Revista Iberoamericana de Evaluación Educativa

ISSN: 1989-0397

La Efectividad y la Eficacia de las Mediciones Estandarizadas y de las Evaluaciones en Educación

Juan Enrique Froemel

Revista Iberoamericana de Evaluación Educativa 2009 - Volumen 2, Número 1

http://www.rinace.net/riee/numeros/vol2-num1/art1.pdf

Fecha de recepción: 15 de marzo de 2009
Fecha de comunicación de dictamen: 20 de abril de 2009
Fecha de aceptación: 20 de abril de 2009

1. Introducción: el auge actual de la medición y de la evaluación del rendimiento escolar

Entre los elementos más difundidos a la vez que debatidos, en la educación formal de niños y jóvenes, en los últimos veinte años, en el mundo, se cuentan la evaluación del aprendizaje escolar y asimismo uno de sus instrumentos fundamentales, la medición estandarizada. Si bien la mayoría de los investigadores y actores en el tema les reconocen su gravitación, no todos están de acuerdo en la real dimensión de su aporte al proceso educativo y lo que es más, variadas percepciones se han construido respecto de ambos, no siempre basadas en bases reales ni menos científicas, generando así mitos y prejuicios.

En este sentido lo que señalara Bloom (1972) acerca de la educación en general, en uno de sus emblemáticos artículos, es aplicable hoy en día a la evaluación y a la medición estandarizada, en cuanto a que "…*seguimos siendo seducidos por el equivalente de los remedios de curanderos, las falsas curas del cáncer, la invención del movimiento perpetuo y las supersticiones…*". Es así como junto al creciente uso de la medición estandarizada, han surgido legiones de "expertos" en el tema que prometen, por medio de enfoques de medición y evaluación más "humanistas" y "educativos", ir mucho más allá de lo que los métodos basados en la teoría psicométrica son capaces de lograr, desacreditando de paso a estos últimos por ser supuestamente rígidos y estrechos. Lo anterior pese a que existe una masa importante de investigación y de desarrollo teórico, ambos rigurosos, sobre el tema.

El daño real y potencial a la educación y a sus sujetos, los alumnos, es por ello, de tal magnitud, que para colaborar en su prevención, uno de los propósitos centrales de este artículo es intentar precisar y explorar el valor real que, para el proceso educativo formal, tienen la medición y la evaluación de raíz científica. Esto ha implicado centrar la mirada en unos aspectos que no constituyendo una visión exhaustiva, sí tocan aspectos críticos del tema en cuestión. Es así como este trabajo no pretende convertirse en una visión integral acerca de la situación actual de la medición estandarizada en educación, sino que más bien aportar algunas pistas para señalar dónde fijar la mirada.

Aunque mucho se habla hoy en día acerca de medición estandarizada en educación, se ha estimado conveniente precisar aquí, primeramente, qué es lo que hace a un instrumento que sea considerado o no estandarizado.

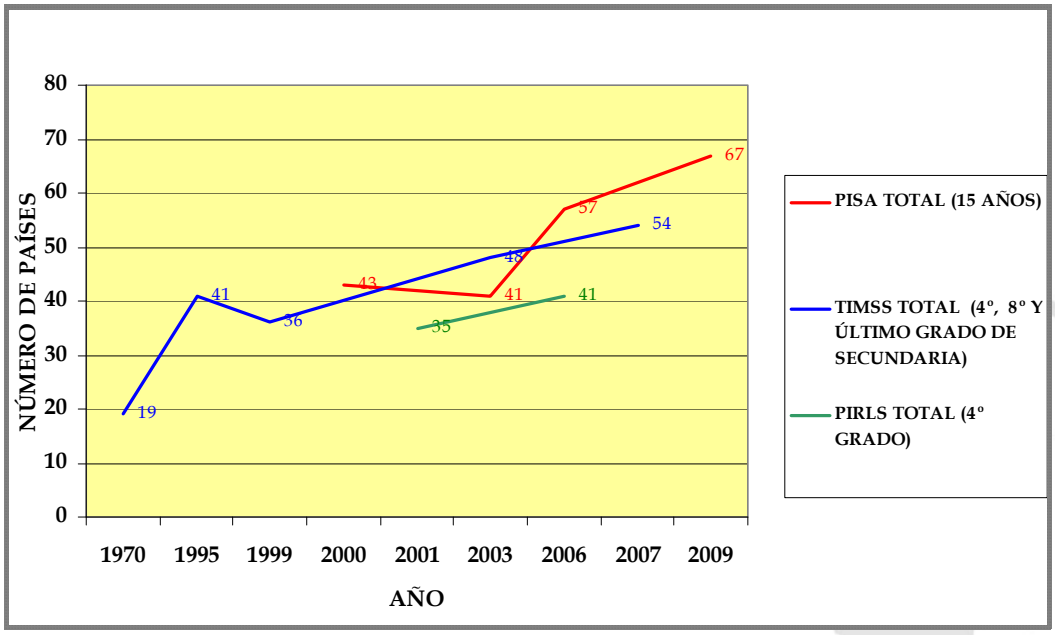
Esta condición implica el cumplimiento de varias características, entre las cuales se cuentan, en primer lugar, el hecho de que los ítems deban ser los mismos y aplicados en un formato equivalente a todos los examinados; en segundo lugar, que las condiciones de administración deban ser también comunes y que, por ende, las instrucciones de aplicación deban ser exactamente a la letra, las mismas, excepto en el caso de las llamadas "acomodaciones" usadas para estudiantes con limitaciones visuales, auditivas o motoras; en tercer lugar, que el procedimiento y proceso de revisión y asignación de puntajes sea también uniforme, no existiendo posibilidad de discrecionalidad en los criterios a aplicar; y finalmente, que la forma en que los resultados son informados también haya de ser común y uniforme. Como es posible apreciar, las condiciones que hacen a un instrumento estandarizado, por una parte, proveen una condición necesaria, aunque no suficiente, para que los resultados sean comparables y por la otra, excluyen de la pertenencia a esta categoría a muchos instrumentos, como es el caso de las pruebas construidas por el profesor o hacen bastante más difícil la estructuración de otros, como ser los *portafolios* si se espera que sus resultados sean comparables.

Una buena forma de visualizar la gravitación de la evaluación y la medición estandarizada es refiriéndose a los estudios internacionales del rendimiento escolar, dado que estos constituyen una de las expresiones actuales más representativas de las dimensiones y evolución de aquellas.

Aunque desde sus orígenes la medición en educación ha estado marcada por una procura incesante de la posibilidad de comparar validamente sus resultados entre sujetos, escuelas y países, la que se ha plasmado en el desarrollo de la condición estandarizada de los mismos, el auge mundial de esta forma de medición es asunto de no más de veinte años. Sus expresiones previas, aun las más antiguas con casi un siglo de existencia, estuvieron restringidas casi exclusivamente a los Estados Unidos, surgiendo inicialmente en el campo de la psicología y trasladándose luego al de la educación.

Las primeras acciones para internacionalizar la evaluación estandarizada surgieron en la década del setenta con el primer estudio realizado por la Asociación Internacional para la Evaluación del Rendimiento Educacional (IEA)[1] siendo ésta entidad la que llevó a cabo la primera serie de estudios internacionales formales en la áreas de ciencia, literatura, comprensión lectora, Inglés y Francés como lenguas extranjeras y educación cívica, involucrando, en ese entonces, a diecinueve (19) países de todo el mundo (Comber y Keeves 1973).

FIGURA 1. NÚMERO DE PAÍSES PARTICIPANTES EN TRES ESTUDIOS INTERNACIONALES DEL RENDIMIENTO ESCOLAR, AGRUPADOS POR AÑOS



Como es posible apreciar en la Figura 1, la evolución en la participación de los países en algunos de los principales estudios internacionales de evaluación del rendimiento escolar ha crecido de manera apreciable desde la primera experiencia antes mencionada. En el caso del Programa Internacional de Rendimiento Estudiantil (PISA)[2], coordinado por la OCDE[3], en un lapso de nueve años se observa un incremento en el número de países participantes cercano al sesenta por ciento. En la serie del programa

[1] International Association for the Evaluation of Educational Achievement.
[2] Programme for International Student Achievement.
[3] Organización de Cooperación y Desarrollo Económico.

denominado Estudio de Tendencias en Matemática y Ciencias (TIMSS)[4], desarrollado por la IEA y que se inició en 1970 con el denominado Primer Estudio de la IEA, antes mencionado, se aprecia un crecimiento en el número de países participantes de un ciento ochenta por ciento, en treinta y siete años. Finalmente, en el caso del Estudio Internacional de Progreso en Lectura (PIRLS), también a cargo de la IEA, el número de países participantes ha crecido en casi un veinte por ciento, en sólo cinco años.

Sin embargo como no siempre la demanda por un producto y su eficacia van juntas, cabe en este caso la pregunta: ¿puede el aumento en la participación nacional en estos estudios interpretarse como un aporte concreto al avance del aprendizaje de los alumnos -fin último de todo proceso que se concibe como educativo-? La respuesta a la pregunta recién planteada pasa antes por la de esta otra interrogante: ¿es la evaluación estandarizada en educación, en primer lugar, efectiva, -ello es cumple con los requisitos para lograr lo que se supone debe lograr- y, en segundo, de ser así, es ella eficaz, -lo logra en la realidad?

2. ¿Por qué eficacia y efectividad?

Dos de los principales requisitos que establecen el valor de cualquier acción son, en primer lugar, el que ella sea lo que dice ser, en otras palabras si es o no **efectiva** y si acaso como tal cumple con su propósito, en otros términos si es o no **eficaz**. De paso, es importante señalar que de hecho, la **efectividad** condiciona la **eficacia**.

Sin embargo, hay pocos términos en el campo de la educación cuyo sentido sea confundido con mayor frecuencia que el de estos dos. Se habla, por ejemplo, de "escuelas efectivas" cuando se quiere hacer referencia a "escuelas eficaces" y todo ello debido a una errónea traducción desde el idioma Inglés, del cual nos llegan la mayoría de las expresiones técnicas en educación. En efecto, tendemos a confundir los términos, ya que por una parte, la traducción correcta del término inglés "effectiveness" al castellano es eficacia -lo distintivo de aquello que cumple con su propósito- y por la otra, el término "efectivo" en castellano describe a aquello que es real.

3. Efectividad y Eficacia versus Validez y Confiabilidad

Son precisamente la efectividad y la eficacia los elementos que se ubican en la raíz misma de gran parte del debate acerca de la medición y la evaluación estandarizadas. Avanzando en este razonamiento si transferimos estos términos, provenientes del campo de la planificación al de la psicometría, nos encontramos con que la **efectividad** se relaciona mayoritariamente con la **validez**, condición que "grosso-modo" califica si un instrumento mide lo que pretende medir y **eficacia** se traduce primordialmente en **confiabilidad**, ello es define el grado en que tal instrumento mide adecuadamente. De forma equivalente a lo señalado anteriormente, en psicometría la **validez** condiciona la **confiabilidad**.

De los dos conceptos señalados, aquel al que este artículo pretende dedicar mayor atención es al de la validez. Existe hoy un debate acerca de si la definición originaria y más propia de esta cualidad de los instrumentos en educación, se enfocaba en un solo ámbito (validez de contenido) (Lissitz y Samuelsen, 2007) o si esa definición implicaba alcances más amplios (validez de constructo y validez concurrente)

[4] Trends in Mathematics and Science Study.

(Moss, 2007). En términos concretos, si acaso al hablar de validez se aludía primariamente a la consistencia entre el instrumento y el mapa de los contenidos y/o competencias por medir que se hubiesen identificado como relevantes o si ella también incluía con igual importancia, entre otras, a la concordancia entre lo que esa herramienta mide y una estructura teórica pre-definida y/o con lo que miden otros instrumentos.

Es preciso señalar, primeramente, que más allá de las diferencias antes señaladas, no existen límites absolutos entre los diferentes ámbitos de la validez y que estos en la realidad se superponen. En segundo lugar, es importante dejar claro que la validez es una variable continua e incluso, las más de las veces imposible de cuantificar con precisión a diferencia de la confiabilidad y en tercero que no existen instrumentos totalmente válidos o inválidos, sino que estos pueden alcanzar diversos grados de validez no siempre nítidos. Finalmente, como lo señala una publicación conjunta de tres entidades profesionales norteamericanas (AERA[5], APA[6] y NCME[7], 1989) citada por Moss (2007), es más propio hablar de **roles** de los variados tipos de evidencia de la validez o de formas de validación, que de **tipos** de validez propiamente tal.

Sin perjuicio de lo recién señalado y más allá de propugnar aquí la adhesión a una concepción unitaria y focalizada o desagregada y abierta de la validez, aunque éste autor se pronuncia más bien por la primera de ellas, se plantea a continuación un intento, de definición de los posibles tipos más básicos de evidencia en este sentido. Es así como en este aspecto es posible reconocer los siguientes ámbitos:

- La **validación por "constructo"**[8] tiene lugar "…cuando un investigador sustenta que un instrumento refleja una particular estructura teórica o "constructo", a los cuales se atribuyen ciertos significados. La interpretación propuesta genera hipótesis verificables, las que constituyen medios de confirmar o rechazar el planteamiento (contenido en el "constructo")…" (Cronbach y Meehl, 1955, citado por Moss, 2007). Un ejemplo concreto corresponde al caso en que se pretende evaluar conocimiento de matemática con pruebas de lápiz y papel, en niveles de grado o curso donde la comprensión lectora está aún en estado de consolidación y por ello es válido sustentar que el "constructo" que mide la prueba no es el conocimiento de matemática sino que una mezcla de éste con la comprensión lectora.
- La **validación por contenidos** es aquella que se lleva a cabo verificando la medida en que una prueba "…constituye una muestra representativa de la clase de situaciones o materias acerca de las cuales se deben extraer conclusiones…" (American Psychological Association, 1966, citado por Moss, 2007). Ejemplo específico, también negativo y típico, es aquel en que los contenidos curriculares del instrumento no corresponden a la tabla de especificaciones y la prueba ya sea, mide lo que no está en la tabla o no mide lo que sí está.
- La **validación por concurrencia** verifica la medida en que hay coincidencia "cuando se comparan los resultados de una prueba con una o más variables externas que se considera que proveen una medida directa de la característica o conducta en cuestión" (American Psychological Association, 1954, citado por Moss, 2007).

[5] American Educational Research Association (Asociación [Norte] Americana de Investigación Educacional.
[6] American Psychological Association (Asociación [Norte] Americana de Psicología.
[7] National Council on Measurement in Education (Consejo Nacional [de EEUU] para la Medición en Educación).
[8] El término inglés "construct" no tiene una traducción precisa en castellano, siendo la que más se acercaría a su sentido conceptual la de "estructura teórica", por lo cual en este artículo se ha optado por utilizar el anglicismo "constructo" puesto entre comillas.

A estas definiciones primigenias, hay que agregar lo expresado en una versión de los estándares para la medición en psicología y educación de AERA, APA y NCME (1999), en la cual se menciona que, al fin de cuentas, "…existen cinco categorías de evidencia acerca de la validez, a ser consideradas al construir un argumento …" "el contenido del instrumento, los procesos de respuesta, la estructura interna, las relaciones con otras variables y las consecuencias de la (aplicación de la) prueba". Es posible sustentar que éste planteamiento coincide con lo señalado por Messick (1989) acerca de que "La validez es un juicio integrado acerca del grado en el cual la evidencia empírica y las concepciones teóricas respaldan a las *inferencias* y *acciones* basadas en los resultados de las pruebas y otros modos de medición, en cuanto que (aquellas) sean *adecuadas* y *apropiadas* (cursivas en el texto original).

Estos últimos dos alcances tienen gran importancia ya que extienden significativamente la gama de evidencias que son necesarias para respaldar la validez, incluyendo aspectos externos tales como las consecuencias de la aplicación de los resultados de los instrumentos. Un reciente artículo de Nichols y Williams (2009) explora este tema en mayor profundidad aún, discutiendo temas tales como la atribución de responsabilidad a los distintos actores involucrados en la medición, por de los distintos tipos de evidencia relacionados con las pruebas en educación.

Una consecuencia práctica muy importante de destacar acerca de la validez, de acuerdo a lo indicado en el párrafo anterior, es que hoy en día se considera que ésta cruza una muy amplia gama de aspectos relacionados con la prueba a la que califica, los que van desde la fidelidad de la misma al mapa de contenidos o competencias que la originó, hasta los efectos que sus resultados pueden generar, pasando por combinaciones de ellos. Un buen ejemplo es el caso de las pruebas de base curricular, como las del estudio TIMSS que, aún cuando su estructura pueda mostrar una alta relación promedio con los currículos de la mayoría de los países, ésta suscitará cuestionamientos en términos de su validez para medir los rendimientos en países en los cuales no existe un currículo único, sino que la docencia se basa en estándares más ligados con competencias. En casos como éste la conveniencia de adherir a un estudio como PISA, basado en estándares de validez universal, aparecería como más plausible.

4. Intentando describir el "iceberg" de la medición en educación

Seguidamente, la exploración en mayor profundidad del tema de la validez de los instrumentos nos lleva a constatar una realidad que por más que pueda no ser del agrado de los especialistas en medición, no tiene mayor posibilidad de ser desmentida y más aún, es preferible que sea reconocida y explicada para contribuir a desvirtuar críticas en contra de la medición estandarizada en educación. Tales críticas son producto, por una parte, de que los especialistas, a veces, en nuestro entusiasmo por los avances psicométricos perdemos de vista las limitaciones de nuestro oficio y le atribuimos mayores capacidades de las que realmente tiene y por la otra, por cuanto muchos sujetos afectados por resultados adversos en estas mediciones tratan de desvirtuar tales resultados desacreditando las herramientas que los originaron.

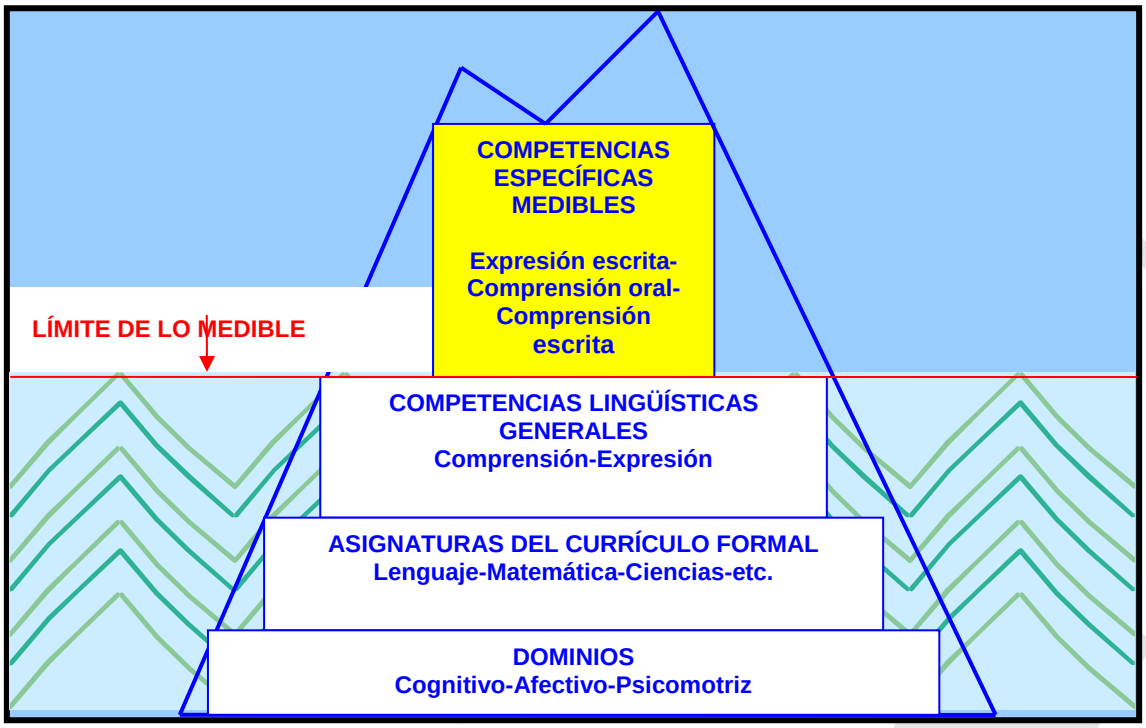
El caso es que las mediciones estandarizadas en educación están limitadas por diversos factores, siendo uno muy gravitante el que surge precisamente de su restringida validez relativa, producto de su naturaleza intrínseca de constituir un intento indirecto para detectar características internas de los sujetos a través de una muestra de evidencias externas. Es importante señalar que si bien este efecto es posible de atenuarse no logra, como podría pensarse, ser totalmente corregido por medio de un uso más intenso

de la tecnología, como podría ser el caso del CA[9] y del CAT[10], dado que su limitación es fundamental y reside, como se dijo antes, en que es un mecanismo indirecto y externo para detectar rasgos internos. Sin embargo, pese a sus limitaciones, hasta hoy no existen métodos científicamente aceptables que hagan un mejor trabajo, para estos efectos, que la medición estandarizada.

Es por esta razón que la proporción del total del aprendizaje, que realmente es posible medir con las evaluaciones estandarizadas corresponde sólo a la cima del "iceberg", que es aquella parte que se ubica por sobre el límite que separa a las características mensurables de forma estandarizada de las que no lo son, que al igual que en el caso de los témpanos que flotan en los mares fríos, corresponde sólo a una parte menor de su volumen total. Esto equivale, en el caso de la educación, también a una expresión minoritaria de lo que realmente ella debió haber logrado en los estudiantes.

Un análisis descriptivo relativamente simple permitirá apreciar esto más claramente si se toma como ejemplo la medición estandarizada de los aprendizajes en Lenguaje. Si se considera la educación formal, ésta abarca al menos tres dominios de competencias, mostrados en el nivel de la base de la Figura 2, por lo cual todo aprendizaje, realmente educativo, debe incluir a estos tres dominios aunque en proporciones variables. Es por ello que al utilizar la medición estandarizada se está capturando información de los dominios (cuando es posible hacerlo) de manera artificialmente segregada. Esto implica que en el ejemplo del Lenguaje, la medición estandarizada habitualmente sólo captura los aprendizajes cognitivos.

FIGURA2. EL "ICEBERG" DE LA MEDICIÓN EN EDUCACIÓN. AQUELLO QUE ES REALMENTE POSIBLE DE SER MEDIDO EN LENGUAJE, POR MEDIO DE PRUEBAS ESTANDARIZADAS ORDINARIAS



[9] CA. Computerized testing (prueba computarizada). Situación en que la prueba es administrada a través de un computador, la que permite el uso de una mayor variedad de estímulos y por ello la medición de competencias más complejas, fundamentalmente usando técnicas de simulación.
[10] CAT. Computerized adaptive testing (prueba adaptativa computarizada). Situación que incluye todas las ventajas del CA, pero que además posibilita adaptar las preguntas subsiguientes en base de las respuestas a las previas de cada sujeto que responde, logrando así una mayor coincidencia entre la prueba y las habilidades del examinado y con ello una medición más precisa.

Ascendiendo un nivel desde la base se aprecia que dentro del espectro cognitivo el aprendizaje involucra a varias asignaturas las cuales, si bien como todo aprendizaje humano también interactúan, son las más de las veces abordadas de manera también separada, por lo cual el Lenguaje como asignatura representa también sólo una fracción, esta vez del aprendizaje cognitivo formal total.

Yendo a un nivel adicional más arriba y entrando en el área curricular del Lenguaje se aprecia que existen al menos dos tipos de competencias generales que debieran ser parte de cualquier diseño curricular en la asignatura: la comprensión y la expresión. Y ambos debieran medirse si la intención es verificar los aprendizajes en esta área curricular.

Sin embargo, al abordar estas dos competencias básicas es posible comprobar que las competencias específicas posibles de medir de forma estandarizada son solamente las siguientes: la expresión escrita, a través de preguntas de respuesta abierta; la comprensión oral, por medio de ítems cerrados o abiertos; y la comprensión escrita, por estos mismos medios. Aunque algunos autores (Rodriguez, 2002) hacen depender la elección del tipo de ítem de aspectos tales como las recomendaciones de la comunidad de especialistas, el costo e incluso de las repercusiones del uso de un tipo u otro en términos de la política educacional, ella siempre dependerá antes que nada del nivel de complejidad de las competencias que se busque medir. Las más de las veces se opta, debido a razones de costo, por escoger sólo ítems cerrados, con la consiguiente reducción en la validez. Similarmente, la expresión oral se deja fuera debido a la complejidad y costo, si es se quiere garantizar su objetividad, de medirla en forma estandarizada.

En conclusión, de tres dominios se mide uno, sucediendo lo mismo con el espectro de asignaturas incluidas y dentro de la única abordada, de sus dos competencias generales se logra medir parcialmente una y de forma relativamente integral la otra.

Si bien es posible complementar la medición estandarizada con otros enfoques como la observación y los portafolios, estos presentan limitaciones a la hora de tratar de comparar sus resultados entre alumnos de distintas escuelas y aulas, ya que no constituyen, de suyo, métodos estandarizados y si bien es posible aproximarlos a tal condición, ello implica, una vez más, un aumento sustancial del tiempo y del costo involucrados.

5. Medición y evaluación ¿distintos términos para un solo concepto?

Para contribuir a develar la complejidad del tema abordado, otro aspecto que es importante precisar es que dos términos que en su ámbito se usan regularmente de manera indistinta no son, en rigor, sinónimos. Se trata de **medición** y **evaluación**.

Si bien existen múltiples definiciones de ambos, una forma directa de percibir sus diferencias es considerando que la medición consiste en el acto de establecer la posición de un sujeto en la escala de una variable determinada y por lo tanto constituye una visión instantánea y, por ello, estática, de su situación respecto de esa variable. Evaluación, por su parte, implica determinar la trayectoria que, en el tiempo, muestran las posiciones de un sujeto en la escala de una misma variable, por lo cual implica una visión dinámica del comportamiento que tal sujeto ha tenido en esa variable, obtenida por medio de la comparación de los resultados de dos o más mediciones de ese mismo sujeto.

Por lo recién señalado, de acuerdo a esta visión de ambos procesos, la medición constituye un insumo de la evaluación y su calidad determina en gran medida la de ésta. En este último sentido aunque los

métodos de comparación de las varias mediciones sean óptimos, si la precisión, objetividad, validez y confiabilidad de éstas no satisfacen estándares aceptables, el valor de la evaluación será cuestionable.

6. Herramientas para diferentes fines: referencia a normas o a criterios

Los primeros instrumentos de medición estandarizada surgieron de la necesidad de clasificar u ordenar sujetos entre sí y se denominan pruebas **referidas a normas**. Una forma simple de visualizar el mecanismo a través del cual éstas operan es considerar que lo que hacen es estructurar la distribución de los resultados obtenidos por los sujetos en un instrumento, de forma relativa y luego establecer el ordenamiento de estos últimos de manera acorde. Es importante precisar que en este enfoque la distribución de los resultados necesariamente mostrará proporciones de sujetos en las distintas categorías de resultados coincidentes con una distribución típica, como es la normal, ello es que porcentajes menores de los resultados corresponderán a sujetos con muy altos o muy bajos puntajes y que la mayor proporción de los resultados de los examinados se ubicarán en las áreas de puntajes intermedios. Un ejemplo de este tipo de instrumentos son las pruebas de selección para el ingreso a la educación superior, conocidas, en los Estados Unidos, como Scholastic Aptitude Test (SAT)[11].

El segundo enfoque, de desarrollo posterior al anterior, incluye instrumentos cuyo propósito es determinar la posición de los sujetos respecto de un referente externo, absoluto y pre-establecido, no considerando especialmente relevante la posición relativa de los mismos. En esencia lo que interesa establecer, en este caso, es cuáles sujetos pasan una determinada valla, criterio o puntaje, para establecer si son o no competentes de manera absoluta en un aspecto determinado. Este tipo de instrumentos se denominan **referidos a criterios** y en su desarrollo e interpretación ha contribuido significativamente la denominada Teoría de Repuesta al Ítem (TRI) [12]. Las pruebas de rendimiento corresponden a este tipo de instrumentos y un buen ejemplo lo constituyen las usadas por los estudios internacionales mencionados al comienzo de este artículo. Es importante señalar que la distribución de los resultados de este tipo de pruebas no tiene por qué corresponder a la normal y más aún, es esperable que no lo haga, ya que si se considera que la educación es un proceso de crecimiento equitativo para todos los estudiantes, sería esperable que su aprendizaje no se distribuyera normalmente.

7. ¿Pueden los mismos instrumentos cumplir una diversidad de roles?

La necesidad que existe, en muchos países, de reducir los costos de procesos de suyo onerosos y de grandes dimensiones, como son los de medición y evaluación, como asimismo la conveniencia de disminuir la frecuencia con que los estudiantes son sometidos a mediciones -baste considerar que al menos tienen que enfrentar las del docente, las del sistema nacional y las internacionales- ha llevado a intentar en algunos casos la utilización de los mismos instrumentos para ordenar a los sujetos y para determinar su nivel de aprendizaje. Un ejemplo fue el Sistema de Ingreso a la Educación Superior (SIES), en Chile, diferido en su aplicación desde hace algunos años.

[11] En castellano Prueba de Aptitud Académica.
[12] La sigla TRI en castellano deriva de la inglesa "IRT", la que corresponde a Item Response Theory.

Como señalara Kaplan (1963) "el puntaje de una prueba puede entregar interpretaciones relativas o absolutas." Por su parte, abordando el tema desde una perspectiva sutilmente diferente de la utilizada en la sección anterior, Haladyna (2002), agrega que "la interpretación absoluta es denominada usualmente *referida a criterios* (RC) debido a que nos indica cuánto de una habilidad o conocimiento ha adquirido una persona". "La interpretación relativa es llamada *referida a normas* (RN), por cuanto esta interpretación permite la comparación de un puntaje con otros puntajes en la escala de una prueba determinada. Las interpretaciones RN nos indican cuan diferentes unos de otros son los puntajes."

Citando nuevamente a Haladyna (2002), es posible señalar "que aunque RC y RN corresponden a tipos de interpretaciones de los puntajes de una prueba, en sentido estricto no tenemos pruebas RC y RN. Sin embargo, los términos RC y RN se asocian a menudo con pruebas que permiten tales interpretaciones." Y agrega que "Irónicamente cualquier prueba puede aportar interpretaciones NR, pero las interpretaciones CR provienen (sólo) de pruebas diseñadas para informarnos cuánto ha aprendido un estudiante."

Si bien ésta es, como se dijo antes, una forma leve pero fundamentalmente distinta para la conceptualización de las diferencias entre los enfoques referidos a normas y a criterios, en esencia la conclusión es la misma en cuanto a que ambos tipos tienen propósitos distintos. El punto siguiente abordar es la aclaración de si acaso tal diferencia descalifica a un tipo para cumplir el rol del otro, o si es posible pensar en una figura híbrida.

Por una parte, la afirmación antes citada acerca de que sólo las pruebas CR pueden aportar información acerca del aprendizaje real de los estudiantes ya indica que, por definición y construcción, éstas son capaces de aportar elementos que las NR no son, por lo cual éste ya constituye un elemento de juicio a considerar en cuanto a que las pruebas NR no están en condiciones de aportar a la medición de los aprendizajes respecto de criterios absolutos.

Por otra parte, cada uno de estos tipos de interpretaciones enfatiza ciertas características que hacen que en la construcción de las pruebas mismas se privilegien ciertos aspectos en desmedro de otros. Tal es el caso de que las pruebas CR tienden a poner el acento en la validez y las NR en la discriminación.

El caso es que, además, los dos tipos de herramientas descritas en la sección anterior se basan en modelos de interpretación diferentes, descansan en supuestos estadísticos distintos (como es el de la normalidad de la distribución de los puntajes en el caso de las NR y no necesariamente en el de las CR) y finalmente entregan información diferente.

En conclusión y en una comparación que puede ser considerada trivial, la utilización indistinta de cualquier de los tipos para cumplir ambos propósitos -jerarquización y verificación del aprendizajes de los sujetos, medidos- puede tener el mismo efecto que emplear un coche de ciudad a campo traviesa, caso en el cual es posible que éste logre efectuar el recorrido, pero con serios riesgo para su integridad y la de sus tripulantes.

8. Formas de expresar puntajes y facilitar su comparabilidad

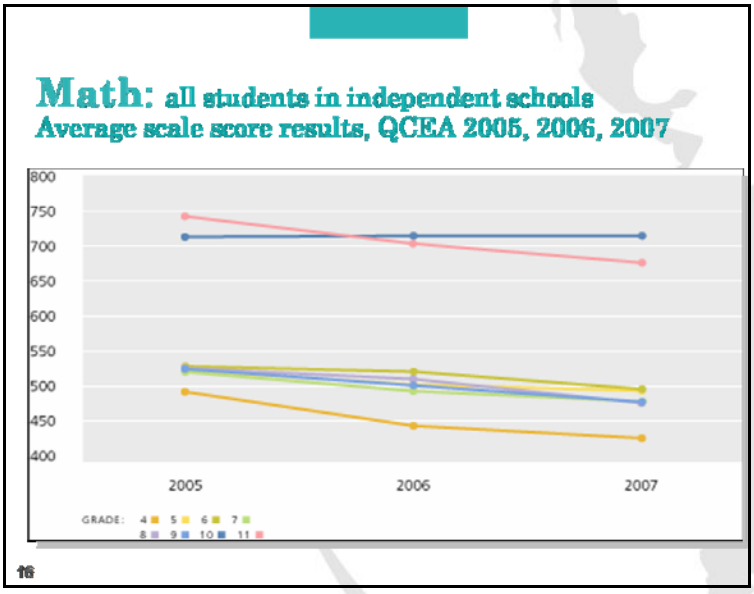
El aspecto que se busca explorar en esta sección es la facilidad relativa que implica la comparación de resultados expresados, ya sea de forma cuantitativa o cualitativa, sin entrar aún al tema de la comparabilidad estadística de los mismos.

Como se ha mencionado repetidamente en este artículo uno de los aspectos que más ha preocupado y sigue preocupando hoy en día, tanto a los especialistas en medición como a los educadores en general, es la posibilidad de comparar los resultados de las mediciones estandarizadas, posibilitando de esta forma una verdadera evaluación de sujetos, aulas, escuelas o países.

Estas comparaciones, las más de las veces basadas en el cálculo de diferencias entre los resultados de dos o más mediciones, pueden ubicarse a distintos niveles de agregación. La forma de comparación puede ser longitudinal, implicando que los resultados de los mismos sujetos son contrastados en distintos momentos. En otros casos las comparaciones pueden ser transversales, lo que significa que se busca comparar a distintos sujetos, en los mismos niveles de grado, en distintos momentos. Existe también la posibilidad de comparar sujetos en distintos niveles de grado, aunque ello requiere del cumplimiento de determinadas condiciones psicométricas. Finalmente, el método del Valor Agregado, consistente en la comparación de un valor de rendimiento predicho y uno efectivo, puede también aplicarse en estas contrastaciones.

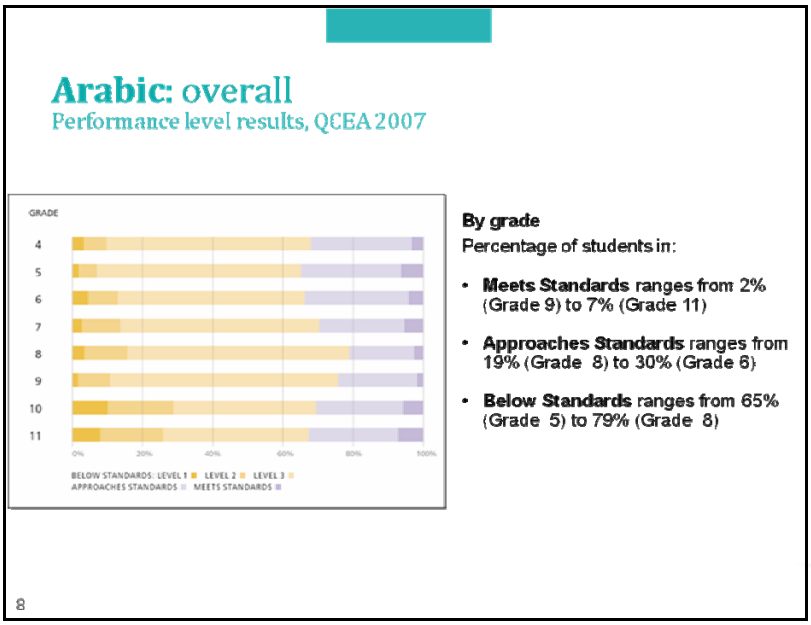
La forma cuantitativa más ortodoxa de presentación de puntajes es en escala IRT para las pruebas referidas a criterios y para estas mismas pruebas la forma aceptada de presentar resultados en forma cualitativa es por medio de niveles de desempeño, lo cual implica la previa determinación de estos por medio de cualesquiera de los métodos universalmente aceptados, tales como el Angoff, el de Marcadores, el de Juicio Analítico u otros (Cizek, 2001).

Figura 3. Comparación longitudinal de resultados entre los años 2005, 2006 y 2007, expresados en escala cuantitativa, en matemática, por niveles de grado, como parte del sistema de medición educacional del Estado de Qatar (QCEA). (SEC, 2007)



La posibilidad de comparación de resultados numéricos, sea ésta vertical u horizontal, está condicionada por la equivalencia psicométrica de los instrumentos. En cambio la comparación de resultados expresados en clave cualitativa es mucho más flexible. A continuación se muestran dos ejemplos: uno correspondiente a resultados expresados de forma cuantitativa (Figura 3) y el otro en forma cualitativa (Figura 4). Ambos ejemplos corresponden al programa denominado Medición Integral en Educación, en el Estado de Qatar[13], en la Península Arábiga.

FIGURA 4. COMPARACIÓN DE RESULTADOS ENTRE GRADOS PARA EL AÑO 2007, EXPRESADOS EN ESCALA CUALITATIVA (NIVELES DE DESEMPEÑO), EN LENGUAJE (ÁRABE), COMO PARTE DEL SISTEMA DE MEDICIÓN EDUCACIONAL DEL ESTADO DE QATAR (QCEA)



Como es posible apreciar, las escalas cualitativas permiten una mayor flexibilidad de comparaciones y una mucha mayor riqueza de información, sin requerir el cumplimiento de condiciones tan estrictas como en el caso de las escalas cuantitativas, aunque lógicamente no logran una precisión comparable con el nivel posible de alcanzar al usar escalas cuantitativas. Sin embargo, incluso es posible, aunque no necesariamente aconsejable, utilizar la información acerca de porcentajes de alumnos en cada nivel para comparar el rendimiento entre distintas asignaturas, lo que constituye un anatema siquiera intentarlo a partir de resultados cuantitativos.

9. DESDE LAS CIFRAS A LOS CONCEPTOS

Como se desprende de la sección anterior, existen razones y muy poderosas, para que se justifique un mayor desarrollo y énfasis en el uso de resultados en escala cualitativa, a la hora de diseminar y analizar los resultados de las mediciones estandarizadas en educación.

[13] Qatar Comprehensive Educational Assessment (QCEA)program

Sin embargo, es necesario señalar claramente que ello no implica licencia para concesión alguna en el rigor de la generación de resultados o para la prescindencia de las escalas cuantitativas, sino que significa tomar ventaja de la mayor flexibilidad de posibilidades de comparación planteada por los resultados expresados en forma de niveles de desempeño, generados por medio de la determinación de puntajes de corte, a partir de resultados cuantitativos, idealmente en escala IRT.

Los procesos de determinación de niveles de desempeño están, hoy en día, definidos y estandarizados en tal medida, que es preciso seguir pasos muy claros para el establecimiento de los niveles y posteriormente de sus límites, a través de la determinación de puntajes de corte. Aunque estos procesos descansan en procedimientos de juicio y no necesariamente en unos de raíz estadística, sí implican un rigor tanto o mayor que el propio de estos últimos.

La posibilidad de contar con información cualitativa acerca de los resultados de las mediciones estandarizadas en educación permite alcanzar tres fines. El primero, como se señaló antes, consiste en la ampliación del ámbito de los tipos de comparaciones posibles de realizar a partir de los resultados de las mediciones estandarizadas. El segundo es la asignación, a tales comparaciones, de un grado mayor de inteligibilidad para los destinatarios de la información, tales como directores, maestros, padres y alumnos, despejando la habitual duda acerca de cuál es el sentido que, en términos educativos, tiene un puntaje numérico determinado. Finalmente, establece un puente entre los resultados de las mediciones y la docencia de aula, en el sentido de que al establecer cuáles son las competencias concretas que los alumnos logran y cuáles las que no alcanzan, se genera una herramienta efectiva de retro-información y por ende de mejoramiento del aprendizaje.

La integración, actualmente posible, de resultados cuantitativos, generados con todo el rigor psicométrico propio de ellos, con esquemas de interpretación cualitativos generados también a partir de procesos igualmente rigurosos, plantea una línea de explotación de los resultados de las mediciones estandarizadas, ausente hasta ahora en el ámbito educacional, pero que promete nuevas y grandes posibilidades en beneficio de los estudiantes.

10. Predicción, alineamiento y "ecualización" de puntajes

Como se adelantara dos secciones más atrás en este artículo, la respuesta a la necesidad de comparar resultados generados por las mediciones estandarizadas en educación, ha constituido y constituye actualmente, una necesidad que hace posible el cumplimiento de un propósito fundamental de la medición, cual es la evaluación.

La principal razón de esta necesidad surge del hecho que, las más de las veces, lo que se intenta contrastar son distintos elementos (sujetos) en un mismo o distinto momento (comparaciones transversales) o los mismos elementos (sujetos) en diferentes momentos (comparaciones longitudinales). En ambos casos surge como condicionante la necesidad de que los instrumentos usados sean comparables y para ello se utilizan procedimientos psicométricos denominados de "vinculación[14]".

[14] En inglés "linkage".

No son pocos los casos en que erróneamente se han generado conclusiones alentadoras o negativas respecto de la situación de sistemas escolares, producto de que se ha intentado comparar puntajes que no son comparables, dado que estos provenían de instrumentos que no eran equivalentes.

La denominada "vinculación" de pruebas hace su aparición a comienzos del siglo veinte, con los trabajos de varios autores, entre ellos, Starch, Weiss y Kelley, citados por Holland (2005).

La vinculación de puntajes de pruebas, de acuerdo al mismo autor recién citado, comprende tres grandes sub-categorías, a saber:

- **Predicción.** Lo que se obtiene es la "mejor predicción"[15] de un puntaje Y a partir de un puntaje X, utilizando procedimientos de regresión.
- **Alineación de escalas**[16]. Es un proceso más complejo que la predicción y consiste en la ubicación conjunta de los puntajes X e Y en una escala común, en vez de asignar a X un puntaje en la escala de Y. Los métodos más comunes son la alineación de batería, la alineación respecto de un ancla[17], la alineación respecto de una población hipotética, la calibración, la alineación vertical y la concordancia.
- **"Ecualización"**[18]. Es el proceso más complejo de todos y se lleva a cabo cuando los diferentes sujetos rinden diferentes formas o versiones de la misma prueba y sus resultados en ellas requieren ser comparados. Este proceso es posible abordarlo también con Teoría Clásica, IRT, ecualización por post-estratificación, ecualización en cadena, método de puntajes observados de Levine y método IRT de puntajes observados.

Para finalizar esta breve revisión acerca del origen y de la práctica de los métodos que permiten la comparación válida de puntajes de instrumentos de medición del rendimiento, es preciso mencionar que una de las modalidades comprendidas dentro de los métodos de alineamiento, debe destacarse un método que recientemente ha ganado en popularidad, cual es la alineación vertical.

Su mayor mérito radica en que hace posible la estimación del crecimiento o progreso del aprendizaje de los alumnos en el tiempo, entregando información, por ejemplo, de cuál ha sido la trayectoria de un alumno en matemática desde tercero a cuarto de educación básica. Un método alternativo y menos exigente en términos psicométricos es el cálculo de valor agregado, aunque sí lleva aparejadas otras exigencias, esta vez operacionales. En ambos casos la limitación mayor es la "distancia" en el tiempo que pueda existir entre ambas mediciones, la que se expresa en el número de niveles de grado en el caso de la alineación vertical y en el tiempo transcurrido entre la mediciones en el del valor agregado.

11. Condiciones para que una evaluación constituya una herramienta de política en educación

Uno de los aspectos más controversiales de la medición en educación es su utilización para la toma de decisiones en materias de política educacional, dado que el volumen de datos y de información que se acumula en los anaqueles de los ministerios y otras entidades de los sistemas educacionales es enorme,

[15] En inglés "best prediction".
[16] En inglés "scale alignment" o "scaling".
[17] En inglés "anchor test".
[18] En inglés "equating".

como asimismo lo son los bancos de datos que rebosan de resultados de operaciones de medición. Es también muy grande la carga que se impone sobre profesores y alumnos a raíz de someter a estos últimos a evaluaciones, generadas al nivel escuela, municipio, provincia o estado y/ país y no devolverles mayores beneficios.

Sin embargo, el número de ocasiones en que las decisiones en materia de política educacional son basadas en la información dura, generada por la medición estandarizada, es inversamente proporcional al crecimiento del volumen de información generada. Si bien autores como Gladwell (2005) abogan por el valor de las decisiones producto de la inspiración súbita, cabe la duda de si acaso aquellas en educación puedan y sea aconsejable que sean adoptadas de tal modo.

Lamentablemente las razones por las que la información generada a partir de los procesos de medición estandarizada no es utilizada en favor de decisiones mejor fundadas en educación es, en gran medida, responsabilidad de quienes estamos involucrados en tales procesos. Un análisis, a la vez sistemático y descarnado, de la situación imperante en este sentido, nos permitirá aislar las causas principales y por ese medio intentar subsanarlas.

La primera condición, que no siempre se cumple, es la que constituye la raíz del tema de este trabajo, la medida en que la información generada por la medición estandarizada es válida. En otras palabras, mientras las pruebas sólo se enfoquen en las competencias más simples de la escala de complejidad de los aprendizajes, su relevancia será discutible y por ende el estímulo implícito para la utilización de sus resultados será también limitado. Es imprescindible que los sistemas de medición desplieguen esfuerzos por ampliar el ámbito de validez de sus instrumentos, aún en tiempos en que los presupuestos no son del todo abundantes.

Una segunda condición se relaciona con la necesidad de que la información proveniente de las mediciones estandarizadas esté disponible para los diferentes actores de manera oportuna y sea inteligible para ellos. Son muchos los casos en que los informes de las mediciones toman años en ver la luz, cuando la ven, y lo que es más, suelen estar escritos en código de especialistas, siendo por ello incomprensibles para docentes, padres y alumnos y consecuentemente no concitan su interés.

Las más de las veces se invierte un enorme esfuerzo en la elaboración y análisis de los instrumentos de una medición, lo cual es sin duda lo pertinente de hacer, sin embargo no se asigna a la elaboración y producción de los informes el mismo esfuerzo e interés. Las Figuras 5 y 6 muestran un ejemplo de un informe para profesores que, desde hace dos años, está siendo provisto a los docentes en el Estado de Qatar, basándose en los resultados del sistema de medición del rendimiento, en ese país. Tal informe, junto con el correspondiente para padres y alumnos, el informe para los tomadores de decisiones a nivel superior y el informe técnico para especialistas, se procura que esté en sus manos no después de seis meses de aplicadas las pruebas, los cual no si bien no es un plazo breve, sí implica una voluntad de hacer llegar la información en la forma lo más oportuna y útil posible.

FIGURA 5. PORTADA Y CONTRAPORTADA DEL INFORME PARA DOCENTES DE LENGUAJE (ÁRABE), UTILIZADO POR EL SISTEMA DE MEDICIÓN DEL RENDIMIENTO DEL ESTADO DE QATAR, PARA EL PERÍODO 2005-2006. (SEC QATAR, 2007)

Otra condición limitante en la utilización de los resultados de las mediciones estandarizadas es la falta de desarrollo de la capacidad y voluntad para utilizar la información de quienes toman las decisiones en educación. De hecho, quienes deciden en este ámbito, desde padres hasta ministros, no están, las más de las veces, capacitados para utilizar información dura para sustentar sus juicios. Contra esta capacidad conspira también lo que se manifestara antes, en esta misma sección, en cuanto a que no siempre el lenguaje de los informes resulta adecuado para los actores en cada nivel que lo recibe.

Pero habiendo subsanado lo anterior deben hacerse esfuerzos para generar acciones concretas en pro de que estos actores, por una parte reciban capacitación para utilizar los informes basados en la medición estandarizada como base para sus decisiones, como asimismo para motivarlos a que lo hagan.

Lo más habitual hoy en día es que los padres tomen decisiones respecto de la educación de sus hijos basados en rumores y comentarios de sus conocidos. En el caso de quienes deciden en los niveles altos de los sistemas el caso es que las presiones de grupos relacionados con la educación tienen más efecto en lo que se resuelve que la información proveniente de fuentes objetivas y empíricas. Mientras esto siga ocurriendo, como lo señalara Bloom, en la cita al comienzo de este trabajo, seguiremos "…siendo seducidos por el equivalente de los remedios de curanderos, las falsas curas del cáncer, la invención del movimiento perpetuo y las supersticiones…".

FIGURA 6. PÁGINAS INTERIORES DEL INFORME PARA DOCENTES DE LENGUAJE (ÁRABE,) UTILIZADO POR EL SISTEMA DE MEDICIÓN DEL RENDIMIENTO DEL ESTADO DE QATAR, PARA EL PERÍODO 2005-2006. (SEC QATAR, 2007)

12. MITOS Y ELABORACIONES RETÓRICAS: ¿LA "APARICIÓN" DE LA EVALUACIÓN FORMATIVA? ¿LA MEDICIÓN DEL APRENDIZAJE VERSUS LA MEDICIÓN PARA EL APRENDIZAJE?

Uno de los aspectos que se mencionó antes como problemático respecto de la medición estandarizada de los resultados de la educación, en la actualidad, es la mítica aparición de "nuevas" tendencias y enfoques en este campo. Si bien los ejemplos son numerosos, un caso emblemático es el de la evaluación formativa. Es sorprendente como para muchos "especialistas" en evaluación y actores en educación, hoy en día, éste es un término que llegó muy recientemente a este campo por primera vez. O bien no estuvieron presentes cuando estos términos vieron la luz por primera vez, o estaban distraídos en ese entonces o simplemente no han hecho sus deberes para la casa.

Lo concreto es que, como señala una autora (Dwyer, 2008) "En 1967, Michael Scriven propuso el uso de los términos "formativo" y "sumativo" (comillas en el original) para distinguir entre los diferentes roles que puede jugar la evaluación. Por una parte, él señaló que la evaluación "puede tener un rol en el mejoramiento continuo del currículo" (p.41), mientras que por la otra, la evaluación "puede servir para habilitar a los administradores para decidir si el currículo completo y terminado, refinado por el uso del proceso evaluativo en el primero de sus roles, representa un avance suficientemente significativo entre las alternativas disponibles, para justificar su adopción por un sistema escolar (determinado) (pp. 41-42). Luego propuso "utilizar los términos evaluación "formativa" y ¨sumativa" para calificar a los roles de de la evaluación".

"Dos años más tarde, Benjamin Bloom (1969) aplicó la misma distinción a las pruebas usadas en el aula."

Como quedó demostrado, la evaluación formativa fue definida por Scriven, desde una perspectiva sistémica, como la herramienta para el mejoramiento continuo del currículo y, por Bloom, desde un punto de vista pedagógico del aula, como un elemento vital para mejorar el aprendizaje de los alumnos, en ambos casos hace más de cuatro décadas.

Hoy se intenta utilizar este concepto para desacreditar a la medición estandarizada, como asimismo a la evaluación que se basa en ella, ya que se las considera como opuestas a lo "formativo". Baste decir que, por un lado, la idea de lo formativo califica las acciones evaluativas, por lo cual ellas no son más o menos formativas en su esencia, sino que en su uso. Sería como decir que un bisturí es esencialmente malo por cuanto puede usarse para matar, además de ser una herramienta curativa. Por otro lado, uno de los artífices más distinguidos de la introducción del rigor en la medición en al aula fue, precisamente, Benjamin Bloom. Mal pudo él propugnar que la medición estandarizada pudiese percibirse como antagónica del rol formativo de la evaluación.

Entrando, ahora, al ámbito de la manipulación retórica de los términos en este campo, segundo fenómeno de reciente aparición, el caso más destacado es el de la distinción artificial y artificiosa entre evaluación **del** aprendizaje y evaluación **para** el aprendizaje, las que se hacen aparecer como antagónicas.

Un simple análisis lógico baste para desvirtuar este intento sofista. En primer lugar, cualquier procedimiento y entre ellos, uno evaluativo, antes que nada debe poseer un objeto sobre el que se lleva a cabo, el que en este caso es el aprendizaje. Por lo tanto, no es posible que exista una acción evaluativa "in abstracto", sino que esta debe evaluar algo en concreto.

En segundo lugar, esa misma acción debe tener un propósito, en este caso la mejora del aprendizaje posterior y éste propósito no tiene razón para ser considerado opuesto al objeto antes definido. Por lo señalado, es perfectamente aceptable que la evaluación del aprendizaje en un cierto momento, vaya en pro del aprendizaje posterior.

A lo señalado, se agrega que una revisión mínima de la historia de la medición estandarizada y de la evaluación ligada a ella, permitirá comprobar que, desde al menos los años sesenta del Siglo XX, en que surge en la educación la idea de retro-información, la evaluación ha sido definida como una herramienta de mejoramiento del aprendizaje. Cualquier otra elaboración retórica no es más que eso, ya que revela desconocimiento del campo de la disciplina o bien una falacia en el razonamiento, que va a parejas con lo negativo de la intención.

Al finalizar este trabajo, sólo resta señalar que se ha procurado mostrar aquí que, si bien la medición estandarizada y su consecuente, la evaluación, distan mucho de ser exhaustivas e irreprochables en su validez y precisión, no existe, hoy por hoy, un enfoque científicamente sustentado que permita plantear una alternativa válida.

La conclusión última es entonces que lo sabio y prudente de hacer es trabajar para mejorar estas dos herramientas de la educación y ser, a la vez, extremadamente cuidadosos en que al aplicarlas se evite, a toda costa, asignarles mayores cualidades que las que realista y científicamente demuestren poseer.

REFERENCIAS BIBLIOGRÁFICAS

American Educational Research Association, American Psychological Association y National Council on Measurement in Education. (1985*). Standards for educational and psychological testing.* Washington: American Educational Research Association.

American Psychological Association (1966). *Standards for educational and psychological tests and manuals.* Washington: APA.

American Psychological Association (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin [Supplement], 51 (2, part 2).*

Bloom, B.S. (1972). Innocence in Education. *School Review*, 80, pp.333-352.

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. En R.W. Tyler (Ed.), *Educational Evaluation: new roles, new means. The 63rd yearbook of the National Society for the Study of Education, 69,* pp. 26-50.

Cizek, J.G. (2001). *Setting performance standards. Concepts, methods, and perspectives.* Mahwah Erlbaum.

Comber, L.C. y Keeves J.P. (1973). *Science education in nineteen countries*. Stockholm: Almqvist & Wiksell.

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

Dwyer, C.A. (2008). *The future of assessment. Shaping teaching and learning*. New York: Erlbaum.

Gladwell, M. (2005). *Blink. The power of thinking without thinking*. London: Penguin.

Haladyna, T.H. (2002). Supporting documentation: assuring more valid test interpretations and uses. En Tindal y T.M. Haladyna (Eds.), *Large scale assessment programs for all students*. Mahwah: Erlbaum.

Holland, P.W. (2005). Linking scores and scales: some history and organizing ideas. En *Linking and aligning scores and scales. A conference in honor of Ledyard R. Tuckers´s approach to theory and practice.* Princeton: Educational Testing Service.

Kaplan, A. (1963). *The conduct of inquiry: Methodology for behavioral science.* New York: Harper & Row.

Lissitz, R.W y Samuelsen K. (2207). A suggested change in terminology and emphasis regarding validity an education. *Educational Researcher, 36, pp.* 437-448.

Messick, S. (1989). Validity. En R.L. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

Moss, P.A. Reconstructing validity. (2207). Educational *Researcher, (36,* pp. 470-476)

Nichols, P.D. y Williams, N. (2009). Consequences of test score use as validity evidence: roles and responsibilities. *Educational Measurement: Issues and Practice, (Spring,* pp. 3-9).

Rodriguez, M.C. (2002). Choosing an item format. En G. Tindal y T. M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 213-231). Mahwah: Erlbaum.

Scriven, M. (1967). *The methodology of evaluation* (Vol. 1). Washington: AERA.

Supreme Education Council. (2007). Qatar Comprehensive Educational Assessment (QCEA). *QCEA 2007 Results.* Disponible en: http://www.english.education.gov.qa/section/sec/evaluation_institute/sao/_qcea