Psicothema
2000. Vol. 12, Supl. nº 2, pp. 95-99

ISSN 0214 - 9915 CODEN PSOTEG

La evaluación convencional frente a los nuevos modelos de evaluación auténtica

Amaia Bravo Arteaga y Jorge Fernández del Valle
Universidad de Oviedo

Durante los últimos veinte años, han ido apareciendo nuevas tendencias en evaluación que tratan de responder a las demandas que la evaluación convencional, basada en el uso de tests estandarizados no ha logrado cubrir. Estas demandas han tenido un especial protagonismo en el ámbito educativo, donde el problema de la evaluación del logro escolar ha sido objeto de numerosas investigaciones y publicaciones que ofrecen alternativas al uso de tests de elección múltiple. El término evaluación auténtica agrupa todo este conjunto de alternativas, y se define por oposición a la evaluación estandarizada, a la cual considera no-auténtica e incapaz de detectar el verdadero aprendizaje. Desde este nuevo modelo se reivindica la importancia del contexto, el realismo de las demandas, de la situación instruccional, y un mayor protagonismo del proceso frente a los resultados. A lo largo del trabajo se ofrece una discusión sobre las ventajas, inconvenientes y aplicación de ambos modelos.

Traditional assessment versus authentic assessment. During last years, new trends on evaluation have appeared on educational context. These new alternatives pretend to replace traditional assessment based on standarized tests. In USA, standarized tests have been nearly the only way to assess academic performance during last decades, in fact, an only test was enough to decide educational and labour opportunities of a person. This kind of evaluation has received many criticisms by researchers and teachers, who think this model does not detect real learning. The term authentic assessment gathers some alternatives to assess people by a different way. People are asked for constructing their responses in a significant context. These assessments remark not only result but process too.

La evaluación estandarizada en el ámbito de la Psicología, tiene sus orígenes en las demandas sociales de principios de siglo. La primera de ellas surge en el ámbito educativo, donde al establecerse la enseñanza obligatoria se requería algún método para clasificar a los niños en diferentes niveles. La respuesta a este problema la ofreció Binet con el primer test mental que evaluaba procesos superiores de pensamiento (test Binet-Simon, 1905). Un segundo origen se sitúa en la Primera Guerra Mundial, cuando un grupo de psicólogos de la APA elaboró los primeros tests colectivos de inteligencia (Army Alpha y Army Beta), con el fin de seleccionar entre miles de soldados a los más adecuados para desempeñar diferentes tareas.

Las nuevas demandas de clasificación y selección impulsaron el desarrollo de toda una tradición en la medición psicológica, la tradición psicométrica, basada en tres *principios*: estandarización, diferenciación y el uso de tareas generales. Estos principios son bien reconocidos en los tests de elección múltiple, compuestos por un conjunto de ítems (tareas generales) que según el modelo reflejan el dominio de una habilidad. Instrucciones estándar, un determinado número de ítems, tiempo limitado y obtención de puntuaciones basadas en los resultados, son algunas características de los tests de lápiz y papel.

A principios de los ochenta, los problemas del sistema educativo americano impulsaron la elaboración de una reforma educativa que dio mucha importancia a la evaluación y convirtió a los tests estandarizados en su herramienta estrella (Linn, 1995). Este tipo de evaluación ofrecía importantes *ventajas*, entre las que sobresalen: la gran cantidad de información que permiten recabar en grupos grandes de personas, en poco tiempo y bajo coste económico, además del gran desarrollo de los estudios de validez y fiabilidad de estas pruebas.

Los procedimientos de evaluación estandarizados se convirtieron en la norma para evaluar el logro de los estudiantes y en el único punto de referencia para la toma de decisiones tan importantes como el permitir o no a un alumno pasar de curso, a la universidad u obtener un diploma. En muchos casos la puntuación obtenida en un único test ha sido suficiente para determinar las oportunidades educativas y económicas de una persona. Los efectos personales y sociales negativos que se derivan de estas decisiones están bien documentados en la literatura (Oakes, 1986a; Oakes, 1986b; Jaeger, 1991).

Sin embargo, la reforma de los ochenta no cumplió sus expectativas, las investigaciones mostraron sólo modestas ganancias en habilidades básicas. La creciente insatisfacción fomentó numerosas críticas hacia los métodos de evaluación estandarizados, críti-

Correspondencia: A. Bravo Arteaga
Facultad de Psicología
Universidad de Oviedo
33003 Oviedo (Spain)
E-mail: abravo@correo.cop.es

ca que se fortaleció con los estudios que demostraban el poder que los procedimientos de evaluación tienen sobre el currículum y los estándares educativos (Frederiksen, 1984; Madaus, 1988). La preocupación sobre el impacto de las evaluaciones en lo que aprenden los alumnos, ha dirigido uno de los frentes más importantes contra los métodos de evaluación estandarizados. Sobre este tema, Terence Crooks (1988) concluye: «La evaluación en el aula… guía el juicio de los estudiantes sobre lo que es importante aprender, afecta a su motivación y a la percepción de su propia competencia, estructura su acercamiento al estudio, consolida el aprendizaje y afecta al desarrollo de estrategias de aprendizaje» (p. 467). El mismo impacto se ha señalado en la decisión de los profesores sobre el qué y cómo deben enseñar. Su trabajo se ve dirigido por las evaluaciones desarrolladas desde los cargos administrativos y políticos, sean municipales, regionales o nacionales (Lomax, 1992; Moss, et al. 1992; Nolan, Haladyna & Hass, 1992). Esto sucede en nuestro sistema educativo con la Selectividad, los profesores centran su instrucción en las habilidades y contenidos exigidos en ese único examen, enseñan habilidades específicas para responder mejor y utilizan el mismo formato en sus propias evaluaciones. Las evaluaciones dirigen y limitan el currículum a los objetivos que se reflejan en los tests, por ello, parece necesario evaluar el amplio rango de conocimientos, habilidades e intereses que queremos fomentar en nuestros estudiantes. En este sentido, Wiggins (1989) anima a los educadores a: «evaluar aquellas capacidades y hábitos que creemos esenciales, y evaluarlas en su contexto». En la misma línea Resnick y Resnick (1991) apuntan uno de sus principios: «construye evaluaciones sobre aquello que quieres que los profesores enseñen» (p. 59), la idea es aprovechar el poder que la evaluación tiene sobre la enseñanza, dicha influencia puede convertirse en virtud y no en debilidad.

Recientes investigaciones sobre aprendizaje y cognición consideran al aprendiz, al estudiante, un participante activo en la construcción del conocimiento y en la comprensión, y no un mero receptor de hechos y reglas de proceder. Como dicen Resnick y Resnick (1991), incluso la memorización requiere organización para ser efectiva. Esto cambia el rol del profesor de transmisor de conocimiento a mediador del aprendizaje. Los estudiantes deben participar activamente en el proceso de pensamiento, en la organización y reorganización del conocimiento y en su propia evaluación.

Críticas a la evaluación estandarizada

A la luz de estas preocupaciones, en los últimos años se han multiplicado las experiencias, investigaciones y artículos donde se mencionan otros tipos de evaluación que suelen etiquetarse como «auténticos», «alternativos», «directos», «basados en actuaciones». Todos ellos pretenden superar las críticas dirigidas a los métodos de evaluación estandarizados. A continuación se describen algunas de las *críticas* citadas con mayor frecuencia:

1. *Miden sólo conocimiento declarativo y no procedimental* (Mehrens, 1992). Los ítems suelen referirse al recuerdo de hechos y resultados y rara vez evalúan estrategias o habilidades procedimentales.

2. *Se centran en el resultado y no en el proceso* (Mumford, Baughman, Supinski y Andersen, 1998). No pueden explicar el porqué de las diferentes ejecuciones, ni el proceso por el cual llegan a obtener un resultado correcto o incorrecto.

3. *No cubren adecuadamente el dominio evaluado* (Mehrens, 1992; Wiggins, 1991). Esta crítica, dirigida a la validez de contenido, no se centra exclusivamente en este tipo de medición. Muchas alternativas a los tests siguen sin representar adecuadamente lo que se pretende evaluar.

4. *Existen diferentes habilidades e incluso inteligencias que no son evaluadas por este tipo de tests.* Desde la teoría de Gardner de las inteligencias múltiples, donde incluye junto a las tradicionales inteligencias lingüística, lógico-matemática y espacial, las inteligencias corporal-kinestésica, musical, interpersonal e intrapersonal, se sugiere un sistema de evaluación con medidas más auténticas. También Powell (1990) critica la limitación del contenido de estos tests a las áreas académicas.

5. *Son medidas relativas* (Powell, 1990). Los tests clasifican a las personas dentro del grupo de referencia según la puntuación obtenida, pero en muchas ocasiones el objetivo de la evaluación no es clasificar, sino averiguar lo que ha aprendido el estudiante, o su nivel de dominio en una tarea o campo concreto, independientemente del nivel de su grupo. Además, se debe añadir la dificultad de conseguir una adecuada muestra de referencia.

6. *La estandarización supone una muestra homogénea.* Sin embargo, la muestra evaluada es heterogénea, con diferente dominio del lenguaje, actitudes y valores.

7. *El formato de elección múltiple limita las evaluaciones y supone otras habilidades distintas a las evaluadas* (Powell, 1990). Las inferencias realizadas suponen habilidades subyacentes a las respuestas del test. Además, con frecuencia dependen demasiado de las habilidades verbales de las personas evaluadas o en una única forma de ver el mundo, lo cual perjudica los resultados obtenidos por algunos grupos.

8. *Se alejan de las verdaderas demandas contextuales* (Mumford, Baughman, Supinski y Anderson, 1998). Al buscar simplicidad y estandarización, el diseño de estos tests hace difícil detectar cómo la gente desarrolla habilidades específicas en ambientes complejos, más reales.

Este tipo de evaluación no ofrece a los estudiantes la oportunidad de mostrar sus verdaderas competencias (Johnston, 1987; Newman, 1990). Los objetivos son homogeneizados y no representan adecuadamente los intereses de cada estudiante en su currículum; la puntuación también se centraliza y se basa en criterios establecidos por personas que no conocen a los estudiantes, sus metas, o sus oportunidades de aprendizaje.

Cada vez son más las razones que impulsan a los expertos en el campo de la medición a buscar alternativas. Bajo la precisión, objetividad y ahorro en tiempo y dinero que ofrece el uso de tests estandarizados, se realizan selecciones que reflejan sólo una o pocas dimensiones de las diferencias individuales, no se mide a «toda la persona», cuando es precisamente «toda la persona» la que es contratada para un trabajo o seleccionada para realizar determinada tarea.

Nuevas alternativas: la evaluación auténtica

Con el fin de superar estos problemas han ido apareciendo nuevas tendencias en evaluación bajo la denominación de *evaluación auténtica*, que es definida por sus defensores por oposición a la evaluación tradicional (no-auténtica), a la que culpan de algunos de los problemas del actual sistema educativo, donde no se detecta el verdadero aprendizaje. Este término agrupa todo un conjunto de alternativas a la evaluación tradicional, donde la respuesta no está limitada a la elección de una de las alternativas presentadas y donde el contexto es significativo. La persona evaluada hace, crea

o produce algo durante un tiempo suficiente para poder evaluar el proceso, el resultado o ambos (Messick, 1998). En los tests tradicionales la respuesta era correcta o incorrecta, sin posibilidad de conocer el proceso por el cual esa opción era elegida.

Algunos autores apuntan otros requerimientos para hablar de evaluación auténtica: la tarea debe dejar libertad al examinado; el material estimular no debe estar estandarizado; han de ser las propias personas evaluadas las que elijan el momento de actuar para estar verdaderamente motivados; incluso, y especialmente en el ámbito educativo, el evaluador debe conocer a las personas que está evaluando, sus circunstancias vitales y su historia de ejecución de la tarea (Sackett, 1998).

El nuevo modelo se centra en actuaciones más realistas, siendo su objetivo evaluar en una escala absoluta (no relativa según el grupo de referencia) cómo las personas dominan tareas específicas. Prácticamente cualquier alternativa al test de lápiz y papel entraría dentro del nuevo modelo: respuestas abiertas (construidas) frente a la mera elección de una alternativa; ensayos; realización de tareas que pueden simular el desempeño de un trabajo o ser verdaderas muestras del trabajo que está realizando la persona evaluada (portafolio, en el ámbito educativo).

Todos ellos, son indicadores más convincentes de lo que realmente sabe un estudiante o un candidato para desempeñar un trabajo. Siempre ha existido este tipo de medidas, pero ahora se proponen como un nuevo sistema de evaluación a gran escala. Hasta ahora se buscaban evaluaciones baratas, breves, fáciles de puntuar y objetivas, el nuevo modelo tiene *características* muy diferentes:

1. Se realizan observaciones y registros de la ejecución de tareas pertenecientes a un *dominio específico*, que proporcionan una base para hacer inferencias sobre las personas, sin pretender evaluar habilidades subyacentes.

2. La demanda se asemeja más a una *situación instruccional real,* donde se presenta un problema, pero no alternativas cerradas para resolverlo. La persona evaluada, no sólo tiene que acabar de definir el problema, sino además elaborar su respuesta.

3. Superan la simplicidad de las preguntas de alternativa múltiple, requiriendo que la persona actúe en *situaciones más complejas y reales.*

4. Los resultados son percibidos como *más válidos* por los profesores, representan mejor los verdaderos conocimientos y habilidades del alumno.

5. Pueden examinarse tanto el *proceso como el resultado.*

6. Se observa la *calidad de la ejecución* observada, sin valorar tanto la restrictiva estandarización de otras evaluaciones.

Para poner en práctica este modelo, es necesario superar importantes retos a los que se enfrentan tanto profesores como investigadores: las aulas son pequeñas, no se dispone de los materiales necesarios y las escuelas tienen esquemas de trabajo muy delimitados. Sin embargo, con más o menos dificultades, cada vez son más las experiencias donde la evaluación se aparta de los cánones de la estandarización para reflejar las características del nuevo modelo. Dos ejemplos de su puesta en marcha son: la propuesta de Solano-Flores y Shavelson (1997) para evaluar la realización de una tarea y el uso del portafolio en el ámbito educativo.

Solano-Flores y Shavelson configuran el proceso de evaluación a través de tres componentes y tres dimensiones.

Los *componentes* necesarios son siempre: a) una *tarea* donde se presenta un problema bien contextualizado, cuya solución requiere el uso de materiales concretos que han de ser utilizados por los estudiantes; b) un *medio* para recoger las respuestas de los estudiantes (grabaciones, gráficos que reflejan la solución, redacción de conclusiones…) y c) *un sistema de puntuación,* con el fin de valorar el razonamiento y exactitud de las respuestas.

Las *dimensiones* tienen un carácter metodológico y práctico, y se encuentran muy vinculadas unas a otras: a) el *contenido* hace referencia a aspectos como la adecuada representación del dominio, asegurar que los contenidos sean significativos y comprensibles para las personas evaluadas y que exista una amplia variedad de soluciones de diferente grado de corrección; b) el *equipamiento,* tiene en cuenta los materiales y recursos disponibles; y c) la *puesta en práctica*, alude a la disponibilidad de las condiciones físicas y de tiempo necesarias, la fiabilidad interjueces, el tiempo para realizar la tarea o el entrenamiento con el sistema de puntuación. Son frecuentes las tensiones entre las dimensiones debido a su mutua influencia, por ejemplo, si se reduce el coste de los materiales pueden aumentar los errores de medida, y si estos son de alta calidad el coste puede ser muy alto. Se trata de un proceso cíclico, donde es imposible optimizarlas todas, se debe buscar la combinación que maximiza las ventajas y minimiza los inconvenientes.

El *proceso* comienza identificando las actuaciones que demuestran el dominio de ciertas habilidades y seleccionando las tareas que pueden licitarlas. Decididos los aspectos relativos a cada dimensión, el problema o tarea se presenta a las personas que se van a evaluar, y se van registrando los pasos que realizan durante el ejercicio. Las observaciones o registros son valorados por un grupo de jueces, expertos en el campo, quienes evalúan la expresión de dichas habilidades y todo el proceso de ejecución de la tarea.

Esta propuesta es una opción entre muchas de cómo llevar a la práctica la filosofía de la evaluación auténtica, ahora bien, la alternativa que más atención ha recibido, especialmente en el contexto educativo, es el *portafolio*. Se trata de un método muy útil en este ámbito y que representa un buen ejemplo para explicar algunas características del nuevo modelo. Arter y Spandel (1991), lo definen como el grupo de trabajos realizados intencionalmente por el estudiante, donde se muestran sus logros en una o más áreas. Según Meisels y Steele (1991), el portafolio permite a los estudiantes participar en la evaluación de su propio trabajo, permite seguir mejor la pista de su desarrollo y proporciona una base para realizar una evaluación cualitativa de todos los logros de cada niño.

El portafolio recoge múltiples pruebas de trabajo: redacciones, cuentos, dibujos, libros leídos, vídeos, fotografías, grabaciones,… Idealmente, debería incluir distintos tipos de observaciones (Grace y Shores, 1991):

a) Registros de actividades, hechos e intervenciones espontáneas realizadas por el estudiante, sin ser juzgadas. Dan una idea de su progreso diario.

b) Inventario de objetivos. Es una de las herramientas más útiles para seguir el progreso de los estudiantes, donde se especifican diferentes objetivos educativos, entre ellos los diferentes pasos para la adquisición de habilidades. En general, las observaciones se basan en actividades cotidianas de clase.

c) Escalas de evaluación. Son apropiadas cuando el comportamiento observado tiene diferentes aspectos o componentes.

d) Preguntas y peticiones. Una de las maneras más directas de recoger información es preguntar directamente al alumno sobre su opinión o conocimientos sobre un tema determinado. Demuestra no sólo sus conocimientos sino también sus habilidades lingüísticas.

e) Tests de prueba. Ayudan a identificar las habilidades y conocimientos adquiridos, con el fin de guiar la planificación del proceso de enseñanza.

El portafolio supone un buen ejemplo de lo que se viene llamando evaluación auténtica. Es un método que permite conocer mejor a los estudiantes y sus verdaderos logros, además de ofrecerles una visión más realista de las demandas que recibirán fuera del ámbito escolar. Un estudio de Calfee y Perfumo (1993) sobre la utilización del portafolio y basado en la opinión de los profesores, reveló dos aspectos importantes: por un lado, los profesores reconocen haber introducido el portafolio a su práctica educativa como un intento de renovación personal, impulsados por un nuevo compromiso que ha incrementado su estatus, haciéndoles más responsables de su propia instrucción; por otra parte, reconocen que los fundamentos técnicos de esta evaluación son aún débiles, todavía no está claro cómo medir los logros conseguidos.

Limitaciones de los nuevos modelos

A pesar de todas las ventajas mencionadas, no son pocos los *problemas* asociados a estas alternativas en evaluación:

1. *Son métodos mucho más costosos.* Al mayor coste económico, hay que añadir el tiempo, siendo además mucho menor el número de personas que pueden ser evaluadas simultáneamente.

2. *Dificultad de elaborar evaluaciones paralelas.* A diferencia de los tests, encontrar tareas cuyos requerimientos sean idénticos es muy difícil.

3. *Falta de acuerdo en los constructos que han de evaluarse* en el proceso de resolución de problemas.

4. *El uso de jueces para puntuar la ejecución de tareas supone mayor subjetividad.* La probabilidad de error aumenta y el coste es mayor. Aunque se entrene a los jueces y se les muestren ejemplos de una buena y mala ejecución, los jueces no siempre siguen estas reglas. Sus evaluaciones están sometidas a factores situacionales, características de los examinados y asunciones del estereotipo de lo que es una buena ejecución.

5. *La compleja naturaleza de muchos ejercicios.* Esta complejidad no sólo hace que la evaluación sea más cara y difícil de puntuar, sino que dificulta la obtención de una muestra adecuada de la expresión de estas habilidades. Por otro lado, el tiempo no puede alargarse más de dos horas, para asegurar un buen rendimiento, y esto limita el número de tareas a un máximo de tres o cuatro, lo que difícilmente asegura una adecuada representación de esa habilidad.

6. *El efecto del contexto en la evaluación.* Separar la influencia del contexto de evaluación de la ejecución realizada es muy difícil. Los resultados pueden verse influidos por la presencia de otras personas y en el caso concreto del portafolio, nos encontramos con el problema de cómo averiguar si las muestras presentadas han sido realizadas por el niño.

7. *Generalización de las inferencias.* Al centrarse en habilidades específicas es difícil generalizar a otros dominios. Se requerirían evaluaciones muy amplias, extendidas a diferentes dominios, lo que encarecería el coste y se perdería efectividad.

Los métodos utilizados en este modelo no son novedosos, de hecho responden al tipo de evaluación que se ha realizado dentro del aula durante muchos años, la novedad se encuentra en su aplicación a las evaluaciones a gran escala, y es aquí donde aparecen sus limitaciones. Aspectos como la validez y fiabilidad de las mediciones tienen gran relevancia, lo que es válido dentro del aula, puede no serlo en una evaluación a gran escala. Muchos autores han tratado el problema de la validación de estas mediciones, sin embargo, no sucede lo mismo con la fiabilidad, olvidada por los defensores de la evaluación auténtica.

Tabla 1
Diferencias entre las dos líneas de evaluación (Mumford, Baughman, Supinski y Anderson, 1998)

	EVALUACIÓN TRADICIONAL	EVALUACIÓN AUTÉNTICA
OBJETIVO	• Clasificar a las personas en función de la puntuación alcanzada en el test.	• Evaluar el nivel de desarrollo de la habilidad a través de la actuación manifestada.
MEDIDA	• Tareas muy específicas y concretas (ítems). • Las respuestas señalan habilidades subyacentes. • Puntuación objetiva.	• Tareas más amplias (ej. resolución de un problema). • Las respuestas señalan el dominio de una/s habilidad/es. • Puntuación a partir del juicio de expertos.
INFERENCIAS	• Discrimina a las personas en función del nivel alcanzado en cierta habilidad a partir de las puntuaciones obtenidas.	• Discrimina a las personas en función del dominio demostrado en habilidades específicas.
DESARROLLO	• Especificar los constructos que se pretenden evaluar con el test. • Ítems que maximicen la discriminación entre las personas. • Ítems relacionados entre sí.	• Especificar el tipo de actuación y conocimiento relevantes para cierta habilidad. • Tareas realistas y referidas a un dominio específico.
VENTAJAS	• Coste bajo. • Puntuación objetiva. • Fiabilidad. • Imparcialidad. • Validez predictiva. • Generalización de inferencias. • Validez de constructo.	• Puntuación en escala absoluta según el criterio de referencia (dominio de habilidad). • Recoge la complejidad de las habilidades. • Refleja las diferencias cualitativas. • Buena información diagnóstica. • Alta validez ecológica.
DESVENTAJAS	• Puntuación en escala relativa, según grupo normativo. • No recoge toda la complejidad de las habilidades evaluadas. • Insensible a las diferencias cualitativas. • Limitada información diagnóstica. • No está directamente relacionado con los conocimientos y habilidades del mundo real.	• Coste alto. • Puntuación subjetiva. • Muestra del dominio no representativa. • Efectos del contexto sobre las puntuaciones. • Escasa generalización de las inferencias. • Injusta para las personas pertenecientes a bajo nivel socioeconómico.

Discusión

Cada modelo responde mejor a unos determinados objetivos (tabla 1), mientras la evaluación estandarizada permite evaluar a grandes grupos de personas simultáneamente de forma objetiva, la evaluación auténtica busca mayor profundidad y exhaustividad, centrándose más en el caso individual, a pesar de hacerlo a través de evaluaciones más subjetivas. No tiene sentido sustituir un modelo por otro cuando los dos han demostrado su efectividad en diferentes contextos.

La primera ha resultado de gran utilidad cuando hay que seleccionar a un número de personas dentro de una amplia muestra. De hecho, dentro de la tradición psicométrica se está desarrollando una nueva línea que supera muchos de los inconvenientes de la Teoría Clásica de los Test, se trata de la Teoría de Respuesta a los Ítems. Una de sus aplicaciones, los tests adaptativos informatizados, ha aportado grandes ventajas a la hora de realizar evaluaciones objetivas a un grupo muy amplio con fines de selección o clasificación (pruebas de acceso a la universidad, M.I.R., P.I.R., etc.).

Por otro lado, en contextos como el educativo, la evaluación auténtica va ganando adeptos, lo cual tiene sentido si el verdadero objetivo de la escuela es enseñar y no clasificar a los estudiantes según el nivel alcanzado en un test. Un profesor debe conocer lo que realmente ha aprendido cada uno de sus alumnos, con el fin de desarrollar mejor el proceso de enseñanza, para ello un único examen a fin de curso (o trimestral) no es el mejor indicador. De aquí proviene la defensa de la olvidada evaluación continua mediante métodos como los antes indicados por Grace y Shores (1991), si bien, para ello no vale sólo defender este nuevo modelo, sino adecuar el contexto educativo para poder aplicarlo (número reducido de alumnos, características del aula, contactos profesor-alumno).

No faltan en ambos modelos las ventajas e inconvenientes de su aplicación (tabla 1), aunque el espíritu de renovación de la evaluación auténtica ha provocado una cierta visión romántica de sus posibilidades. Para rechazar la evaluación estandarizada, conviene primero analizar cuál es el objetivo de la evaluación y los recursos disponibles, en palabras de Frechtling (1991): «antes de asumir que disponemos de una alternativa que puede solucionar los problemas a los que nos enfrentamos, debemos estudiar esta nueva herramienta de manera más analítica y menos emocional, y preguntarnos qué puede y no puede hacer».

Referencias

Arter, J. y Spandel, V. (1991). *Using portafolios of student work in instruction and assessment.* Portland, OR: Northwest Regional Educational Laboratory.

Calfee, R.C., y Perfumo, P. (1993). Student portafolios: opportunities for a revolution in assessment. *Journal of Reading, 36,* 532-537.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 85(4),* 438-481.

Frechtling, J.A. (1991). Performance assessment: moonstruck or the real thing?. *Educational Measurement: Issues and Practice,10(4),* 23-25.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 39(4),* 193-202.

Grace, C. y Shores, E.F. (1991). *The portafolio and its use: developmentally appropiate assessment of young children*. Little Rock, AR: Southern Early Childhood Association.

Jaeger, R.M. (1991). Legislative perspectives on statewide testing. *Phi Delta Kappan, 73(3),* 239-242.

Johnston, P. (1987). Assessing the process, and the process of assessment, in the language arts. En J.R. Squire (Ed.) *The dynamics of language learning: research in reading and English.* (pp.335-357). Urbana: ERIC Clearinghouse on Reading and Communication Skills.

Linn, R.L. (1995). High-stakes uses of performance-based assessments. Rationale, examples, and problems of comparability. En Oakland, T., Hambleton, R. K. (Eds.). *International perspectives on academic assessment.* (pp. 49-73). Boston/ Dordrecht/London: Kluwer Academic Publishers.

Lomax, R.G. (1992). Appendix A: Nationwide teacher survey. En G.F. Madaus, M.M. West, M.C. Harmon, R.G. Lomax, y K.A. Viator (Eds.). *The influence of testing on teaching math and science in grades 4-12.* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Public Policy, Boston College.

Madaus, G. (1988). Testing and the curriculum: from compliant servant to dictatorial master. En L. Tanner (Ed.), *Critical issues in curriculum: 87th NSSE yearbook.* (pp. 83-121). Chicago: National Society for the study of Education.

Mehrens, W.A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11(1),* 3-20.

Meisels, S. y Steele, D. (1991). *The early childhood portafolio collection process.* Ann Arbor, MI: Center for Human Growth and Development, University of Michigan.

Messick, S.J. (1998). Alternative modes os assessment, uniform standards of validity. En M.D. Hakel (Ed.) *Beyond multiple choice: evaluation alternatives to traditional testing for selection,* (pp. 59-74) Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Moss, P.A., Beck, J.S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., Herter, R. (1992). Portafolios, accountability, and interpretive approach to validity. *Educational Measurement: Issues and Practice, 11(3),* 12-21.

Mumford, M.D., Baughman, W.A., Supinski, E.P., Anderson, L.E. (1998). A construct approach to skill assessment: procedures for assessing complex cognitive skills. En M.D. Hakel (Ed.) *Beyond multiple choice: evaluation alternatives to traditional testing for selection.* (pp. 75-112). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Newmann, F.M. (1990). Higher order thinking in teaching social studies: a rationale for the assessment of classroom thoughtfulness. *Journal of Curriculum Studies, 22(1),* 41-56.

Nolan, S.B., Haladyna, T.M. y Hass, N.S. (1992). Uses and abuses of achievement test scores. *Educational measurement: Issues and Practice, 11(2),* 9-15.

Oakes, J. (1986a). Beyond tracking. *Educational Horizons, 65(1),* 32-35.

Oakes, J. (1986b). Tracking, inequality, and the rethoric of school reform: why schools don´t change. *Journal of education, 168(1),* 61-80.

Powell, M. (1990). *Performance assessment: panacea or pandora´s box.* Rockville, MD: Montgomery County Public Schools.

Resnick, L.B. y Resnick, D.P. (1991). Assessing the thinking curriculum: New tools for educational reform. En B.G. Gifford y M.C. O´Conner (Eds.) *Changing assessments: alternative views of aptitude, achievement and instruction.* (pp. 37-75). Boston: Kluwer Academic Publishers.

Sackett, P.R. (1998). Performance assessment in education and professional certification: lessons for personnel selection? En M. D. Hakel (Ed.) *Beyond multiple choice: evaluation alternatives to traditional testing for selection,* (cap. 8, pp. 113-129) Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Solano-Flores, G., Shavelson, R.J. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16(3*), 16-25.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 79*, 703-713.

Wiggins, G. (1991). A response to Cizek. *Phi Delta Kappas, 72,* 700-703.