

CLAVES PARA EL DISEÑO DE UN NUEVO MARCO CONCEPTUAL PARA LA MEDICIÓN Y EVALUACIÓN EDUCATIVAS

Joan Mateo¹

Universidad de Barcelona

RESUMEN

En este artículo se analiza en profundidad la evolución que se ha producido en el marco conceptual de la medición y evaluación educativa. Se inicia con una reflexión en torno a las limitaciones más importantes del modelo actual basado fundamentalmente en la psicometría clásica y en la teoría conductista. Se establecen a continuación los cambios más significativos ocurridos en los contextos asociados a la medición y evaluación educativa, para finalmente abordar los elementos constitutivos básicos sobre los que construir un nuevo paradigma.

Palabras clave: Medición, evaluación, nuevos paradigmas en medición y evaluación.

SUMMARY

In this article the evolution produced in the conceptual framework measurement and evaluation of the education has been analysed in depth. It starts with a reflection around the most important limitations of the present model, based on the classic psychometrics and behaviourist theory. Furthermore, it is stated that the most significant changes occurred in the associated contexts to the educational measurement and evaluation. Finally, the basis for a new paradigm are presented.

Key words: measurement, evaluation, new paradigm in measurement and evaluation.

¹ Departamento de Métodos de Investigación y Diagnóstico en Educación. Facultad de Pedagogía. Dirección electrónica: jmateo@ub.edu

INTRODUCCIÓN

Es evidente que en el campo de la medición y evaluación educativa se ha producido un cambio fundamental en su marco conceptual. La manera en la que en la actualidad se genera el discurso en esta área de conocimiento ha evolucionado profundamente incorporando un conjunto de nuevos elementos que apenas gozaban de consideración anteriormente, mientras que otros que habían ocupado un espacio nuclear de forma tradicional se han visto relegados a un papel más complementario.

Así y, en primer lugar, se ha observado un claro desplazamiento de los planteamientos evaluativos basados en exclusiva en los principios psicométricos a otros centrados en la evaluación educativa. Se ha pasado de la cultura de los tests a la cultura de la evaluación. Todo ello ha producido un abandono sistemático del término medición educativa a favor del más comprensivo de evaluación educativa.

Pero, sería injusto reducir el problema a un simple cambio tecnológico, el desplazamiento terminológico responde, por encima de todo, un conjunto de transformaciones de mucho mayor calado, de carácter paradigmático, así se han modificado nuestras concepciones sobre la naturaleza del aprendizaje, del rendimiento y de la evaluación. Nuestras definiciones sobre estos tres elementos son radicalmente distintas de las que se han ido sosteniendo de forma tradicional, también se ha incorporado el factor ético como un factor consustancial a la realidad evaluativa.

Nuestro mensaje a lo largo del artículo ha sido básicamente que la evaluación constituye una parte fundamental del hecho educativo y su uso se justifica en tanto en cuanto optimizamos su impacto sobre la calidad de los aprendizajes. Posiblemente y desde una postura más radical, podríamos decir que las actividades evaluativas no son sino actividades educativas y la distinción entre unas y otras es puramente metodológica o académica, pero en ningún caso establece diferenciación por razón de su naturaleza.

Pretendemos con esta aportación recapitular los elementos constitutivos de lo que podría constituir un nuevo marco para la medición y evaluación educativa. Es evidente que la mayoría de ellos han ido apareciendo a lo largo de la última década, pero posiblemente es necesario significarlos de forma conjunta para establecer la línea argumental y estructural que los aglutina y poder orientar definitivamente nuestro pensamiento y acción hacia la necesidad de diseñar y establecer de forma generalizada un nuevo paradigma para la evaluación educativa.

CRISIS DE UN MODELO: LIMITACIONES INHERENTES AL MISMO

La pretensión clásica de la medición ha consistido históricamente en tratar de cuantificar los atributos que están presentes en objetos o individuos. Con ello se intenta objetivar nuestros juicios valorativos respecto de los mismos y facilitar su estudio y su conocimiento.

Con la medición se intenta hacer comparables los individuos en relación a aquellas propiedades que previamente hemos analizado. La medición empezó a aplicarse a las Humanidades a partir del siglo XIX cuando se intentó introducir este tipo de conocimientos en la misma metáfora en la que ya estaban situadas las denominadas discipli-

nas científicas. Medir era la condición previa necesaria para poder aplicar el método científico a cualquier área específica de conocimiento que pretendiera conseguir estatus científico.

Con todo, la analogía aplicada a las Humanidades, entendiendo que las características medibles de estos conocimientos eran de naturaleza similar a otros saberes como, pongamos por caso la Física, esta hoy en clara crisis y se han ido evidenciando un gran número de limitaciones, algunas de carácter conceptual, otras instrumental y finalmente también de tipo metodológico, que trataremos de presentar brevemente a continuación.

Limitaciones de carácter conceptual

La medición educativa adoptó como propios en sus inicios los principios de la psicometría, sin embargo hemos de considerar que la psicometría como construcción científica se realizó a partir de los trabajos en torno a la inteligencia y a su medición.

La noción que subyacía era que la inteligencia era una característica de naturaleza similar a otras características físicas de las personas y que podía ser también observada y consecuentemente medida.

Su propia naturaleza establece su primera dificultad para trasladarse sin más al entorno educativo. Así desde la psicometría, se considera que los atributos medidos son de carácter fijo, estableciéndose un concepto de la inteligencia muy cristalizado y poco fluido, en contraste con el uso que la medición debería adoptar en el contexto educativo, donde la medición de los aprendizajes debería estar orientada a su propio desarrollo.

La medición de cualquier aprendizaje incide inevitablemente en su proceso de cambio y mejora. La medición adopta por tanto un carácter mucho más dinámico cuando se aplica a la educación.

Otro factor de la psicometría que puede considerarse como una limitación desde la perspectiva educativa, es la necesaria interpretación de las puntuaciones en función de la norma establecida por el propio grupo. Así en el marco de la psicometría clásica las puntuaciones de cualquier sujeto se establecen por comparación con sus pares y por tanto la ejecución del sujeto en la prueba siempre se expresa en términos relativos, renunciando por las características de su propio diseño a establecer ningún tipo de interpretación absoluta del atributo medido.

Las pruebas de referencia normativa se diseñan para producir agrupaciones proporcionales de sujetos con altos, medios y bajos resultados. Desde el momento que los sujetos no tienen el control de las puntuaciones del resto de individuos tampoco lo tiene de sus propios resultados, su aplicación es a todas luces injusta, y sin embargo es un formato que ha sido, y sigue siendo, ampliamente utilizado en educación.

Aún podemos encontrar dos supuestos problemáticos más que se asumen desde el paradigma psicométrico (Berlack, H. et al., 1992; Goldstein, 1992, 1993).

El primero es el supuesto de universalidad, que presupone que la puntuación en una prueba tiene esencialmente el mismo significado para todos los individuos. Así, por ejemplo, bajo esta asunción, pretendemos decir que una puntuación determinada en un test estándar de lectura representa la habilidad individual para leer y que este significado es universalmente aceptado y entendido.

El factor clave que subyace bajo este tipo de argumento es el concepto de «constructo». Un constructo es un término usado para denominar habilidades y atributos subyacentes. Un constructo constituye en definitiva un concepto puramente artificial con el que se pretende explicar la naturaleza de la conducta humana.

Se supone que en el proceso de desarrollo de una prueba, el primer paso consiste en definir con precisión el constructo y posteriormente construir el reactivo que nos permitirá capturar su esencia y medirla. Comprobar la relación entre ambos forma parte de los procesos de validación.

Cualquier proceso de definición supone siempre acotarlo. Generalmente los conceptos en que nos movemos tienen múltiples acepciones y la pretensión de que las pruebas que diseñamos capturan sin fisuras su esencia, se nos antoja como una pretensión francamente desmesurada.

El presupuesto de universalidad constituye en la actualidad, uno de los elementos más discutibles, mucho más si se analiza desde los planteamientos postmodernistas centrados en la preeminencia de lo local frente a lo universal, tan en boga en el pensamiento actual.

Finalmente cabe señalar la otra presunción, la de unidimensionalidad. Bajo este supuesto se asume que los ítems que constituyen un test deberían medir un solo atributo subyacente.

Generalmente se usan modelos estadísticos muy sofisticados para hacer posible el cumplimiento de esta característica. Así se fuerza a una simplificación exagerada de la realidad por tal que la prueba diseñada concentre su acción sobre un solo atributo, eliminando aquellos ítems que en las comprobaciones estadísticas aparecen como de naturaleza distinta al resto.

Han sido muchos los autores que han criticado esta manera artificial de forzar a la unidimensionalidad a atributos que en el caso de la conducta humana son en su mayoría de naturaleza pluridimensional (Goldstein y Wood, 1989; Goldstein 1992 y 1993).

Limitaciones de carácter instrumental

El enfoque clásico de la medición ha basado casi en exclusiva la recogida de la información en la aplicación de tests estandarizados de carácter presuntamente objetivo. Modelo de actuación que incluso en la actualidad sigue siendo el preponderante.

Esta perspectiva tal como se plantea, nos proporciona un tipo de información cuantificada que es fundamentalmente útil para comparar las realizaciones de unos grupos de estudiantes frente a otros y no facilita, en realidad, otros tipos de información también relevante que nos permitan una gestión más eficaz de los juicios extraídos sobre el conocimiento de nuestros alumnos.

Siguiendo a Wolf y Reardon (1996), podríamos resumir las características más básicas del enfoque de medición clásico en las siguientes:

- Tipos de muestras: pruebas de elección múltiple, de emparejamiento, de verdadero o falso, de ordenamiento, de completar frases, etc.
- Juicio evaluativo basado en la recogida de información objetiva e interpretación de las puntuaciones.

- Focalizada fundamentalmente sobre la puntuación del alumno en comparación con la puntuación de los otros alumnos.
- Permite al evaluador sintetizar el conocimiento del alumno mediante una puntuación.
- La medición tiende a ser generalizable.
- Provee la información evaluativa e tal forma que inhibe la acción curricular o instruccional.
- Coloca la evaluación bajo la decisión del profesor u otra fuerza externa al propio alumno.

Es evidente que mediante la aplicación de las estrategias y de los instrumentos del enfoque clásico de la medición, no sabemos a ciencia cierta cómo están aprendiendo nuestros estudiantes y cómo les afecta la aplicación del currículum, aún es más, el tipo de capacidades que ponen en juego no siempre están relacionadas con el tipo de conocimientos o habilidades que tienen verdadera significación en los contextos de realidad.

Limitaciones de carácter metodológico

Los dos pilares metodológicos sobre los que se sustenta la calidad de los procesos de medición son la validez y la fiabilidad. Vamos a continuación a analizarlos desde una perspectiva crítica.

En torno a la validez

Le corresponde a la fiabilidad la categoría de ser el término más controvertido y sin embargo la validez ha sido, de entre los que conforman la medición y evaluación educativa, el término más ampliamente tratado a lo largo de los últimos años y posiblemente Messick (1994, 2004) ha sido el autor que más profusamente ha reflexionado sobre el mismo. En líneas generales, desde su perspectiva, la validez se debe fundamentar sobre dos tipos de aportaciones: en las evidencias que aporta el enfoque de validez de constructo y en las consecuencias sociales que se derivan del uso de las pruebas.

Sin embargo existen un conjunto de autores que consideran necesario expandir aún más el marco conceptual de la validez (Linn, Baker y Dunbar, 1991; Harnish y Mabry, 1993) e instan a los constructores de pruebas a que se sumen a esta iniciativa. Con todo hemos de ser conscientes que aceptar, el principio de la importancia de la validez de constructo es introducir a los técnicos en un proceso que por su propia naturaleza no tiene fin.

Es por esta última razón que cada vez más surgen autores como Shepard que ya desde 1993 sugiere como modelo a seguir, el establecimiento de la validez para cada una de las ocasiones en que se use una prueba en concreto.

Parece evidente que si aceptamos los principios de singularidad de las pruebas de ejecución y de los contextos donde se aplica, parece lógico derivar la necesidad de validarla en cada ocasión, teniendo siempre en cuenta los propósitos específicos que se persiguen.

Los autores actuales sugieren que las críticas sobre la aproximación de Messick no se basan en la idea de que su conceptualización sea errónea, sino en que es, en la práctica inoperable.

Sin embargo, hemos de ser conscientes de las consecuencias que se derivan de pensar que, la alternativa a los principios de Messick reside en la comprobación continua de la validez de una prueba. Supone en definitiva aceptar la imposibilidad de hacerla posible dadas las limitaciones reales que acostumbran a tener los usuarios de las pruebas.

Consecuentemente aún aceptando el principio, parece que lo razonable es compartir las responsabilidades, posiblemente en el siguiente sentido:

- Posiblemente esta parece que pudiera ser la vía intermedia. Al diseñador de la prueba le corresponde establecer con claridad los constructos que subyacen bajo la prueba y disponer el modo apropiado de aplicarla.
- A la administración le toca analizar las consecuencias sociales del uso de la prueba.
- Al editor de la prueba debe presentar de forma clara y honesta la información contextualizada de forma amplia y profunda de manera que permita su interpretación en cada uno de los posibles subcontextos de aplicación.
- Al usuario conocer y comprender la información surgida en los tres apartados anteriores y usarla con precisión y honestidad.

y de compromiso entre aplicar estrictamente los principios de Messick o bien de renunciar a ellos, por inoperables, y pretender comprobar siempre y cada vez «ad hoc» la validez de las pruebas.

En torno a la fiabilidad

Si bien la reconceptualización de la validez la hemos resuelto convirtiendo el tema en una cuestión de priorización y especificación de las responsabilidades que competen al diseñador, al administrador, al editor y al usuario de la prueba, repensar la fiabilidad precisa de una aproximación mucho más radical.

La evaluación educativa no opera bajo la asunción de unidimensionalidad de los constructos que pretende medir, la distribución de las puntuaciones correspondiente a los resultados difícilmente se ajusta a la curva normal, usualmente trabaja con muy pocos elementos y las condiciones de evaluación no son fácilmente estandarizables. Con todo ello pretendemos señalar que la aproximación estadística, usualmente aplicada en la estimación de la fiabilidad de los tests estandarizados es simplemente inadecuada para las pruebas basadas en ejecuciones.

En cualquier caso no podemos seguir asociando los conceptos de «medición precisa» y «puntuación verdadera» a las pruebas basadas en ejecuciones. Necesitamos urgentemente sustituir el término «fiabilidad», en su concepción clásica por otros, posiblemente, y tal como sugieren numerosos autores (Gipps, 2000), deberíamos hablar de la «comparabilidad» como alternativa y base para comprobar la consistencia de una prueba.

Bajo este principio la consistencia entre diferentes mediciones se obtiene presentando a los estudiantes a evaluar las tareas a realizar de la misma manera, interpretando los criterios evaluativos del mismo modo, y evaluando las ejecuciones de los estudiantes bajo los mismos estándares por parte de todos los correctores. Con ello obtenemos resultados que a pesar de efectuarse en contextos diversos pueden ser perfectamente comparables porque mantienen un importante grado de consistencia interna a pesar de los cambios externos introducidos.

Con todo existen numerosos autores que entienden que posiblemente el camino a seguir está en desdramatizar la importancia de algunos factores de calidad de la medición psicométrica como son la fiabilidad y la generalización. Para estos autores debemos alejarnos definitivamente del modelo clásico de medición y avanzar hacia los nuevos modelos de ejecución donde la calidad de su aplicación se basa en la calidad de la ejecución y la honestidad y limpieza del proceso de puntuación, pero en cambio no se consideran cruciales ni la fiabilidad ni la generalización (Moss, 1992, p. 250).

Parece por tanto necesario por un lado repensar el sentido otorgado a los indicadores clásicos de calidad de las pruebas y por otro desarrollar unos nuevos estimadores que sustituyan o complementen los anteriores. Este será el camino que trataremos de realizar en el siguiente apartado.

BASES PARA UN CAMBIO DE PARADIGMA

La Psicometría clásica original se basó en la teoría de la inteligencia para fundamentarse, ante sus insuficiencia en el campo educativo se buscó posteriormente solución en los modelos conductistas y aparecieron todo el universo de las prueba denominadas objetivas, que tampoco solucionaron la mayor. El desarrollo de un nuevo paradigma debería huir claramente de cimentarse en alguna teoría en concreto y más bien perseguir fundamentalmente ayudarnos a comprender los nuevos modelos y procesos de aprendizaje.

Desde el punto de vista histórico Wood (1986) cita el trabajo de Glaser en 1963 sobre las pruebas referidas al criterio, como el momento clave de inicio de la separación respecto de los planteamientos métricos clásicos y el tránsito conceptual de pasar de la medición a la evaluación educativa como concepto nuevo e integrador. Para Wood se inicia en ese momento la construcción de un nuevo paradigma para la medición y evaluación educativa.

En el marco del paradigma emergente se buscaba diseñar las pruebas de forma que la información obtenida permitiera analizar el sujeto por referencia a sí mismo y no por referencia a los demás individuos del grupo y facilitar el uso de los resultados de forma constructiva, identificando las fortalezas y debilidades de los individuos de manera que orientara la acción para su mejora educativa.

Según Wood (1986) el nuevo paradigma se podría definir por su acuerdo con los siguientes principios:

- Trata sobre el rendimiento del sujeto en relación consigo mismo, más que por referencia al rendimiento de otros.
- Busca comprobar la competencia más que la inteligencia.

- Tiene lugar bajo condiciones relativamente incontroladas y consecuentemente no produce datos conductualmente correctos.
- Busca la «mejor ejecución» más que la «típica ejecución»
- Es más efectivo cuando las normas y regulaciones características de los tests estandarizados son más flexibles.
- Permite una mirada más constructiva respecto de la evaluación donde el objetivo es ayudar al sujeto más que sentenciarlo.

En los últimos años los cambios en el paradigma se han hecho evidentes en muchos aspectos, a lo largo de nuestro artículo trataremos de señalarlos, sin embargo vamos a indicar uno inicial que entiendo corresponde citar en este momento.

Nos referimos al cambio muy visible que se ha producido en el paulatino cambio de denominación del área de conocimiento. Hemos pasado de la medición educativa a la evaluación educativa (entendida en términos de «assessment», recordemos que la evaluación tiene otras acepciones en las que no entramos en esta reflexión). Los procesos que antes estaban vinculados con la medición se han ido modificando y traspasado al paradigma de la evaluación, de tal forma que de manera generalizada se usa casi en exclusiva el término evaluación como único e inclusivo.

Gipps (2000), justifica el cambio a partir que mientras la medición implicaba «cuantificación precisa», la evaluación actual no esta totalmente interesada en este tipo de proceso y ha encontrado nuevas vías de capturar la información evaluativa sustantiva y de garantizar su calidad.

Vamos a continuación a presentar y analizar los elementos más significativos del cambio a que hacemos referencia

Cambios en la naturaleza de los aprendizajes

La naturaleza de los aprendizajes ha experimentado un cambio radical en el nuevo contexto conceptual de la medición y evaluación educativa. La calidad de un aprendizaje ya no se basa fundamentalmente en conocer más sobre un contenido concreto, sino en nuestra capacidad de usar holísticamente nuestros conocimientos, y habilidades para solucionar tareas específicas. Hacemos referencia al proceso conocido como desarrollo competencial.

Tal como señala Weinert (2001), adquirir una competencia es algo más que poseer una habilidad respecto de un dominio básico. Implica regulación, monitorización y capacidad de iniciativa en el uso y desarrollo de dicha habilidad.

Desarrollar el uso competencial de un conocimiento o habilidad sobrepasa su mero dominio o la aplicación mecánica y habilidosa del mismo sobre un contexto de realidad. Supone interacción, mediación y gestión entre el conocimiento y la realidad física, social y cultural y actuar con efectividad y eficiencia no tan sólo en la realización de la aplicación sino también en la interpretación del contexto y sus significados.

El desarrollo de esta nueva capacidad supone, de fondo, saber a ciencia cierta lo que sabemos pero simultáneamente y para que se dé de forma plena, exige también tener conciencia de lo que no sabemos, capacidad compleja que en términos más científicos se denomina metacognición.

- El desarrollo de esta nueva competencia asociada a la metacognición exige un importante grado de introspección respecto del propio proceso de cognición y su desarrollo se obtiene, entre otras acciones, mediante la observación y reflexión crítica sobre el propio aprendizaje y el de los otros durante la instrucción, la retroalimentación continuada durante el proceso de aprendizaje por parte del profesor, y la transmisión de estrategias efectivas y eficientes respecto de los procesos de resolución de problemas.
- Lleva asociada capacidades como planificación, monitorización, evaluación y manipulación de los propios procesos cognitivos y de las acciones respecto de tareas específicas.
- Desarrollar la competencia metacognitiva supone ser capaz de orientar inteligentemente la resolución de tareas específicas, lo cual implica (Sternberg, 1996):
 - Desarrollar estrategias para organizar las tareas y los problemas de forma que resulte más sencillo su resolución (organizar la tarea en una estructura más comprensible, dividir un texto en unidades más sencillas, etc.).
 - Aprender a usar eficazmente ayudas o instrumentos que faciliten la comprensión de la tarea que estamos realizando (gráficos, analogías, etc.).
 - Aprender a aplicar los recursos cognitivos que exigen más esfuerzo en tareas realmente significativas y al nivel de dificultad más apropiado.
 - Registrar y evaluar de forma continua todos los progresos alcanzados en la resolución de tareas.
- Aprender a usar el conocimiento sobre el propio conocimiento supone también ser abierto y flexible para ser capaz de incorporar amplias experiencias anteriores en la resolución de problemas (Nuthall, 1999). Requiere también un nivel alto pero realista de auto-confianza.
- El objetivo último del desarrollo de la metacompetencia cognitiva se sitúa en la base del «aprender a aprender». La competencia más asociada a la metacognición es consecuentemente la capacidad para aprender de forma autónoma (Boekaerts, 1999).

En cualquier caso para desarrollar la capacidad competencial se precisa de la interacción entre el conocimiento y la realidad. Hay que graduar simultáneamente la ampliación de nuestro campo cognoscitivo y su aplicación a entornos de realidad cada vez más complejos.

El uso competente de conocimientos y habilidades exige, por un lado, la comprensión profunda de los mismos desde su propia lógica interna, por otro saber leer e interpretar la realidad donde se han de aplicar, por compleja que sea y, finalmente una actitud decidida por incidir significativamente en ella ampliando, como consecuencia, nuestra propia percepción y comprensión de la misma y nuestra capacidad de actuar por transferencia en otros contextos.

Podríamos sintetizar con Borghesi (2005), que la finalidad del desarrollo competencial del conocimiento, es conseguir que las personas conviertan el conocimiento en experiencia y la información en vida.

Parece evidente que desde la Psicometría no vamos a encontrar sistemas para evaluar este tipo de aprendizaje, pero vamos a tratar en el apartado siguiente de señalar los cambios que habrá que inducir en el paradigma para conseguirlo.

Cambios en la lógica de la evaluación

La evaluación educativa de forma histórica había estado centrada en la medición y el control de los resultados. Posteriormente desplazó su preocupación a los procesos de petición de responsabilidades, que añadía, al principio anterior, el implicar a toda la comunidad educativa en la responsabilidad de la consecución de la calidad de los procesos y los resultados educativos.

Es básicamente en la última década que se descubre el enorme potencial de la evaluación como instrumento para gestionar la calidad educativa. Se establece definitivamente la importancia de asociar los procesos evaluativos a los del desarrollo y potenciación de nuestra capacidad para aprender.

Finalmente se produce un nuevo cambio, cuando se vincula la evaluación a la generación de cultura evaluativa. Bajo esta última percepción, se entiende que evaluar supone por encima de todo un proceso de construcción de un tipo específico de conocimiento, el axiológico.

Evaluar supone siempre el acto de establecer el valor de algo y para ello no basta con la mera recogida de información evaluativa, esta debe ser interpretada en contextos en el que también los valores y las actitudes juegan un papel importante.

Evaluar bajo esta nueva acepción, implicará por encima de todo, la creación de una cultura evaluativa donde ubicar adecuadamente esta forma de conocimiento.

La cultura evaluativa es el magma que permite dinamizar la acción de todos los implicados en la mejora de la calidad de cualquier proceso educativo. Renunciar a este principio, supone sumergir la evaluación en procesos burocratizados difícilmente asociables a procesos de mejora.

Obviamente la lógica que rige las estrategias evaluativas será absolutamente distinta según sea nuestra posición respecto del papel que juega la evaluación. No se diseña igual un procedimiento evaluativo orientado al control que a la mejora. Y ya no hablemos como serán de distintos si valoramos el papel que juegan los valores en el mencionado proceso y no aceptamos aproximaciones que no supongan incrementar la cultura evaluativa de todos los implicados.

En todo caso supone una ingenuidad absurda concebir la evaluación como un mero proceso técnico, cuya lógica se aplica de acuerdo a pautas lineales. La evaluación moderna es fundamentalmente una actividad matricial e interactiva que implica culturalmente a los sujetos y a sus contextos.

Cambios de enfoque de la instrumentación evaluativa y las estrategias de evaluación

En el apartado anterior señalaba las limitaciones del enfoque clásico basado en exclusiva la aplicación de pruebas de carácter objetivo que han provocado la aparición de movimientos que enfatizan el uso de metodologías que faciliten la observación directa

del trabajo de los alumnos y de sus habilidades. Este nuevo enfoque es conocido como «evaluación alternativa».

Como acostumbra a pasar cuando aparece un nuevo término, rápidamente aparecen variaciones en su significado y distintas argumentaciones según sea el caso. Así, es bastante usual utilizar como sinónimos los términos evaluación alternativa, auténtica y de ejecución).

Evidentemente, cualquier estrategia que aplique instrumentos de aprehensión del rendimiento distintos de las pruebas clásicas de elección múltiple o de construcción basada en el desarrollo de objetivos predeterminados, se puede considerar propiamente como una formula alternativa al método tradicional.

Con todo Meyer (1992), afirma que la evaluación denominada de ejecución y la denominada como auténtica no son lo mismo. Para que una evaluación pueda ser calificada de auténtica es preciso que los alumnos estén comprometidos en la ejecución de tareas pertenecientes a la vida real. Por su lado esta condición no es en absoluto necesaria en una prueba de ejecución.

En definitiva, la evaluación basada en ejecuciones («performance based assessment») y la auténtica son dos modalidades dentro del denominado enfoque alternativo, en el cual, la evaluación de ejecución requiere que los estudiantes, en un momento determinado, demuestren, construyan, desarrollen un producto o solución a partir de unas definidas condiciones y estándares (Rodríguez, 2000), y la auténtica se basa en la realización de tareas reales.

Por su lado Khattri y Sweet (1996) señalan que adoptar la evaluación de ejecución implica los siguientes pasos a realizar por los estudiantes:

- Estructurar las tareas objetos de evaluación.
- Aplicar una información previa.
- Construir respuestas.
- Explicar el proceso que les ha llevado a una determinada respuesta.

Es evidente que las tareas reales incluyen todas las condiciones contenidas en a evaluación basada en ejecuciones, pero añade la realidad como elemento diferenciador, y es obvio que la evaluación alternativa debería tender a la auténtica; sin embargo, es el todo ingenuo pensar que seremos capaces de generar tantos escenarios de realidad como necesidades evaluativas tengamos.

Resumimos con Wolf y Reardon (1996), las características más básicas del enfoque alternativo:

- Muestras: experimentos de los alumnos, debates, portafolios, productos de los estudiantes.
- Juicio evaluativo basado en la observación, en la subjetividad y en el juicio profesional.
- Focaliza la evaluación de manera individualizada sobre el alumno a la luz de sus propios aprendizajes.
- Habilita al evaluador a crear una historia evaluativa respecto del individuo o grupo.

- La valuación tiende a ser idiosincrásica.
- Provee de información que facilita la acción curricular.
- Permite a los estudiantes a participar en su propia evaluación.

Cambios en las tipologías evaluativas

Debemos a Scriven (1967) la distinción entre evaluación formativa y sumativa. La evaluación formativa es aquella que centra su intervención en los procesos de manera que trata desde su inicio de incidir en la mejora, mientras que se asigna el calificativo de sumativa para la evaluación focalizada sobre los resultados y en la que se persigue fundamentalmente el control de los mismos, reservándose la acción mejoradora para futuras intervenciones.

El valora de la aportación de Scriven, si inscribe en la oportunidad el momento en que la hizo. Una época en que la única lógica existente para aplicar la evaluación, era la sumativa. El enfoque formativo permitió ampliar conceptualmente el sentido y significado de la evaluación.

Sin embargo, con el paso del tiempo, se ha ido pervirtiendo el concepto, de manera que lo que son dos funciones de la evaluación se han convertido en dos tipos de evaluación (Mateo, 2000), que muchos autores presentan incluso como enfrentados.

La evaluación es una acción conceptualmente única que, sin embargo, dependiendo del momento o de la necesidad, adopta funciones diversas (diagnóstica, formativa, sumativa). Lo que le confiere calidad no es el calificativo, sino la oportunidad el momento de aplicación, la adecuación de acuerdo al objetivo reseguído y nuestra capacidad de utilizarla complementariamente.

Ambas funciones se necesitan mutuamente, y responden a un planteamiento totalmente obsoleto el pensarlas, por separado y como excluyentes; así, desconocer el valor formativo de la evaluación sumativa nos parece una incongruencia e ignorar la aportación sumativa de la evaluación formativa un error.

Con todo, desde el cambio de paradigma que preconizamos, creemos que ambos términos deberían ser reconsiderados y situarlos en el contexto más amplio del concepto de evaluación continuada.

La evaluación continuada constituye un nuevo enfoque en el que se diluye la posibilidad de uso excluyente de los dos tipos de evaluación ya mencionados. En el nuevo marco conceptual esta tipología evaluativa representa la coherencia con los nuevos planteamientos respecto de la naturaleza de los contenidos y de las estrategias para evaluarlos (Mateo, J. y Martínez, F. 2006).

Se corre el peligro de pensar que la evaluación continuada no es sino meramente un conjunto de evaluaciones puntuales repartidas de forma consistente a lo largo de la aplicación del currículum. A mi juicio esto no es sino una evaluación aplicada repetidamente de forma puntual, por fascículos, y no responde en absoluto a las necesidades del nuevo paradigma.

En los nuevos modelos de desarrollo competencial de los estudiantes, el currículum se estructura no por unidades temáticas sino por actividades de aprendizaje. Estas facilitan dos tipos de acciones, por un lado favorecen el dialogo inter y multidisciplinar, ayudando a la construcción de un tipo de pensamiento global menos fragmentado

y por otro, estructura la relación de ese conocimiento más holístico en su dimensión aplicada sobre contextos de realidad de complejidad creciente y hábilmente pautada en los programas formativos.

En ese contexto la distinción entre actividades de formación y de evaluación son absolutamente irrelevantes y de carácter puramente académico. Toda actividad debe llevar incorporada un conjunto de estrategias evaluativas que incluyan todo tipo de pruebas (de enfoque tradicional y alternativo), donde la información continuamente recogida se utiliza con fines formativos, sumativos y de mejora de los aprendizajes. Bajo este enfoque holístico no puede haber acción evaluativa que no tenga carácter formativo y no puede haber acción formativa o sumativa que no sea evaluada.

Responde en definitiva a la necesidad de plantearse los procesos evaluativos como simbióticamente unidos a los procesos de aprendizaje, donde se atiende a la complejidad de estos desde una visión plural, multidimensional y comprensiva de la evaluación.

Cambios en la conceptualización de las cualidades métricas básicas: fiabilidad, validez y generalización

Posiblemente el reto más grande que afrontamos cuando hablamos de cambio de paradigma en la medición y evaluación educativa, reside en la necesidad de reconceptualizar el término fiabilidad. Bajo esta necesidad, sin embargo, subyace un cambio importante de nuestra propia cosmovisión, así el modelo psicométrico va asociado fuertemente a la noción de objetividad y consecuentemente se entiende que la habilidad es una propiedad individual que puede ser fiablemente medida y donde la puntuación resultante no se ve afectada por el contexto o las circunstancias de aplicación del test.

Sin embargo es del todo evidente que la realidad no es así. Cualquier ejecución se ve afectada por el contexto, por la motivación del evaluado y por el propio modo de evaluar.

La evaluación no puede seguir siendo considerada como una ciencia exacta y debemos posiblemente de parar de otorgarle este tipo de «estatus». Esta aseveración forma parte de la condición post-moderna que postula la suspensión de la creencia respecto del estatus absoluto del conocimiento científico (Gipps, 1993, Torrance 1993, 95).

De forma parecida desde el paradigma constructivista se indica que nosotros construimos el mundo de acuerdo a nuestros valores y percepciones y desde esta visión es difícil sostener la existencia de lo que en psicometría se denomina «puntuación verdadera».

El cambio producido en el sentido de resituar el modelo psicométrico y colocarlo de forma más periférica y complementaria respecto del modelo educativo es paralelo al efectuado en los modelos experimentales respecto de los naturalísticos.

Desde la nueva posición paradigmática se cuestionan fuertemente los tradicionales criterios de fiabilidad, validez y generalización que son sustituidos por los de integridad y autenticidad (Guba y Lincoln, 1989). En el caso específico de la medición y evaluación educativa se hace cada vez más patente la necesidad de una reconceptualización de los mencionados conceptos.

La integridad se basa a su vez en tres conceptos: la credibilidad, la transferibilidad y la confiabilidad.

La credibilidad en la medición y evaluación puede surgir como resultante de un prolongado compromiso con el proceso evaluativo y fruto de una persistente observación de los hechos objeto de evaluación.

La transferibilidad podría reemplazar la noción de generalización, y hace referencia a que si la prueba de ejecución es llevada a término en un contexto claramente delimitado y especificado, sus resultados pueden ser transferidos (generalizados) a otros contextos de características similares.

La confiabilidad se entiende como el resultado de abrir el proceso evaluativo a un examen minucioso del mismo (Guba y Lincoln sugieren que debería ser auditado en el sentido más estricto del término) y podría sustituir la noción de fiabilidad en su concepción más tradicional.

Finalmente la autenticidad hace referencia a determinar en qué medida la evaluación abarca clara y adecuadamente los constructos que son realmente relevantes en el contexto educativo de interés. Cabe preguntarse siempre, ante una prueba de ejecución determinada, si ésta responde a los intereses reales del contexto educativo e incluye todos los constructos que son relevantes al mismo o simplemente los de interés particular de los responsables de desarrollar y aplicar la prueba.

ELEMENTOS CONSTITUTIVOS BÁSICOS DEL NUEVO MARCO CONCEPTUAL. HACIA UN NUEVO PARADIGMA PARA LA MEDICIÓN Y EVALUACIÓN EDUCATIVA

Vamos par finalizar nuestra reflexión en torno a la necesidad de diseñar un nuevo paradigma para la medición y evaluación educativa, a señalar y desarrollar aunque sea brevemente sus elementos constitutivos más importantes, que a nuestro juicio pasan por reconceptualizar de forma definitiva y generalizada los tres ámbitos siguientes:

- Una nueva definición el término. Hemos de adoptar definitivamente el término evaluación educativa (entendida como «assessment»), como único y comprensivo concepto de los procesos de medición y evaluación. El nuevo concepto de evaluación educativa integra todos los procesos históricamente enmarcados en ambos campos y les dota de valor añadido al hacer corresponder su función principal al desarrollo de los aprendizajes de los estudiantes más que a su descripción objetiva.
- Unos nuevos criterios de calidad. Hemos de superar los conceptos clásicos de validez y fiabilidad enmarcados en el pensamiento psicométrico y manejar nuevas acepciones para gestionar la calidad de la evaluación educativa.
- Incorporación definitiva del factor ético en el pensamiento y en la gestión de la evaluación educativa. Las consecuencias sociales de la evaluación y los valores que se ven implicados en la interpretación de las pruebas evaluativas, no pueden ser ajenos al nuevo paradigma. Es absolutamente necesario interpretar en clave ética todos los elementos que configuran y determinan la acción evaluativa y establecen los factores determinantes de su calidad.

Hacia una nueva definición del término

ACEPCIONES EN TORNO AL CONCEPTO DE EVALUACIÓN EDUCATIVA. EN BUSCA DE UNA DEFINICIÓN COMPRENSIVA DEL TÉRMINO.

Vamos a tratar de presentar, de forma desarrollada, un conjunto de acepciones acerca del sentido de la nueva medición y evaluación educativa.

- La nueva medición y evaluación educativa reconoce el carácter multi-dimensional y complejo de los dominios y los constructos implicados en su acción. Evaluar los rendimientos no es una ciencia exacta y las interacciones que se establecen entre el contexto, las actividades a realizar y el propio estudiante comprometen la posibilidad de generalizar fácilmente los resultados obtenidos para otros contextos o tareas. Estas afirmaciones, perfectamente establecidas en la literatura son, con frecuencia, ignoradas por los generadores de pruebas, posiblemente por considerarlas inconvenientes a sus intereses (Satterly, 1994).
- En la evaluación basada en ejecuciones es necesario establecer previamente los estándares de calidad que se persiguen. Dichos estándares orientan el diseño de la pruebas y ayudan a determinar el nivel de logro deseable de la ejecución.
- El proceso evaluativo debe ser compartido con los estudiantes y es absolutamente imprescindible lograr que sean capaces de auto-monitorizar de forma meta-cognitiva su propio proceso de aprendizaje. La retroalimentación se convierte en el factor clave para lograr que los estudiantes entren en este tipo de proceso.
- La evaluación educativa debe inducir a lo estudiantes a organizar e integrar ideas, a interactuar con los materiales que componen las pruebas de evaluación, a criticar y evaluar la lógica de las argumentaciones y no simplemente a reproducir hechos o ideas.
- Una evaluación de calidad precisa concretarse en tareas que sean también de calidad, no en ejercicios que únicamente sirven para hacer perder el tiempo a los alumnos. Las actividades propuestas en las pruebas deben estar orientadas a comprobar objetivos valiosos concretados en tareas significativas y relevantes.
- La evaluación debe tratar que se base en propuestas que permitan al estudiante mostrar el máximo de su capacidad de ejecución. Ha de tratarse de tareas muy bien concretadas, que estén dentro del ámbito común a la propia experiencia del estudiante, presentadas con claridad y relacionadas con sus intereses y deben realizarse bajo condiciones que no sean percibidas como amenazantes.
- En general los criterios de evaluación deben tener un carácter muy amplio, han de ser más holísticos que los referentes criterios específicos de la prueba. Con esto se persigue dar mayor cobertura a la interpretación de los resultados por parte de los evaluadores, cuando se juzgan especialmente competencias de carácter complejo.
- La evaluación educativa implica operaciones como puntuar y clasificar las ejecuciones e los estudiantes. Si queremos mejorar la consistencia de las puntuaciones entre los evaluadores, es preciso que éstos entiendan perfectamente las categorías de puntuación y los niveles de ejecución asociados a ellas. Todo ello se puede

conseguir mediante ejemplificaciones que deberían estar insertadas en el marco de un esquema general de evaluación. Las ejemplificaciones y los estándares son dos instrumentos muy potentes para que los evaluadores comprendan en profundidad la naturaleza de la competencia que están evaluando.

- En la cultura del «testing» es el número de ítems correctamente contestados, no la calidad general de la respuesta, la que determina la puntuación. En la cultura de la evaluación educativa nos desplazamos de la mera puntuación como forma general para describir la calidad de la ejecución a otras maneras de describir el rendimiento, lo que Wolf (1991, p. 62) denomina «marcos diferenciados de la ejecución de los estudiantes». Con esto significamos la necesidad de asociar a los sistemas de puntuación categorías descriptivas de los niveles de ejecución de los estudiantes a las pruebas propuestas para su evaluación.

Hacia unos nuevos criterios de calidad alternativos

Algunos autores sugieren introducir un conjunto de elementos que, a nuestro juicio podrían actuar de forma complementaria a los ya clásicos, una vez reconceptualizados en el sentido que le hemos conferido a lo largo del artículo, y que podrían añadirse a ellos para asegurar la calidad de la medición y evaluación educativa. Citamos entre otros los siguientes:

- Fidelidad curricular.- El currículo constituye un criterio alternativo para valorar la propia evaluación cuando esta está asociada a un diseño curricular concreto. La noción de fidelidad curricular se puede alinear perfectamente con el concepto de validez de constructo: concretamente se puede interpretar la idea de sub-representación con la de una deficiente fidelidad curricular.
- Confiabilidad.- La confiabilidad puede considerarse como un concepto situado en la intersección entre la fiabilidad y la validez (William, 1993), diríamos que un evaluación es confiable, es decir merece nuestra confianza, en la medida que es fiable y válida.
Harlen (1994), define la calidad evaluativa como la provisión de información de alta validez y de óptima fiabilidad adecuada a un propósito particular y a un contexto específico.
En relación con la evaluación basada en ejecuciones, se debe asegurar la comparabilidad, en el sentido de asociar debidamente la evaluación a los criterios adecuados, proveer del debido entrenamiento a los evaluadores para hacer más consistentes sus observaciones y sus entrevistas y evitar al máximo la tendencia a valerse de estereotipos en la emisión de juicios por parte de profesores y evaluadores.
- Credibilidad pública. Este concepto debe diferenciarse del término credibilidad acuñado por Guba y Lincoln. Se entiende que en evaluaciones orientadas a la petición de responsabilidades y con alta incidencia en la confección de las políticas educativas, el que el global del proceso evaluativo goce de la aceptación y del máximo respeto público es absolutamente esencial.

- Descripción del contexto.- Si pretendemos conseguir la transferibilidad de los resultados evaluativos, y abordar de esta manera una cierta generalización de los mismos, es absolutamente imprescindible basarnos en descripciones precisas y detalladas de los contextos de actuación, que es lo que nos permitirá decidir si es posible transferir los resultados de una realidad a otra, ya que estaremos en condiciones de juzgar su grado de similitud y la posibilidad de extender las conclusiones.
- Equidad.- Se requiere que en cualquier evaluación el rango de elementos que usemos para establecer nuestros juicios sea tan amplio que permita contemplar la diversidad de alumnos y circunstancias y elaborar interpretaciones que sean justas y honestas respecto de sus ejecuciones.

Concretando la lista de nuevos indicadores de calidad de la evaluación y medición educativa queda resumida en:

Fidelidad curricular, que implica que el constructo, el dominio o el currículo están bien especificados y ampliamente cubiertos por el proceso evaluador.

Comparabilidad, que se alcanza mediante la consistencia en la aproximación de los evaluadores a la evaluación y una comprensión profunda y compartida de los criterios evaluativos.

Todo ello es posible mediante una sabia combinación de, entrenamiento de los evaluadores, prudencia general y la provisión de muchas y variadas ejemplificaciones.

Confiabilidad, que emerge a partir de la evidencia proporcionada por la fidelidad curricular, la consistencia y la comparabilidad.

Credibilidad pública, que surge de la aplicación continuada de la evaluación, bajo condiciones de revisión pública de procesos y resultados.

Descripción del contexto, que requiere que una detallada información del contexto donde se ha llevado a término la evaluación está a disposición de los expertos para poder analizar las posibilidades de transferencia de resultados, conclusiones y de la propia instrumentación a otros contextos similares.

Equidad, que requiere que los factores de análisis para la emisión de juicios contemplen la diversidad de los sujetos y sus circunstancias.

Con todo estos criterios de calidad juntamente con las reconceptualizaciones propuestas para la fiabilidad y la generalización no han sido realmente llevados a la práctica de forma sistemática y continuada, seguimos usando de forma mimética e indebida los modelos clásicos, que aplicados a la nueva realidad evaluativa deforman la información cuando se aplican indebidamente.

Confiamos que algunas de las reflexiones aquí introducidas servirán para iniciar una nueva singladura en la creación de un nuevo marco conceptual para la medición y evaluación educativa. Sabemos que hay una nueva generación de evaluadores comprometida en ello y esperamos que nuestra comunidad universitaria sea capaz de introducirse con decisión en la nueva corriente que se está generando.

El factor ético como factor constitutivo básico del nuevo paradigma

ELEMENTOS ÉTICOS BÁSICOS. CLAVES INTERPRETATIVAS PARA EL NUEVO PARADIGMA.

Conviene ser conscientes que la medición y evaluación constituyen, también, un instrumento muy poderoso de acción sobre la realidad educativa. Configura el currículo, el estilo de enseñanza y el de aprendizaje, afecta la visión que tienen de sí mismos los estudiantes y como consecuencia de las tomas de decisiones que se derivan del proceso evaluativo con gran frecuencia en el mundo académico (certificación, selección) afecta también el cómo los estudiantes se ven unos a otros. En otro orden de cosas controla el acceso a la educación superior y determina la orientación profesional de muchas personas.

Es por todo ello que finalizar este artículo dedicado a reflexionar respecto del nuevo marco conceptual de la medición y evaluación educativa, sin hacer referencia a los elementos éticos y los riesgos implícitos que implica el hecho evaluativo, nos parecía una frivolidad.

Validez consecuente

Messick (1989, 2004) en su ya clásico capítulo dedicado a la validez, argumenta que las consecuencias sociales de la evaluación y los valores que se ven implicados en la interpretación de las pruebas son aspectos integrales de la validez. Las concepciones actuales sobre la validez mantienen una estructura unitaria y por tanto todos los aspectos que la configuran, incluyendo la relevancia, la utilidad, los valores implicados y las consecuencias sociales, son aspectos fundamentales a considerar cuando se explora la validez de constructo.

Cuando se diseña y aplica una prueba, los responsables deben advertir a los usuarios, no tan sólo de las bondades de la misma, sino de los peligros que se derivan de su mal uso. Se deben describir profusamente los constructos objetos de medición, pero también como se deben interpretar y usar las puntuaciones obtenidas y las potenciales consecuencias de todo tipo que se derivarían de optar por usos alternativos de las mismas.

En sociedades tan diversas como las nuestras, donde un importante contingente de estudiantes tiene lenguas nativas distintas de la vehicular escolar, que provienen de culturas alejadas, con valores diferenciados, etc., la interpretación de los resultados debe hacerse a la luz de las nuevas situaciones. Los signos de la postmodernidad son la pluralidad y la multipertenencia y ello nos obliga a tratamientos complejos de la información evaluativa y a aplicar estrategias amplias e integrales para establecer la validez de una prueba.

Equidad

Otro de los factores que se puede ver gravemente comprometido cuando aplicamos una prueba, es garantizar un trato equitativo a las personas afectadas. La situación con la que tenemos que lidiar es la de aplicar un reactivo diseñado bajo unas circunstancias y dirigido a un tipo de población, cuando posteriormente podemos topar (cada vez más) con realidades donde la misma prueba se aplica a una población muy heterogénea, con biografías educativas muy diversas, procedentes de contextos sociales y culturales muy diferenciados. Evidentemente bajo estas condiciones garantizar un trato equitativo a través de la evaluación resulta difícilmente sostenible.

Cada vez más en la construcción de pruebas debe recurrirse no tan sólo a expertos en construcción de pruebas, sino también a mediadores socioculturales que traten de analizar las pruebas propuestas para ver en que pueden resultar ininteligibles a los estudiantes por razón de su diferencia específica.

Por otro lado, los resultados obtenidos deben también analizarse con inteligencia y de forma muy comprensiva y tratando de explorar o identificar los posibles sesgos y establecer las correcciones más oportunas. La interpretación debe hacerse desde visiones muy integrales y cabe actuar con enorme prudencia, muy especialmente, cuando, de la evaluación efectuada, se derivan consecuencias difícilmente reversibles para los estudiantes.

En general se aconseja como elemento fundamental para suavizar los posibles efectos negativos para la equidad, el generar estrategias evaluativas en las que los estudiantes tengan múltiples oportunidades para alcanzar los estándares de calidad establecidos y generar una importante diversidad de caminos para lograrlo. Ofrecer la posibilidad de rectificar favorece el aprendizaje, cada oportunidad evaluativa es un elemento dinamizador de los recursos personales y la multiplicidad de vías para acceder a un objetivo ayuda a que cada estudiante pueda personalizar su propio itinerario.

Los resultados

Otro elemento que debe analizarse en clave ética es el que hace referencia al uso de los resultados. Los resultados de la medición y evaluación educativa se usan para un amplio rango de situaciones, así se pueden focalizar para decisiones personales respecto de los estudiantes o bien para diseñar políticas educativas, en cualquier caso dependiendo como se presentan pueden introducir elementos de dudoso carácter ético.

Fomentar políticas de tratamiento de la información evaluativa que pueda ser humillante para sujetos determinados o ciertos colectivos. O bien presentar datos que afectan a la totalidad del sistema y que de una lectura superficial puedan derivarse situaciones injustas, constituyen graves irresponsabilidades en las que no debemos entrar o colaborar.

La discreción y la prudencia han de ser elementos básicos en la manera de actuar de los evaluadores. El respeto a la persona y a su derecho de privacidad ha de ser la guía de nuestra actuación ante nuestros estudiantes y cuando la información afecta al conjunto de la sociedad, es del todo obligado el diseño de políticas de comunicación que gradúe el tipo de información que se ofrece a los diferentes colectivos afectados

por los resultados, con profusión de elementos clarificadores de la misma para generar visiones integrales y contextualizadas.

Evitar la superficialidad y la frivolidad en el tratamiento y difusión de la información evaluativa ha de constituir los fundamentos de la actuación de los profesionales de la evaluación.

A MODO DE SÍNTESIS

Vamos de forma breve a recapitular los elementos más significativos de nuestra aportación con el ánimo de establecer el hilo conductor argumental y facilitar con ello una visión global y comprensiva de la misma.

En primer lugar hemos reflexionado sobre el cambio producido en la denominación de la medición y evaluación educativa, que ha sido paulatinamente sustituido por el de evaluación educativa, dotándole de un carácter inclusivo y en el que quedan comprendidos todos los principios de la medición y evaluación.

Con todo y tal como señalábamos el cambio no es simplemente nominal ni coyuntural, responde al hecho que de fondo se ha ido produciendo una transformación continuada en el paradigma que servía de base a la medición y evaluación educativa clásica hasta trasladarse hacia territorios donde aparecen y se estabilizan un conjunto de principios que responden a una cosmovisión radicalmente distinta a la anterior.

Hemos justificado la necesidad del cambio basándonos en las limitaciones de carácter conceptual, instrumental y metodológico en las que se encuentra inmerso el paradigma tradicional y hemos identificado los factores de cambio que deberían orientar el sentido de las nuevas propuestas.

Los factores que más profusamente han cambiado y que son los que están forzando la necesidad de un nuevo paradigma son los cambios producidos en la naturaleza de los nuevos aprendizajes, en la lógica de la evaluación, en la instrumentación y estrategias evaluativas, en las tipologías evaluativas y en la conceptualización de las cualidades métricas de las pruebas.

Finalmente abordamos los elementos constitutivos básicos que deberían configurar el nuevo marco conceptual que en la medida que se desarrollen y se compruebe su bondad y consistencia teórica, avalarían el relevo del paradigma. Con este objeto redefinimos el término y señalamos sus acepciones más relevantes, presentamos los nuevos criterios de calidad alternativos a los clásicos conceptos de calidad métrica y finalmente incorporamos y justificamos la presencia del factor ético como factor constitutivo fundamental del nuevo paradigma.

REFERENCIAS BIBLIOGRÁFICAS

- Berlack, H., Newmann, F., Adams, E., Archibald, D., Burgess, T., Raven, J., y Romper, T. (1992) *Towards a New Science of Educational Testing and Assessment*, New York: State University Press.
- Boekaerts, M. (1999). Self regulated learning. *International Journal of Educational Research*, 31, 443-457.

- Borghesi, M. (2005). *El sujeto ausente. Educación y escuela entre el nihilismo y la memoria*. Madrid: Fondo Editorial.
- Gipps, C. (1993). *Reliability, validity and manageability in large scale performance assessment*. Ponencia presentada en la conferencia AERA, Abril, San Francisco.
- Gipps, C. (2000). *Beyond Testing*. Londres: The Falmer Press.
- Glaser, R. (1963). Instructional technology and the measureme of learning outcomes: Some questions. *American Psychologist*, 18, 519-521.
- Glaser, R. (1990). Toward new models for assessment. *International Journal of Educational Research*, 14, 5, 475-83.
- Goldstein, H. (1992). *Recontextualising Mental Measurement*. Londres: ICRA Research Working Paper, ULIE. (Publicado posteriormente en *Educational Measurement: Issues and Practice*, 1994, 13(1)).
- Goldstein, H. (1993). Assessing group differences. *Oxford Review of Education*, 10 (2), 141-150.
- Goldstein, H. y Word, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, 41, 139-167.
- Guba, E. y Lincoln, Y. (1989). *Fourth Generation Evaluation*. Londres: Sage.
- Harlen, W. (Ed.) (1994). *Enhancing Quality in Assessment*. (BERA Policy Task Group on Assessment). Londres: Paul Chapman Publishers.
- Harnisch, D. y Mabry, L. (1993). Issues in the development and valuation of alternative assessments. *Journal of Curriculo Studies*, 25 (2), 179-187.
- Khattari, N. y Sweet, D. (1996). Assessment Reform: Promises and Challenges. En M.B. Kane y R. Mitchell (Eds.). *Implementing Performance Assessment*, pp. 1-21. Mahwah (NJ): Lawrence Erlbaum.
- Linn, R.L., Baker, E. y Dunbar, S. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20 (8), 15-21.
- Mateo, J. (2000). *La evaluación educativa, su práctica y otras metáforas*. Barcelona: Horsori.
- Mateo, J. y Martínez, F. (2006), *Medición y evaluación educativa*. Madrid: La Muralla (en prensa).
- Messick, S. (1989). Meaning and values in test evaluation: The science and ethics of assessment. *Educational Researcher*, 18(2), 5-11.
- Messick, S. (1994). The Interplay of Evidence and Consequences in the validation of Performance Assessments. *Education Researcher*, 23(2), 13-23.
- Messick, S. (2004), «Validity» en Linn, R. (Ed.) *Educational Measurement*. American Council of Education, Washington, Macmillan.
- Meyer, C.A. (1992). What's the Difference Between Authentic and Performance Assessment. *Educational Leadership*, (Mayo), 39-40.
- Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62(3), 229-258.
- Nuthall, G. (1999). Learning how to learn: the evolution of students' mind through the soial processes and cultura of the classroom. *International Journal of Educational Research*, 31(3), 139-156.
- Rodríguez, S. (2000). La evaluación del aprendizaje de los estudiantes. Ponencia presentada al I Congreso Internacional: *Docencia Universitaria e Innovación*. Barcelona: Instituts de Ciències de l'Educació (UAB, UB, UPC).

- Satterly, D. (1994). The quality of external assessment. En W. Harlen (Ed.). *Enhancing Quality in Assessment*. Londres: Paul Chapman Publishers.
- Scriven, M.S. (1967). The methodology of Evaluation. En *Perspectives of Curriculum Evaluation (AERA Monograph Series on Curriculum Evaluation, n° 1)*. Chicago: Rand McNally.
- Shepard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.
- Sternberg, R.J. (1996). *Successful intelligence*. Nueva York: Simon & Schuster.
- Torrance, H. (1993). *Assessment, curriculum and theories of learning: Some thoughts on assessment and postmodernism*. Ponencia presentada en el seminario ESCR/BERA, Junio, Liverpool.
- Torrance, H. (1995). *Evaluating authentic assessment: problems and possibilities in new approaches to assessment*. Philadelphia (NJ): Open University Press.
- Weinert, F.E. (2001). Concept of Competence: A conceptual clarification. En D. Rychen y L. Salganik (Eds.). *Defining and selecting key competences*. (pp. 45-66). Kirkland: Huber-Hogrefe Publishers.
- Wiliam, D. (1993). *Reconceptualising validity, dependability and reliability for National Curriculum Assessment*. Ponencia presentada en la conferencia BERA, Junio, Liverpool.
- Wolf, A. (1993). *Assessment Issues and Problems in a Criterion-Based System*. Londres: Further Education Unit.
- Wolf, D.P. y Reardon, S.F. (1996). Acces to Excellence through New Forms of Student Assessment. En J.B. Baron y D.P. Wolf (Eds.) *Performance Based Student Assessment: Challenges and Possibilities*. Ninety-fifth Yearbook of the National Society of Education, Part 1. Chicaho, IL: University of Chicago Press.
- Wood, R. (1986). The agenda for educational measurement. En D. Nuttall (Ed.). *Assessing Educational Achievement*. Londres: The Falmer Press.