nombre *Education at a Glance* y, desde entonces, se publican anualmente. La versión en español está a disposición del público desde 2005 con el nombre de *Panorama de la educación*.

En la página web del Instituto de Evaluación se puede consultar y descargar los informes de todos los estudios indicados en la Tabla 4.

La calidad de las medidas en el ámbito educativo

Solo es posible obtener medidas que proporcionen la información deseada si se trabaja con instrumentos con una adecuada calidad métrica y con profesionales cualificados para ello.

Dado que el error es consustancial a la medida, es preciso evaluar la calidad de los instrumentos de medida, esto es, determinar si cumplen o no los criterios métricos de calidad que todo instrumento debe satisfacer para poder ser utilizado con garantías. Esto supone que habrá que estudiar su fiabilidad y validez –criterios métricos de la calidad global de la prueba– y proceder a analizar convenientemente la calidad métrica individual de cada una de sus preguntas, para poder garantizar que se dispone de medidas adecuadas –fiables y válidas– del nivel en que poseen los examinandos la característica evaluada. Solo en este caso se pueden

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Tabla 4. Estudios de evaluación en el sistema educativo español

NACIONALES		
Evaluaciones de diagnóstico	LOE 2/2006	Al finalizar 4º de Primaria Al finalizar 2º de ESO
Evaluación general	Sistema estatal de indicadores de la educación	De resultados educativos De procesos educativos De escolarización De recursos De contextos
	De etapas educativas	
	Infantil	Estudio piloto para desarrollar y validar los instrumentos de evaluación (2007)
	Primaria	Al finalizar la etapa (2007) Conocimiento del medio, Matemáticas, Lengua inglesa y Lengua castellana y Literatura
	ESO	Al finalizar la etapa (2000) Ciencias de la Naturaleza, Ciencias Sociales, Geografía e Historia, Matemáticas, Lengua Castellana y Literatura
INTERNACIONALES		
Auspiciados por la OCDE	PISA	Estudiantes de 15 años Lectura, Matemáticas y Ciencias
	TALIS	Profesores y directores ESO Condiciones de la enseñanza y el aprendizaje
	INES	Sistema internacional de indicadores de la educación
Auspiciados por la IEA	TIMSS	Estudiantes de 9 y 13 años y en el último curso de la enseñanza secundaria Matemáticas y Ciencias
	PIRLS	Estudiantes de 4º de Primaria Lectura
	ICCS	Estudiantes de 14 años Educación cívica y ciudadana

interpretar las puntuaciones obtenidas en la prueba en términos de dicha característica y ser utilizadas con el fin previsto.

Son numerosas las instituciones preocupadas por mejorar la práctica de los profesionales que trabajan en la evaluación psicológica y educativa, así como la formación/información de sus usuarios: en España el Colegio Oficial de Psicólogos (COP) a través de su Comisión de Tests y fuera de nuestras fronteras instituciones homólogas (*American Psychological Association*, *International Test Commission*) e importantes organizaciones profesionales como *American Educational Research Association* y *National Council on Measurement in Education*. Esta preocupación se ha traducido en la elaboración de distintas normativas (véase Tabla 5) que, a modo de códigos o guías de conducta, recogen los principios generales implicados en el uso adecuado de los instrumentos de medida y las directrices que han de regir una práctica profesional responsable, tanto desde un punto de vista ético como técnico.

Un aspecto compartido por todas las normativas es la asignación de responsabilidades no solo a los que construyen y comercializan instrumentos de medida sino también a sus usuarios, esto es, a las personas o entidades que seleccionan y aplican (o encargan la administración de) tests o toman decisiones en base a éstos. De hecho, alguno de estos códigos desdobla sus recomendaciones para usuarios y constructores de pruebas. Los *Estándares para la evaluación psicológica y educativa* (AERA, APA y NCME, 1999) son taxativos a este respecto y sostienen que "la validación es responsabilidad conjunta del constructor y del usuario del test. El constructor es responsable de proporcionar evidencia rele-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Tabla 5. Normativas para una correcta utilización de los instrumentos de medida en el ámbito psicológico y educativo

NORMATIVA	INSTITUCIÓN(ES) QUE LA PROMUEVE(N)
Código de buenas prácticas de evaluación en el ámbito de la Educación	APA (2000)
Código de responsabilidades profesionales en la medición educativa	NCME (1995)
Directrices internacionales para el uso de los tests	ITC (2000)
Estándares para la evaluación psicológica y educativa	AERA, APA y NCME (1999)
Normas mínimas para el uso de los tests	COP
Principios éticos de la evaluación psicológica	APA (2002)

vante así como la base lógica y fundamentos para el uso propuesto del test. El usuario es el responsable último de evaluar la evidencia en el contexto particular en el que se va a utilizar el test" (*op. cit.* p. 11).

La Comisión de Tests del COP propone unas normas mínimas para el uso de los tests, con 12 recomendaciones muy claras y concretas que se pueden consultar en su página web, si bien el documento de referencia por excelencia en el campo lo constituyen los *Estándares para la evaluación psicológica y educativa*, que cuentan con una larga tradición y 6 versiones en menos de 50 años (desde 1954 a 1999), con la séptima revisión en marcha.

Estos estándares abordan el proceso de construcción, evaluación y documentación del test en la parte I, cuestiones relativas a la imparcialidad en la evaluación con tests en la parte II y en la última parte, tras un primer capítulo dedicado a las responsabilidades de los usuarios de los tests, se ocupan de aspectos relativos a la utilización de tests en determinados contextos aplicados (educativo, psicológico, evaluación de programas, acreditación y ámbito laboral).

En el primer estándar formulado en el capítulo dedicado al ámbito de la evaluación educativa (estándar 13.1) se indica que quien ordena la utilización de tests es también responsable de controlar el impacto que tiene su aplicación, así como la responsabilidad de identificar y minimizar las posibles consecuencias negativas a las que pueda dar lugar. La última revisión de los estándares completada hasta la fecha tuvo como núcleo central justamente la incorporación dentro del marco de la validez –el criterio métrico de calidad más importante de un instrumento de medida– de las consecuencias de la utilización de los tests, tras un profundo y prolongado debate no exento de polémica (véase los números monográficos dedicados por la revista *Educational Measurement: Issues and Practice* en el segundo número de los volúmenes 16 y 17, o el excelente trabajo de Moss, 1992).

Sea cual sea la postura defendida y dónde se coloquen los límites de la validez, lo importante es haber traído a primer término las consecuencias derivadas del uso de los tests, como información potencialmente relevante para la toma de decisiones vinculadas a la aplicación de tests. Como es bien sabido, los tests se utilizan para tomar decisiones acerca de los sujetos en distintos ámbitos; en el campo de la evaluación educativa, muchas veces se trata de decisiones muy importantes para la vida de los evaluados y la propia evaluación se está convirtiendo en un instrumento de política educativa. El debate sería hasta qué punto es lícito considerar las consecuencias del uso de los tests como implicaciones no técnicas del trabajo del psicómetra o evaluador y, por tanto, pueden ser confortablemente relegadas. Por otra parte, la separación entre los aspectos técnicos y los que no lo son resulta cada vez más artificial o difusa: la creciente multiculturalidad de buena parte de las sociedades occidentales hace imprescindible que la validación de las puntuaciones de los tests tenga en cuenta los distintos antecedentes lingüísticos y culturales de los sujetos evaluados, lo que ha contribuido a poner a punto una importante tecnología de análisis del funcionamiento diferencial de los ítems y de los tests (más conocido como DIF) o, utilizando un término que parece estar abriéndose paso en el campo, el impacto adverso (Outtz, 2010). La atención a la diversidad también aquí plantea un reto importante (Padilla, Gómez, Hidalgo y Muñiz, 2006).

En ocasiones, una de las consecuencias de la aplicación de tests en evaluaciones diagnósticas prescritas por ley es acomodar el programa de estudios a los objetivos educativos evaluados en esas pruebas, en principio coincidentes con los formulados por las Administraciones educativas pero que, en modo alguno, pueden ser abarcados en su totalidad en las pruebas ni con la profundidad requerida durante el proceso de enseñanza-aprendizaje. Esto suele llevar aparejada una indeseable restricción en el programa de estudios impartido, especialmente en países o estados donde los resultados de las evaluaciones educativas tienen un impacto en la financiación de los centros escolares. A la inversa, puede no tener un efecto perverso sino servir de ayuda a algunos centros a poner el foco en los aspectos más sustantivos y reorientar de manera acorde la instrucción, para lograr esas competencias básicas en sus estudiantes sin tratar de incrementar con atajos indebidos las puntuaciones en las pruebas.

En definitiva, si la validez se concibe como un proceso de acumulación de evidencias que permitan interpretar y utilizar las puntuaciones de un test de un modo determinado, caben pocas dudas de que las consecuencias –previstas pero también imprevistas, actuales y potenciales– de utilizar los tests tienen que ser introdu-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

cidas en la ecuación antes de concluir el proceso de toma de decisión al que suelen nutrir los tests.

No solo es importante estar alerta acerca de las posibles consecuencias de la utilización de los tests sino también y todavía más importante acerca del posible mal uso o utilización inadecuada de éstos, aspecto éste que entra de lleno y por derecho propio dentro de la esfera de la validez. Según Muñiz (1998b), "en esto ocurre con los tests lo mismo que con cualquier otra tecnología, sea del campo que sea, su potencialidad para causar daño proviene más de su uso por gentes sin preparación o principios que de sus propiedades técnicas. El ejemplo está en el garaje, allí mora el mayor homicida de nuestros días, el automóvil; no hay guerra que le haga sombra pero muy pocos de los accidentes son atribuibles a fallos mecánicos, es el uso. Salvando las distancias, con los tests ocurre algo parecido, el problema es el uso que se hace de ellos." (p. 311). Según Anastasi (1987), el mal uso de los tests puede proceder de la negligencia, de unos conocimientos o formación insuficientes o también de un intento deliberado de distorsionar la realidad. Muñiz (1998b) apunta a la segunda como causa más frecuente de utilización inadecuada de los tests. Los estándares 13.10 y 13.13 abundan en la necesidad de que los responsables de programas de evaluación educativa se aseguren de contar con profesionales con la competencia técnica necesaria en el proceso de administración de los tests y en la posterior asignación de sus puntuaciones, así como de que las personas que han de interpretarlas para tomar decisiones dentro del ámbito escolar estén debidamente cualificadas o bien reciban el asesoramiento necesario para ello.

Los *Estándares para la evaluación psicológica y educativa* constituyen la normativa ética y técnica de referencia para el trabajo del psicólogo educativo en el centro escolar y también para el resto de procesos de evaluación examinados, con la única excepción de los exámenes realizados por los profesores en el aula. Esto no significa en modo alguno que esos instrumentos de medida no deban satisfacer unos determinados criterios sino sencillamente que plantean un menor nivel de exigencia métrica, ya que en un buen número de ocasiones se utilizan para tomar decisiones de bajo impacto en los estudiantes (e.g., cómo continuar el desarrollo de un bloque temático) y constituyen una fuente más de información –no necesariamente la de mayor peso– en el proceso. Como ya se indicó al finalizar el apartado dedicado a los exámenes en el aula, los profesores tienen también su propia normativa que desarrolla las competencias que éstos han de mostrar en materia de evaluación educativa (*American Federation of Teachers* y cols., 1990), competencias importantes que demandan conocer en buena medida muchos aspectos recogidos por los *Estándares para la evaluación psicológica y educativa*, en particular, los relativos a la construcción de las pruebas, a su validez, fiabilidad e imparcialidad. DiRanna y cols. (2008) procuran justamente definir o adaptar estos conceptos a la evaluación en el aula, en un texto que describe un programa para ayudar a los profesores a integrar para mejorar la docencia y la evaluación.

Algunas reflexiones finales

El centro escolar no es solo el lugar clave para el aprendizaje sino el punto en el que interactúa el sistema educativo con todos los sectores implicados (estudiantes, familias, comunidad). Por ello, se hará en primer término alguna reflexión acerca del trabajo del psicólogo educativo dentro del centro escolar, para finalizar con alguna otra relativa a la medición en el campo general de la educación.

Dado el amplio espectro de áreas que abarca el trabajo del psicólogo educativo en los centros escolares (evolutivas, cognitivas, afectivas, comportamentales, psicosociales e institucionales), es importante que desde las asociaciones profesionales se faciliten mecanismos de actualización de los conocimientos y de las nuevas herramientas de evaluación que salen al mercado. En esta línea, es de destacar la iniciativa de la APA de crear una nueva base de datos sobre tests y medidas psicológicas (PsycTESTS), centrada básicamente en tests desarrollados por investigadores y no disponibles comercialmente (con información descriptiva y técnica para cada test y con el test incluido en la mayoría de los casos) pero donde se incorporan también tests publicados con enlaces a las empresas que los comercializan. Asimismo, es de gran interés la evaluación iniciada recientemente por la Comisión de Tests del COP (Muñiz, Fernández-Hermida, Fonseca-Pedrero, Campillo-Álvarez y Peña-Suárez, 2011), con 10 tests ya examinados con un modelo de evaluación desarrollado por la Comisión Europea de Tests y adaptado al contexto español por Prieto y Muñiz (2000).

Dentro de las funciones reconocidas como genuinas en el Perfil Profesional del Psicólogo de la Educación, hay dos que pueden tener en la actualidad una relevancia mayor: la orientación y el trabajo con los profesores.

La atención a las demandas individuales relacionadas con la orientación académica y profesional es básica para mejorar la competitividad de la economía del país y para su progreso económico y social. Cada vez resulta más importante trabajar por la inserción eficaz y madura del alumno en la vida activa. De algún modo, éste es el espíritu que anima el conocido estudio PISA, cuyo objetivo es averiguar hasta qué punto los estudiantes de 15 años pueden usar las habilidades y conocimientos adquiridos para vivir en sociedad y enfrentarse a los retos de la vida adulta.

El estudio revela que el éxito de los sistemas educativos parece tener que ver con su capacidad de inclusión y con la capacidad de centros y profesores para dar respuesta a una población escolar diversa mediante una oferta educativa personalizada (Navas y García-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Forero, 2011; OECD, 2010). El apoyo al profesorado en la atención a la diversidad resulta, por tanto, una tarea clave para conseguir adaptar el proceso de enseñanza-aprendizaje a las características de cada educando, como vía prioritaria en el necesario camino hacia la excelencia. Las nuevas tecnologías pueden jugar aquí un servicio impagable.

El ordenador es una herramienta muy útil y asequible que permite instrumentalizar nuevos modos de aprendizaje, monitorizar y evaluar el progreso e informar sobre el mismo (véase Bennett, 2006 y Ponsoda, 2012, para una revisión de la relación entre las nuevas tecnologías y la medición). Las nuevas tecnologías pueden desempeñar un papel muy importante, no solo por las enormes posibilidades que abren en el proceso de enseñanza-aprendizaje sino por el estímulo y efecto motivador que pueden tener en dicho proceso, habida cuenta de la actitud tan positiva que tienen los alumnos hacia ellas y de lo versados que están en su utilización. Sin embargo, tal y como se ha puesto de manifiesto en la edición 2009 del estudio PISA (Navas y García-Forero, 2011; OECD 2011), aunque los centros escolares parecen estar en general bastante bien dotados en los países de la OCDE, la frecuencia de utilización de las nuevas tecnologías en los centros todavía es más bien baja, tanto dentro de clase como fuera de ella.

El psicólogo educativo puede ser también clave para trasladar al profesor los resultados más importantes revelados por los estudios de evaluación realizados a nivel regional, nacional e internacional y que pueden tener una traslación más o menos directa en su trabajo cotidiano en el aula: la labor de formación –en muchas ocasiones soslayada por los imperativos de necesidades perentorias– resulta decisiva para mejorar el proceso y resultados educativos, así como para facilitar el trabajo de todos los colectivos implicados.

Una de las limitaciones más importantes del estudio PISA es su escasa utilidad para reorientar el trabajo del profesor en el aula, al ser un estudio básicamente dirigido a la comparación entre países. Ahora bien, dado que cada día es mayor el esfuerzo que se pide a los centros escolares para que participen en evaluaciones educativas, podría resultar motivador y, sobre todo, muy útil proporcionarles información específica acerca de puntos fuertes y débiles de los alumnos, esto es, información relevante para mantener o corregir el rumbo dentro del centro, más allá de la imagen que ofrece el estudio en cuestión sobre la situación de una región o país en relación a otras regiones y/o países de referencia. El recurso a modelos de diagnóstico cognitivo (Gorin, 2006; Leighton y Gierl, 2007; Rupp y Templin, 2008; Rupp, Templin y Henson, 2010) en la fase de construcción de las pruebas y en el posterior análisis de datos puede proporcionar esa información diagnóstica de utilidad para la práctica en el aula, que relacione los resultados del estudio con el trabajo del profesor.

En resumen, sería deseable que estas macroencuestas educativas proporcionasen una evaluación no solo sumativa sino en alguna medida formativa, una evaluación para el aprendizaje y del aprendizaje. Douglas B. Reeves, presidente del *Center for Performance Assessment*, en su prólogo al excelente manual *Common Formative Assessments* (Ainsworth y Viegut, 2006) dice abiertamente que en EE.UU. "se padece de un exceso de aplicación de pruebas de evaluación pero estamos infraevaluados. Esta distinción es esencial, porque muchos centros escolares están embarcados en evaluaciones sumativas, autopsias educativas que tratan de explicar de qué ha muerto el paciente pero que no proporcionan información de interés para ayudar al paciente a mejorar" (p. ix). En el nuevo escenario que se está gestando en EE.UU. para establecer unos objetivos comunes para preparar a los estudiantes para el acceso a la Universidad y a su carrera profesional (*Common Core State Standards*, firmados ya por 43 estados y con consorcios para trabajar con fondos federales en el diseño de las correspondientes evaluaciones), se incluyen entre otros objetivos el proporcionar información útil para apoyar el trabajo de profesores y directores en la enseñanza, el aprendizaje y la mejora de los programas. Algunos años antes, el comité del *National Research Council* encargado de revisar los avances acaecidos en el campo de la medición y las ciencias cognitivas publicó un informe (Pellegrino, Chudowsky y Glaser, 2001) en el que se presenta un interesante marco de trabajo –el triángulo de evaluación– para poder obtener ese tipo de información. En Navas y Urdaneta (2011) se puede consultar una aplicación de dicho marco al estudio PISA.

La creciente repercusión mediática de estas macroencuestas educativas tiene el efecto positivo de poner temporalmente en primera plana la educación, pero en muchas ocasiones se ofrecen interpretaciones erróneas o indebidas de los datos, tanto por parte de periodistas como de políticos.

Al presentarse los resultados de cada nueva edición de la Prueba de Conocimientos y Destrezas Indispensables realizada por la Consejería de Educación de Madrid, los titulares más habituales suelen ser del tipo "Los peores resultados desde 2005" o "Los resultados no indican ninguna mejoría". Estos titulares están del todo injustificados, ya que no se puede comparar las competencias básicas de los alumnos de este curso académico con las mostradas por los alumnos evaluados en 2005: se ha trabajado con pruebas distintas –que seguramente tendrán un nivel de dificultad diferente– y no se ha utilizado en el proceso de elaboración de las pruebas los mecanismos necesarios para garantizar esa comparación. Puede que sean peores –o no–, sencillamente no hay forma de saberlo dada la forma en la que se han hecho las cosas. Los políticos suelen ir más lejos, al hacer atribuciones de causalidad con los resultados obtenidos (cuando bajan, la oposición lo atribuye a que "cada vez hay menos calidad en el sistema educativo" y el gobierno "al aumento de la dificultad y de los escolares extranjeros en las aulas de Madrid"),

Psicología Educativa
Vol. 18, n.° 1, 2012 - Págs. 15-28
ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

cuando en este tipo de estudios nunca se puede establecer relaciones de causa-efecto sino simplemente de covariación y, en particular, en el estudio citado no se mide nada más que las competencias básicas en Lengua y Matemáticas, por lo que éstas no son correlacionadas con ninguna otra variable.

La presentación de una nueva edición del estudio PISA suele dejar también un buen número de titulares de dudosa justificación. Así, en una carta al director, se valoraba la distancia de 12 puntos de España respecto al promedio de la OECD en la edición 2009 en términos de deplorable mediocridad e indigencia educativa, cuando para interpretar correctamente esa distancia lo que hay que hacer es ponerla en relación con el rango de puntuaciones de los países de la OECD (que es más de 100 puntos) o con otros indicadores estadísticos que señalan que la distancia entre España y la OECD es más bien pequeña y no significativa estadísticamente.

Un ranking que es verdaderamente imposible –pero ofrecido puntualmente todos los años por los medios de comunicación– es el de las universidades en el examen de Selectividad. En este caso ya no es que no se pueda comparar los resultados académicos de un año con el anterior –como sucedía con la prueba de Conocimientos y Destrezas Indispensables– sino que en el mismo año no se puede comparar la tasa de aprobados de los alumnos que se han examinado en distritos universitarios diferentes, al tratarse de exámenes distintos y que no pueden ser equiparados.

Ha llegado la hora de plantearse un mayor nivel de exigencia cuando se realizan pruebas de evaluación a gran escala en cualquier ámbito de aplicación, máxime en uno tan crítico para la vida de miles de estudiantes y, por ende, de la sociedad.

Para corregir esta situación y garantizar la comparabilidad de las calificaciones obtenidas por todos los alumnos presentados a la Selectividad, es preciso trabajar con pruebas estandarizadas elaboradas con todas las garantías por expertos en medición, en estrecha colaboración con profesores de bachillerato, expertos en contenido y autoridades educativas. Estas pruebas deberían ser aplicadas en idénticas condiciones y preferentemente en un llamamiento único a los alumnos de todas las comunidades autónomas en cada convocatoria de la Selectividad. Por su parte, los jueces o calificadores deberían ser formados previamente para poder realizar la tarea de la corrección con las garantías suficientes. Los grandes avances experimentados en las últimas décadas en el campo de la medición educativa permiten disponer de la tecnología necesaria para, partiendo de amplios bancos de preguntas para cada materia a evaluar, construir la prueba que se necesite con arreglo al objetivo marcado y, además, obtener puntuaciones en la misma escala de medida para todos los alumnos.

Wayne J. Camara (2011), ex presidente de la asociación profesional NCME, habla de la dificultad intrínseca de la medición en el ámbito educativo, ya que se trata de medir la conducta de personas y éstas son por naturaleza bastante complejas y en buena medida impredecibles, con un sinfín de factores externos distintos a la calidad de la docencia que pueden influir tanto como ésta en los resultados educativos.

En la misma línea, otro ex presidente de dicha asociación hablaba en su discurso presidencial en la reunión anual 2009 de lo que creía saber acerca de la medición educativa, de lo que en realidad sabía y de lo que creía u opinaba acerca de la misma (Reckase, 2010). Después de reconocer el cambio de paradigma que supuso la introducción de la teoría de respuesta al ítem en los años 60 y de los cambios tan importantes que se han operado en la metodología para diseñar y construir tests, así como a la hora de asignar puntuaciones para estimar el nivel de los sujetos en las características de interés y de avances considerables en campos tan relevantes como la equiparación, el establecimiento de criterios y la evaluación computerizada, reconoce también que la mayoría de los programas de evaluación no utilizan estos avances tecnológicos (en buena medida, porque no resultan fácilmente comprensibles), siguen anclados en la teoría clásica para estimar el rendimiento académico, la calidad de los ítems utilizados sigue siendo manifiestamente mejorable y en el diseño de las pruebas de evaluación continúa siendo necesario trabajar más tanto en la especificación de contenido como técnica, definiendo funciones de información objetivo o distribuciones iniciales de la dificultad y discriminación de los ítems junto a las correspondientes descripciones del contenido y de los procesos cognitivos implicados en la resolución de las tareas planteadas. En un encomiable ejercicio de humildad, Reckase admite saber menos cosas de las que creía saber, dejando al margen todo el bagaje tecnológico que, según él, no es más que "conocimiento sobre un sistema artificial de conocimiento" (p. 3). Lo que sabe con certeza del mundo real –no del mundo paralelo que se construye con los modelos– es que tanto las personas como los ítems (los dos términos básicos de la ecuación que explica la relación entre la actuación observable en los tests y el nivel –inobservable o no directamente aprehensible– de los sujetos en la característica que éstos pretenden medir) son complicados y que los modelos son solo aproximaciones (muy útiles y convenientes) a la realidad, que muchas veces se resiste a ser simplificada en una única puntuación o estimación, proporcionada en ocasiones por la presión ejercida para informar de la manera más sencilla posible.

La dificultad de la tarea nunca puede ser óbice para no acometerla, máxime cuando se dispone de excelentes herramientas y habida cuenta de la importancia de los retornos de la educación y del papel crucial que juega la evaluación en ella. Las reflexiones de Reckase alertan sobre la complejidad de la medición y sobre los peligros de interpretaciones simplistas que desvirtúen los muchos ángulos que caracterizan a la realidad educativa, pero muestran a las claras la utilidad de la medición educativa y la ingente labor que todavía queda por hacer a los profesionales del campo.

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

Extended Summary

The origin of Educational Psychology is closely linked to the administration of standardized tests that require the collaboration of professionals with the necessary training, working on assessments originally meant, above all, for the field of Special Education. The administration of standardized tests continues to be a relevant part of the work of the educational psychologist, and may constitute its most visible identifying marks: it is necessary to evaluate and intervene, and both the evaluation and intervention are based on measurement, which provides objective information, based on which decisions can be made about the subjects. The measurement provides the necessary information to design the best intervention strategy possible, and it also indicates to what degree that strategy has led to the desired results.

Measurement is needed to make decisions about the students, the teaching-learning process, the curriculum, the teachers and the schools. It is possible to distinguish between two quite different types of assessments which respond to different goals and use different procedures to measure or obtain the necessary information: classroom assessments conducted with the purpose of knowing about but, above all, improving the students' learning; and the large-scale assessments that take place periodically in the entire population in question (for example, college entrance tests) or a sample of it (for example, PISA or national assessments of educational progress), with the basic objective of accrediting or accounting for the educational results of the students. Both are essential nowadays.

Assessment is inherent to the teaching practice. Thanks to assessment, the teacher has information about what the student really knows, and can organize the work accordingly in order to facilitate and later document or give proof of his/her learning. It informs students, parents and the corresponding authorities about the students' academic progress; therefore, it is linked to and serves the teaching itself. After reviewing the measurement procedures available to the teacher and the way of choosing the most appropriate one for each case, a reflection is made about the technical preparation and competence of Spanish teachers for carrying out such an important task.

The grade in the college entrance test constitutes at least 40% of the score that finally opens or closes the door to higher education. The system of access to University awards the same value to the grades obtained on the tests carried out in all the universities, a perfectly viable assumption from a technical point of view, but completely unsustainable taking into account the way these tests are designed and administered. Whether a Spanish student passes the entrance test or not depends not only on his or her capacity to successfully handle university studies, but also on the test and the qualifying board that he or she has been assigned to in the university district of his or her region; that is, the grades obtained by students who have taken the test in different geographical areas are not comparable and cannot be equated. The tests are designed locally and for each specific occasion by panels of experts in the corresponding content. The exercises are rated by different judges without sufficiently detailed, clear and fixed scoring criteria, leaving a lot of room for subjectivity. Furthermore, the tests are not subjected to any quality control: nor their reliability or their validity are measured; nor there is an item analysis to establish whether the questions included provide relevant information and function as expected. Many of these problems are shared by the entrance tests to specialized healthcare training in Spain.

However, in recent years the culture of evaluation has taken on great impetus. In fact, title VI of Organic Education Law 2/2006 is entirely dedicated to the evaluation of the educational system, with very ambitious objectives and a scope that goes from the students' results to the evaluation of school inspection and the Departments of Education. The aforementioned law contemplates two assessments to find out the students' basic skills at critical moments in primary and secondary education. These assessments have a formative and orienting nature for the schools, and provide information for the families and the entire educational community. In some autonomous regions the Departments of Education have also promoted regional assessments of basic competencies at other moments in the students' educational journey, and the participation in international assessments is also important (ICCS, INES, PIRLS, PISA, TALIS, TIMSS).

Next, we examine the keys to obtaining sound educational measures: having instruments of adequate metric quality and professionals with the necessary skills to select (or construct) the best instrument for each case and use it appropriately. Many institutions are worried about improving the practice of professionals who work in psychological and educational testing, as well as the training/information provided for their users. This concern has translated into the elaboration of different regulations which, in the form of codes or user guides, include the main principles involved in the correct use of the measurement instruments and the guidelines for a responsible professional practice, from both an ethical and a technical point of view. After a brief review of the regulations, this study focuses on the Standards for Educational and Psychological Testing (AERA, APA and NCME, 1999).

We then reflect on the consequences of bad practice and, more generally, the consequences derived from the use of tests, as potentially relevant information for making decisions related to the administration of tests. Assessment is becoming an instrument for education-

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

al policy making, and, as is clearly indicated in Standard 13.1, "it is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimize potential negative consequences" (p. 145).

In summary, it is necessary to be able to count on professionals with the necessary technical training in the process of administering the tests and later scoring them. Likewise, the people who must interpret them in order to make decisions in the school setting must be duly qualified or receive appropriate help. For this purpose, it is important for the professional associations to have mechanisms for updating knowledge on this issue and the new evaluation tools that come out on the market. In this sense, initiatives like the new APA database (PsycTESTS), and the evaluation of tests recently initiated by the Test Commission of its Spanish counterpart, are excellent examples of steps taken in this direction.

From among the large number of functions performed by the educational psychologist within the school, supporting teachers in dealing with diversity is becoming increasingly important for tailoring the learning-teaching process to the characteristics of each student, as a priority on the necessary path to excellence. New technologies can play an important role here: the computer is a very useful and accessible tool that makes it possible to implement new ways of learning, to monitor and to evaluate student progress and provide information about it. The last edition of the PISA study reveals that the success of educational systems seems to be related to their capacity to be inclusive and the capacity of schools and teachers to serve a diverse school population by offering personalized education.

One of the most important drawbacks of the PISA study is its limited usefulness for reorienting the work of the teacher in the classroom, as it is a study basically focused on comparisons between countries. However, it would be helpful if these types of survey could provide an assessment that was not only summative but also to some degree formative, an evaluation of learning but also for learning.

The paper ends with some reflections about the complexity of measurement and the dangers of simplistic, erroneous or undeserved interpretations of the data. It also discusses the great usefulness of measurement and the large advances made in the last few decades in providing strong technology and excellent work tools to professionals in the field. The next step would be to take advantage of its full potential.

Referencias

Ainsworth, L. y Viegut, D. (2006). *Common formative assessment*. Thoasand Oaks, California: Corwin Press.

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.

American Federation of Teachers, National Council on Measurement in Education y National Education Association (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: Autor.

Anastasi, A. (1987). What test users should know about the interpretation of test scores. Keynote address at Joint Committee on Testing Practices Second Test Publishers Conference, Rockville, Maryland. (citado de Muñiz, 1998b).

Bennett, R. E. (2006). Inexorable and inevitable: The continuing story of technology and assessment. En D. Bartram y R. K. Hambleton (Eds.), *Computer-based testing and the internet: Issues and advances*. Nueva York: Wiley.

Camara, W. J. (2011). The role of NCME and measurement professionals: What role we may choose in policy and accountability. *NCME Newsletter*, *19*, 1-3.

Cattell, J. M. (1890). Mental tests and measurements. *Mind, 15*, 373-380.

DiRanna, K., Osmundson, E., Topps, J., Barakos, L., Gearhart, M., Cerwin, K. y Strang, C. (2008). *Assessment-centered teaching. A reflective practice.* Thoasand Oaks, California: Corwin Press.

Garaigordobil, M. (2009). Papel del psicólogo en los centros educativos. *INFOCOP, 44*, 14-17.

Gorin, J. S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*, 21-35.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation*. Boston: Allyn and Bacon.

Leighton, J. P. y Gierl, M. J. (2007). Why cognitive diagnostic assessment? En J. P. Leighton y M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge: Cambridge University Press.

Martí Recober, M. (1998). *Los sistemas de corrección de las pruebas de Selectividad en España. Análisis y propuestas*. Madrid: CIDE. Memoria de Investigación inédita.

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

McMillan, J. H. (2008). *Assessing essentials for standard-based education*. Thoasand Oaks, California: Corwin Press.

Meliá, J. L. (1990). *Introducción a la medición*. Valencia: Cristóbal Serrano.

Méndez, L. (2011). El psicólogo educativo en España. Algunas propuestas para la reflexión. *Psicología Educativa, 17*, 39-56.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*, 229-258.

Muñiz, J. (1998a). La medición de lo psicológico. *Psicothema, 10*, 1-21.

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012

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

Muñiz, J., Fernández-Hermida, J. R., Fonseca-Pedrero, E., Campillo-Álvarez, A. y Peña-Suárez, E. (2011). Evaluación de tests editados en España. *Papeles del Psicólogo, 32*, 113-128.

Muñoz-Repiso, M., Murillo, F. J., Arrimadas, I., Navarro, R., Díaz Caneja, P., Martín, A. I., ... Fernández, E. (1997). *El sistema de acceso a la universidad en España: Tres estudios para aclarar el debate*. Madrid: CIDE.

Navas, M. J. (2001). La medición de lo psicológico. En M. J. Navas, (Ed.), *Métodos, diseños y técnicas de investigación psicológica*. Madrid: UNED.

Navas, M. J. y García-Forero, C. (2011). *Las competencias básicas de hoy para el mundo de mañana. Informe PISA Canarias 2009*. Consejería de Educación del Gobierno de Canarias.

Navas, M. J. y Urdaneta, E. J. (2011). PISA y el triángulo de la evaluación. *Psicothema, 23*, 701-706.

OECD (2010). *PISA 2009 Results: What makes a school successful?–Resources, policies and practices* (Volume IV). Recuperado el 08 de diciembre de 2010, de http://dx.doi.org/10.1787/9789264091559-en

OECD (2011). *PISA 2009 Results: Students on line: Digital technologies and performance* (Volume VI). Recuperado el 20 de junio de 2011, de http://dx.doi.org/10.1787/9789264112995-en

Outtz J. L. (Ed.) (2010). *Adverse impact. Implications for organizational staffing and high stakes selection*. Nueva York: Routledge.

Padilla, J. L., Gómez, J., Hidalgo, M. D. y Muñiz, J. (2006). La evaluación de las consecuencias del uso de los tests en la teoría de la validez. *Psicothema, 18*, 307-312.

Pellegrino, J. W., Chudowsky, N. y Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Ponsoda, V. (2012). Nuevas tecnologías y medición educativa. *Revista Española de Pedagogía, 251,* 45-60

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

Reckase, M. D. (2010). NCME 2009 presidential address: "What I think I know". *Educational Measurement: Issues and Practice*, 29, 3-7.

Rupp, A. A. y Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state of the art. *Measurement, 6*, 219-262.

Rupp, A. A., Templin, J. L. y Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications*. Nueva York: Guilford Press.

Sanz, J. (1992). *Análisis por asignaturas de las pruebas de acceso a la Universidad*. Madrid: CIDE. Memoria de Investigación inédita.

Savage, L. W. y Ehrlich, P. (Eds.) (1990). *Philosophical and foundational issues in measurement theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Yela, M. (1984). *Apuntes de Psicometría*. Madrid: Universidad Complutense de Madrid.

Manuscrito recibido: 20/09/2011
Revisión recibida: 29/11/2011
Manuscrito aceptado: 02/12/2011

ISSN: 1135-755X - http://dx.doi.org/10.5093/ed2012a2
Artículo publicado Online First: 30/03/2012