



Programa de Promoción de la Reforma Educativa en América Latina y el Caribe  
Partnership for Educational Revitalization in the Americas

## **Grupo de Trabajo sobre Estándares y Evaluación**

**Cuestiones técnicas que  
condicionan las interpretaciones  
de los datos generados por las  
evaluaciones de logros de aprendizaje  
escolar en América Latina**

**Richard Wolfe**  
Abril 2007

## **Cuestiones técnicas que condicionan las interpretaciones de los datos generados por las evaluaciones de logros de aprendizaje escolar en América Latina**

**Richard Wolfe**

**Ontario Institute for Studies in Education of the University of Toronto, Canada**

### ***Resumen Ejecutivo***

*El Grupo de Trabajo sobre Estándares y Evaluación del PREAL está convencido tanto de la importancia de la evaluación de los aprendizajes como de que la sostenibilidad de los sistemas de evaluación depende de que se aprovechen y maximicen los beneficios que ellos prometen a la política educativa. Desde el inicio de sus actividades ha recomendado -entre otros cursos de acción- mejorar la calidad técnica de diversos aspectos de esos sistemas, en especial el diseño de los instrumentos de recolección de información y los modos de procesar y reportar los resultados.*

*En esta oportunidad, se desea especificar de manera más concreta algunos campos en los cuales es particularmente urgente introducir esas mejoras técnicas, a la luz de la creciente necesidad y compromiso de los sistemas de contribuir al mejoramiento de la gestión educativa. Se trata de identificar y describir los principales problemas que afectan la validez de las interpretaciones más comunes de los resultados obtenidos en las evaluaciones latinoamericanas.*

*Las cuestiones aquí seleccionadas para la discusión se derivan de la experiencia de quince o más años que tiene el autor respecto a diversos programas de evaluación nacional y regional en América Latina. Si bien es cierto que a lo largo de esos años, los objetivos de las evaluaciones han ido reformulándose y sus métodos sofisticándose, continúa siendo necesario clarificar sus fines y mejorar sus métodos.*

*Entre los temas aquí tratados se encuentran: (a) el uso de procedimientos para mejorar la validez en el diseño de los ítems, las pruebas y los sistemas de pruebas; (b) la necesidad de asegurar la comparabilidad de resultados a lo largo del tiempo; (c) métodos para orientar el análisis de los efectos del contexto escolar; y (d) sugerencias sobre el análisis y reporte de las escalas de rendimiento que permitan mejorar la interpretación y utilidad de los resultados. En cada caso, se detalla la importancia de esos problemas para la interpretación y uso correctos de los resultados. La parte final del documento incluye algunas recomendaciones sobre cursos de acción a tomar para superar esos problemas.*

## 1. Introducción

El presente texto hace referencia al diseño, implementación, interpretación y uso de evaluaciones a gran escala del logro educativo en América Latina y si bien se encuentra focalizado en las pruebas nacionales, es pertinente tanto para las evaluaciones nacionales como para proyectos regionales o internacionales.

En la mayoría de los países de la región existe actualmente algún tipo de recolección periódica de datos que utiliza pruebas educacionales estandarizadas para proveer análisis y reportes sobre los niveles de aprendizaje escolar. Históricamente, este trabajo se puede considerar como una extensión del ejercicio más tradicional de reportar estadísticas educativas bajo la forma de extensas memorias anuales que producen los Ministerios de Educación brindando información detallada sobre estudiantes, profesores y escuelas, agregados a nivel de distritos, provincias, etc. En dichos documentos, las variables que se analizan son muy simples: grado, sexo, edad. Las medidas de resultados generalmente se restringen a aspectos formales de progreso y certificación (aprobados, desaprobados o repitentes). Las tablas y -a veces- algunos gráficos ilustran los tamaños relativos de distintas unidades, tasas de resultados y cambios a lo largo del tiempo. El único análisis complejo incluido se relaciona con la determinación y proyección de cambios y

crecimiento, especialmente en lo referido a la promoción, retención y deserción.

Las evaluaciones de logros educativos añaden información nueva e importante sobre la *calidad* de los resultados educativos, mientras que las estadísticas se refieren principalmente a las *cantidades* de educación. Ciertamente es que contabilizar la aprobación y desaprobación es un aspecto de la calidad; sin embargo, las estadísticas educativas tradicionales no brindan evidencia sobre el contenido de los logros. Evidentemente, las evaluaciones de logros resultan siendo más complicadas y costosas que la recolección y reporte de estadísticas, razón por la cual frecuentemente se realizan sólo en grados y materias seleccionados y no todos los años. En algunos países, las evaluaciones educativas son realizadas por agencias externas al Ministerio de Educación.

Asimismo, la mayoría de las evaluaciones de logros educativos contienen un componente importante de análisis curricular. Las pruebas de logros se desarrollan a partir de un análisis detallado de los contenidos y expectativas del currículo, siendo su objetivo determinar el grado al cual los alumnos alcanzan las exigencias del currículo y, supuestamente, orientar el mejoramiento de éste, examinando -por ejemplo- si la amplitud y graduación de los contenidos son adecuados.

Otro objetivo de algunas evaluaciones es identificar las relaciones entre el logro educativo y factores correlacionados con la distribución de ese logro entre estudiantes, maestros y escuelas. Algunos de esos factores, tales como la asistencia y atención de los estudiantes, el tiempo y los contenidos de la enseñanza y la calidad de la misma pueden ser considerados como determinantes principales de un logro más alto o más bajo. Otros, tales como la organización de la escuela, la experiencia y calificaciones de los docentes o el origen social y actitudes de los estudiantes, pueden ser factores antecedentes o mediadores.

En años recientes, las evaluaciones educativas en algunos países se han convertido en parte de los sistemas de administración y responsabilización educativa. Los puntajes obtenidos por individuos determinan cuestiones tales como su promoción y los puntajes agregados determinan la asignación de incentivos y premios u orientan la libre elección de la escuela.

Este último tipo de aplicación remite a una distinción crítica entre evaluaciones educativas que están basadas en muestras de estudiantes y escuelas y aquellas que se basan en una recolección de datos censal. Para muchos propósitos, excepto para la administración y responsabilización, las muestras resultan más baratas y rápidas, proporcionando esencialmente la misma información agregada, tal como estimados nacionales y estimados de cambios a

lo largo del tiempo. Llevar a cabo aplicaciones censales resulta claramente más costoso y tiene que justificarse sobre la base de la utilidad, validez y sostenibilidad de políticas administrativas que requieran datos específicos de cada unidad de un sistema escolar.

Cualesquiera sean los propósitos generales de la evaluación educativa, los asuntos referidos a su precisión, validez e interpretabilidad son cruciales. Debido a ello, en el presente documento se da una mirada a los elementos técnicos que ponen en riesgo dichas cualidades determinantes así como a métodos que permitan lidiar con ellos. Entre las razones por las cuales aquí se les da importancia están::

- Los reportes de las evaluaciones pueden contener detalles falsos que sobredimensionan o presentan sin evidencia estadística significativa diferencias de puntajes en distintos contenidos de las pruebas – tales como áreas curriculares – y entre distintas poblaciones escolares – tales como tipos o localizaciones de escuelas. Así pues, se podría llegar a introducir cambios en políticas basándose en diferencias no significativas o debidas puramente al azar, lo cual significa que esas decisiones habrían sido planteadas sin una justificación válida.

- Si la calidad técnica de las mediciones y análisis es inadecuada, existe la probabilidad específica de malinterpretar un progreso como declive o viceversa, generándose de este modo confusión y desorientación política y social.
- El modo en que se ordena y presenta la información genera dificultades que afectan las inferencias derivadas de las evaluaciones de logros. Los datos recogidos de manera transversal o en un momento del tiempo (*cross-sectional*), por ejemplo, proporcionan una base débil para analizar los efectos de programas escolares y prácticas. Resultados obtenidos de datos sincrónicos pueden estar tan sesgados como para reflejar de manera inversa el fenómeno realmente subyacente. Dada esta situación, podría incluso considerarse afortunado el hecho de que los maestros hayan preferido ignorar varios hallazgos sobre cómo aulas más grandes muestran logros más altos.
- La omisión de variables importantes en el análisis, tales

como la oportunidad de aprender, ocasiona que otras variables tales como el nivel socioeconómico absorban estadísticamente sus efectos. A esto se le llama “error de especificación”. Los resultados pueden llevar a grandes equivocaciones en la comprensión de los efectos de factores sociales y de las reformas educacionales.

Los problemas y posibles soluciones mencionados en este texto se consideran relevantes para el caso de la reciente participación de los países latinoamericanos en estudios regionales e internacionales. El hecho de que existen diferencias entre los marcos de contenidos de las pruebas regionales e internacionales y los currículos nacionales representa una complicación adicional que puede hacer que la medición y comparación de las pruebas internacionales resulten de alguna manera sesgadas. Es particularmente preocupante porque los países latinoamericanos se encuentran por lo general en la parte más baja de la escala de logros. El tratamiento psicométrico y estadístico en esa parte de la escala es particularmente inexacta, debido al número pequeño de ítems relativamente sencillos que suelen incluirse en estas pruebas, diseñadas con más altos

niveles de expectativas de aprendizaje que las que contienen los currículos nacionales de países en desarrollo. Ello puede dar lugar a que aparezcan cambios aparentemente grandes entre una medición y la siguiente, mientras que en los países de logros más altos los cambios son generalmente relativamente pequeños. Ello podría deberse a un error de equiparación (*equating*), y tratarse más bien de una variación imprecisa, sin interpretación válida real.

## 2. Marco teórico

La visión clásica sobre medición y evaluación ordena las cuestiones relacionadas con la calidad de las pruebas refiriéndolas a los conceptos de *validez*, *confiabilidad* y *comparabilidad*. También refiere a otros criterios, tales como la validez de constructo y la validez aparente, la confiabilidad interna y de los post-tests, y la calibración y equiparación.

Una conceptualización más moderna e integrada contempla a todas éstas como aspectos distintos de un concepto unificador de *validez de la interpretación y uso de los puntajes de las pruebas*. Esta concepción implica que, más allá de la tecnología utilizada para elaborar los ítems y las pruebas, para recoger las respuestas, para asignar puntajes y para resumirlos estadísticamente, lo que importa a fin de cuentas es la posibilidad de confiar en las

interpretaciones y usos que se harán de los puntajes de las pruebas. Ésta es la perspectiva tomada y articulada en los ampliamente aceptados *Estándares para Pruebas Educativas y Psicológicas* de la Asociación Americana de Investigación Educativa (AERA), la Asociación Americana de Psicología (APA) y el Consejo Nacional de Medición Educativa (NCME) de 1999.

Por ejemplo, luego de una evaluación nacional de logros, digamos en Matemáticas en el sexto grado, se producen puntajes que pretenden medir los logros de los estudiantes con respecto al currículo nacional de matemáticas. Es necesario validar las interpretaciones y usos del puntaje obtenido por un estudiante y los puntajes agregados obtenidos por aulas, escuelas y distritos de la nación. ¿Es válido interpretar los puntajes y usarlos como indicadores de si los estudiantes han alcanzado los estándares educativos establecidos por el currículo? ¿Podemos aplicar con precisión esta interpretación a alumnos individuales? ¿Son confiables las comparaciones que hacemos para evaluar el cambio del sistema a lo largo de los años?

La interpretación depende en primer lugar de la calidad de las pruebas: ¿Son suficientemente precisas y representativas del contenido del currículo? ¿Miden cantidades mínimas de contenidos irrelevantes y cantidades máximas del contenido intencional? ¿Incentivan a los estudiantes a aplicar procesos cognitivos

similares a aquéllos que sugiere el currículo?  
 ¿Es coherente la manera en que se califican los ítems con la naturaleza del constructo que se busca evaluar (v.gr., ¿se incluye o no la calidad de la caligrafía como componente de la capacidad de comunicar ideas por escrito?)?. Éstos son los primeros principios que definió Samuel Messick en “Validez de la evaluación psicológica: Validación de inferencias sobre personas?. Respuestas y desempeños como una exploración científica del significado de los puntajes” [Validity of Psychological Assessment: Validation of Inferences from Persons? Responses and Performances as Scientific Inquiry into Score Meaning], *American Psychologist* 50 (9) (Setiembre 1995).

La validez de las interpretaciones de los puntajes está limitada por su precisión. Desde la perspectiva tradicional, esto se refiere a la confiabilidad de la prueba (v.gr., la correlación entre una prueba y su reaplicación). En cambio, autores como Messick están más preocupados por la *generalizabilidad* de los puntajes y sus interpretaciones, lo que requiere constatar, por ejemplo, que en una prueba de matemáticas haya una muestra suficiente de ítems, tipos de ítems y contenidos, y establecer si las interpretaciones son igualmente válidas para distintos tipos y grupos de estudiantes. En el caso de evaluaciones continuas a gran escala, es necesario otorgar especial consideración a la comparabilidad de las mediciones entre distintas pruebas y distintos períodos de tiempo. Si las

pruebas cambian o se traslapan entre una y otra aplicación, es necesario aplicar técnicas complejas de calibración y equiparación. Si bien la calibración no puede evitar la introducción de errores, éstos tienen que ser minimizados y estimados..

También es importante tener en cuenta que la evaluación educativa a gran escala se da dentro de la realidad de un sistema escolar complejo, jerárquico y diferenciado. Los informes de resultados pueden llegar a estudiantes individuales, a docentes y a escuelas y a otras unidades administrativas hasta el nivel central nacional o incluso internacional. Es una responsabilidad perentoria el asegurar que a cada nivel de reporte haya validez y precisión suficientes como para justificar las interpretaciones y usos que se dará a los datos.

En un sentido aun más amplio, Messick y otros argumentan que es necesario considerar la validez de las *consecuencias* de los usos de los puntajes de logros. Pudiera suceder que los resultados de las pruebas fueran utilizados para tomar decisiones sobre individuos (v.gr., promoverlos), sobre maestros y escuelas, o sobre el progreso de las reformas curriculares y educativas en general. En este caso, es necesario preguntarse si es posible garantizar que tales mediciones tienen las cualidades adecuadas como para constituirse en evidencia suficiente para ese tipo de acciones.

Otra responsabilidad aun mayor consiste en proporcionar información contextual adecuada que permita arribar a interpretaciones razonables de los resultados, especialmente aquéllas que comparan individuos o grupos. En primer lugar se debe identificar si se está evaluando y reportando contenidos que distintos estudiantes o grupos de éstos han tenido oportunidades adecuadas de aprender, pues en la medida en que esas oportunidades no hayan estado disponibles, las diferencias se podrían atribuir a condiciones pre-existentes y no a los esfuerzos de los estudiantes o la efectividad de la enseñanza. Del mismo modo, se debe analizar si las condiciones de aprendizaje, tanto en casa como en la escuela, son las mismas para todos los estudiantes. Resulta evidente que esto no es así, por lo tanto, resulta importante, al momento de analizar e interpretar resultados de pruebas, obtener medidas razonables de dichas condiciones (v.gr., origen socioeconómico de los estudiantes, recursos del aula o de la escuela).

Aunque difícil de evitar, la atribución de causa resulta frecuentemente equivocada, Por ejemplo, cuando los resultados de una evaluación muestran diferencias entre tipos de escuelas (tales como las públicas y privadas) que convocan a distintos tipos de estudiantes (más pobres o más ricos), es probable que los evaluadores o sus audiencias infieran que esos tipos de escuelas tienen efectividad distinta (que los colegios privados ofrecen mejor enseñanza) y que los estudiantes tienen distintas

capacidades (los más ricos saben más o aprenden mejor). Es necesario tener presente que la trayectoria de causalidad es más bien compleja y que las correlaciones mal controladas pueden llevar a conclusiones erróneas.

Un aspecto particular de las pruebas a gran escala es la diferenciación entre *aprender* y *saber*. Lo que los estudiantes saben se mide usualmente a través de aplicaciones en algún momento (generalmente hacia el final) de un año escolar. Resulta obvio que aquello que los estudiantes saben se encuentra altamente determinado por --y correlacionará con-- toda su trayectoria previa de escolaridad y vida familiar. También correlacionará con los esfuerzos y recursos educativos desplegados durante ese mismo año, pero resulta imposible, en principio, separar esos efectos de los factores anteriores. Esto significa que atribuir el conocimiento de los estudiantes a la escuela y los docentes a los que han estado asignados en ese año en particular resulta falaz, como lo discuten Wiley y Wolfe en “Problemas que plantea la concepción del Tercer Estudio Internacional de la IEA sobre Matemáticas y Ciencias,” UNESCO Perspectivas, XXII, 3. 1992 (83). Con estudios transversales, en un momento del tiempo, se puede establecer correlaciones entre logros estudiantiles y prácticas pedagógicas, mas no se puede determinar si fueron tales prácticas las que determinaron dichos logros. Una hipótesis alternativa, que no puede ser puesta a prueba con datos transversales, es que las prácticas se

aplican diferencialmente de acuerdo a niveles previos de logros, los cuales están correlacionados con los actuales logros. Por ejemplo, podría suceder que se esté proporcionando mejores ambientes de aprendizaje a alumnos que ya tienen logros más altos. Si fuere así, no se podría inferir válidamente que son dichos ambientes de aprendizaje los que están *causando* los mayores logros finales.

El uso de datos longitudinales permite realizar análisis más significativos del *aprendizaje* durante un año escolar y de la relación entre éste y los recursos y prácticas de enseñanza. El diseño mínimo incluye un *pre-test* al inicio del año escolar y un *post-test* al final del mismo. El análisis puede ser mucho más fino si se introducen más períodos de tiempo y una serie de tiempo más larga. Desafortunadamente, este tipo de aproximación difícilmente se realiza en América Latina o en cualquier otra parte del mundo y, por lo tanto, es más común encontrarse con sobre-interpretaciones de correlaciones posiblemente espúreas. Ante lo difícil que puede ser contar con un diseño ideal, en el texto de Wiley y Wolfe antes mencionado se discute un diseño de transacción, que consiste en combinar mediciones a grados sucesivos y analizarlos como si fueran cohortes sintéticas.

### 3. Problemas y soluciones

En este texto se da tratamiento a cuatro aspectos principales del diseño, implementación e

interpretación y uso de evaluaciones educacionales a gran escala, tal como éstos se encuentran o están evolucionando en América Latina. Dichos aspectos son:

- Validez de ítems y pruebas
- Comparabilidad
- Análisis e interpretación del contexto
- Análisis e interpretación de logros

Para cada aspecto, se identificará en primer lugar los problemas evidentes en el área y luego se enumerará algunas soluciones que podrían considerarse.

### 4. Validez de ítems y pruebas

*Problema:* Entre los diferentes proyectos y sistemas de evaluación en América Latina, se encuentran pocos casos en los cuales las pruebas reflejan con precisión los contenidos curriculares.

En primer lugar, se observa un uso excesivo de ítems de opción múltiple. Éstos sirven para ordenar individuos en una escala normativa, pero no contienen información suficiente para medir el cumplimiento de criterios de logro en tanto sólo pueden medir una parte limitada de los conocimientos y habilidades de los estudiantes. Por ejemplo, no podemos realmente medir la calidad de la escritura de los estudiantes sin que éstos escriban algo. Tampoco puede evaluarse la habilidad del estudiante para

*comunicar* procesos y conceptos matemáticas con la mera selección de una opción en un ítem de selección múltiple.

Además de esto, la información que se obtiene a través de un ítem al nivel del estudiante (correcto / incorrecto) o al nivel del grupo (porcentaje de aciertos) es esencialmente ambigua en cuanto a una interpretación relacionada con criterios y estándares. No hay evidencia específica del proceso realizado por el estudiante para responder. La tasa de respuesta correcta está relacionada no sólo con el conocimiento de la materia sino también con procesos irrelevantes quizás usados para responder (adivinanza, estrategia de eliminar opciones, etc.). El hecho de seleccionar una opción no brinda evidencia muy fuerte de que se podría haber llegado a una respuesta por rutas distintas a las del dominio de la capacidad medida. La selección de una opción no demuestra por sí misma ni el razonamiento ni el proceso seguido para resolver el problema. Es por estas razones que se necesita respuestas abiertas.

Una posible alternativa al uso de un ítem de respuesta abierta es definir y utilizar conjuntos de ítems de selección múltiple que correspondan a los diferentes pasos y componentes del contenido. Por ejemplo, puede reemplazarse un ítem de respuesta abierta a un problema en matemáticas con una serie de ítems de selección múltiple que corresponden a cada conocimiento previo y etapa necesaria para resolver el

problema. De las diferentes respuestas cerradas puede inferirse algo acerca de la *calidad* de conocimiento. En la práctica, hay que utilizar entre 5 y 10 ítems de selección múltiple para obtener información equivalente a 1 ítem abierto. Esto implica que el costo de reemplazar ítems abiertos con ítems cerrados es tener pruebas muy extensas. Además de esto, el análisis necesario para construir este tipo de serie de ítems y para interpretar la combinación de respuestas es bastante complejo.

En segundo lugar, las pruebas suelen ser demasiado cortas. En términos estadísticos, una prueba consiste en una muestra del universo de ítems posibles correspondiente al contenido. La precisión de una medición varía de acuerdo con:

- la varianza de las dificultades de los ítems en el universo de ítems,
- la varianza en la interacción entre ítems y estudiantes, es decir, el grado de consistencia que se encuentra entre el rendimiento en la prueba global y los aciertos en ítems individuales, y
- el tamaño de la muestra, o sea el número de ítems;

Esto resulta problemático, porque sabemos que aun dentro de un contenido muy específico, habrá bastante variación de dificultad de ítems (algunos son fáciles, otros son difíciles) e interacciones entre ítems y estudiantes (es decir, algunos estudiantes pueden responder a ciertos

ítems y otros pueden responder a otros ítems). Una selección pequeña de ítems implica una baja precisión en la determinación de rendimiento promedio en la población de estudiantes, tanto como una baja confiabilidad en el puntaje individual. Cuando el número de ítems es pequeño, el error es grande. Así, las limitaciones son graves no sólo en cuanto al puntaje individual, sino en cuanto a la determinación de la distribución de conocimientos del grupo o entre diversos grupos.

Aunque las diferentes evaluaciones educacionales en América Latina hacen algún uso de ítems de respuesta abierta, el número de este tipo de ítems suele ser muy reducido, posiblemente uno por estudiante colocado al final de una prueba compuesta de ítems de selección múltiple. La experiencia con respuestas construidas por el estudiante nos indica que las varianzas entre ítems y entre estudiantes, en las respuestas y las calificaciones, son altísimas. Por lo tanto, el número de ítems tiene que ser grande para que haya una mayor precisión de la prueba. Esto quiere decir que uno o dos ítems de respuesta abierta no contribuirían mucho a la validez de la prueba, aunque ayudarían a la interpretación de los criterios de logro, especialmente por proporcionar ejemplos de lo que puede los estudiantes en sus propias palabras.

En tercer lugar, existen aspectos importantes de los dominios de contenido, tales como

- Habilidades para hablar y escuchar
- Capacidad de realizar tareas grandes y de largo plazo
- Saber trabajar en grupos
- Poder realizar experimentos y aprender de la experiencia práctica, que parecen estar quedando excluidas de las actuales mediciones.

*Soluciones.* Mejorar la validez de las evaluaciones es posible, pero tiene costos:

- Aumentar las capacidades institucionales para preparar y mejorar ítems. Dedicar más tiempo al pilotaje y mejoramiento de ítems. Preparar y aplicar más ítems.
- Obtener una mayor cantidad de respuestas por estudiante, utilizando pruebas más largas o aplicaciones múltiples, aunque ello requiera usar más tiempo del alumno y de clases. Por supuesto hay problemas de cansancio y puede haber objeciones en términos de que se estaría “robando” tiempo de aprendizaje. Desde otro punto de vista, sin embargo, gran parte del costo de la evaluación lo genera el simple llegar a las escuelas y no sería una inversión eficiente el obtener sólo poca información en esa visita. El costo del tiempo invertido en las evaluaciones puede

equilibrarse con los beneficios educacionales que puede generar la evaluación—materiales, reportes, etc.

- Establecer como meta el uso de una proporción mayor de respuestas abiertas. En TIMSS, la mitad del total de ítems con que se construyó las pruebas fue de respuesta abierta (aunque la proporción de ítems abiertos que respondía cada estudiante fue menor).
- Utilizar administraciones matriciales (formas rotadas) de ítems para aumentar el número de ítems analizado e incorporarlo en los puntajes agregados a nivel de aula, escuela o nación. Puede usarse diferentes niveles de rotación. Por ejemplo, puede dividirse 200 ítems cortos en 5 formatos de 40 ítems cada uno, aplicar uno con rotación dentro del aula a cada estudiante y a la vez tener dos formatos especiales, cada uno con dos tareas largas, para aplicar éstos a una submuestra muy pequeña, quizás de 100 estudiantes en la población total. La validez y precisión de una medición al nivel del grupo son mucho mayores utilizando un diseño de 200 ítems, cada uno con respuestas de N/5 o

una quinta parte de la muestra total de estudiantes evaluados, que con un diseño de solo 200/5 ítems, cada uno con respuestas de todos los estudiantes evaluados. El número de estudiantes es igual y por lo tanto el error de muestreo de los estudiantes sería igual. Pero el error de muestreo de los ítems se dividiría entre 5. Las mediciones individuales serían un poco más variables, pero esto es menos importante cuando lo que se quiere es investigar la distribución de los rendimientos y su relación con factores asociados.

- El diseño de formas rotadas es un ejercicio estadístico y psicométrico. En teoría, cada forma puede constituir una muestra independiente. Puede obtenerse precisión más alta utilizando estratificaciones por subcontenido e ítems comunes entre las diversas formas.
- Reducir las limitaciones de contenido. Invertir lo necesario para ampliar la cobertura curricular de las mediciones tanto en detalle (utilizar ítems fáciles, medianamente difíciles y difíciles en cada punto del currículo) como en alcance (incluir aspectos de cada

parte del contenido) y en profundidad (utilizar ítems de respuesta abierta, tareas grandes, experimentos, etc.)

## 5. Comparabilidad

*Problema:* Las equiparaciones entre años muchas veces resultan siendo inválidas o muy imprecisas, razón por la cual las inferencias sobre cambios no son confiables. Se pretende medir cambios, ¡pero cambiamos los instrumentos de medición! Por un lado, el número de ítems comunes entre pruebas de diversos años es limitado y cubre solo una parte de sus contenidos. Luego, la metodología estadística y la programación para el proceso de equiparación son bastante complejas. Esto hace que muy frecuentemente se aplique incorrectamente los programas de análisis de datos basados en la TRI (teoría de respuesta al ítem). Dado que los resultados de equiparaciones se utilizan para hacer comparaciones, es decir, para estimar mejoras o deterioros en puntajes, y dado que estas estimaciones incluyen (1) diferenciales reales, (2) errores de muestreo en ambos momentos y (3) *error de equiparación*, es necesario que esta última fuente de error sea calculada y tomada en cuenta.

Por razones prácticas y educacionales, no es factible repetir exactamente las mismas pruebas usadas para una evaluación en una siguiente. Siempre es necesario publicar alguna parte de

una prueba para mostrar al público sus contenidos y respuestas, pero luego los estudiantes podrían “practicar” esos ítems a fin de prepararse para la siguiente evaluación, lo cual sesgaría considerablemente los resultados de ésta. También existe el peligro de que los docentes modifiquen sus clases para preparar a los estudiantes específicamente para la prueba previamente conocida, en lugar de enseñar con referencia a los objetivos generales del currículo. Es sabido que los programas de evaluación que intentan usar la misma prueba a lo largo de varios años siempre se encuentran con que los puntajes se elevan cada año.

Un argumento a favor de diseños matriciales como los anteriormente descritos es que contienen un número grande de ítems y la confidencialidad o seguridad de los mismos no es un problema tan crítico, especialmente si la cobertura curricular de la prueba es amplia. En este caso, el “enseñar para la prueba” podría estar representando efectivamente lo mismo que “enseñar el currículo”.

Los sistemas de evaluación deberían tener un plan claro sobre cómo proceder de año en año en lo que se refiere al muestreo y uso de ítems ya aplicados en anteriores oportunidades, a la construcción y selección de nuevos ítems y a la reserva de algunos de éstos para años posteriores. En primer lugar, el conjunto de ítems que se usarán para la equiparación tiene que mantenerse en secreto y ser incluido de un año de aplicación al siguiente, a fin de mantener

la escala de puntajes. En segundo lugar, tiene que seleccionarse un conjunto de ítems que serán “liberados” para ser publicados como ilustraciones de contenidos y respuestas, que luego tendrán que ser removidos para siempre de las pruebas. Finalmente, en cualquier año de aplicación, todo el resto de los ítems debe ser nuevo.

Este diseño básico puede tener algunas variaciones. El conjunto de ítems para la equiparación puede ser usado para establecer vínculos entre más de dos años de aplicación (AB, AC, AD,...) o puede serlo entre pares de años (AB, BC, CD,...). El tercer conjunto puede incluir tanto ítems viejos como nuevos – es decir, ítems que han sido usados con anterioridad y no fueron liberados, aunque no utilizados explícitamente para realizar la equiparación.

A fin de proveer un vínculo fuerte para las escalas de un año a otro, la muestra usada para la equiparación debe ser una buena muestra de la prueba completa, en dos sentidos: (1) necesita ser representativa de todos los aspectos del dominio de contenidos y medición y (2) necesita ser suficientemente grande. Si la muestra no es representativa de la totalidad de la prueba, la equiparación puede resultar siendo sesgada hacia los contenidos incluidos. Por ejemplo, si tanto la lectura como la escritura están incluidas en una prueba, pero la equiparación se basa solamente en lectura, la equiparación de un año a otro seguirá los cambios en lectura y

representará inadecuadamente los cambios en la escritura. Si la muestra no es suficientemente grande, la equiparación tendrá imprecisiones aleatorias y desconocidas y los cambios *estimados* de un año a otro no tendrán explicación válida alguna.

*Soluciones.* El diseño de las pruebas y sus diversas formas, la determinación de los tamaños de las muestras, la aplicación matricial de formas rotadas a muestras de estudiantes, el análisis de calibración y equiparación de pruebas y la evaluación de los errores de medición y muestreo son tareas técnicamente muy complejas que requieren atención estadística y psicometría de alto nivel.

- Dedicar más tiempo al diseño del sistema de pruebas y del muestreo, con consideración detallada de las necesidades de equiparación y la precisión que necesita tenerse.
- El diseño debe asegurar suficiente disponibilidad de los datos necesarios para la equiparación (ítems, contenidos, muestras).
- Realizar las equiparaciones contando con programas modernos, analistas bien entrenados, análisis rigurosos y asesoramiento de alto nivel.
- Obtener verificación y juicios independientes sobre los análisis.

- Determinar el error estándar de calibración e incluirlo en el análisis del error total.

## 6. Análisis e interpretación del contexto

*Problemas.* En América Latina, el diseño, análisis y reportaje de evaluaciones de logros suele realizarse prescindiendo de un marco conceptual integral referido a cómo ocurre el aprendizaje dentro del sistema escolar. Es necesario desarrollar ese marco, de manera que sea posible organizar adecuadamente qué datos se recogerá, qué variables se medirá y qué análisis se realizará.

Muchos de los “determinantes” o “factores asociados” de los logros de aprendizaje sobre los cuales estamos recogiendo información en América Latina tienen que ver con factores de clase social del alumno. Esto es sumamente importante, ya que uno de los *efectos* deseables del proceso de educación debería ser disminuir las diferencias de rendimiento entre pobres y ricos, pobladores urbanos y rurales, etc. Si bien no es factible modificar en el corto plazo la distribución de estos factores, es importante monitorear las diferenciaciones sociales que el sistema de educación contribuye a reproducir.

Otro conjunto de variables que se suele medir corresponde a las características de las escuelas

y de los docentes que estarían supuestamente correlacionadas con el rendimiento, tales como el tipo de administración, la edad y años de experiencia de los profesores, su formación y credenciales, el número de alumnos por aula, etc. Puede ser que estas variables constituyan *indicadores* de diferencias en el proceso de enseñanza, pero resultan bastante indirectos. La realidad nos indica que las correlaciones entre rendimiento y este tipo de variables son débiles y a veces son incluso inversas a lo anticipado.

En cambio, no se recoge información suficiente sobre partes importantes del ambiente escolar y de la enseñanza. Tampoco se mide de manera precisa procesos fundamentales de la enseñanza tales como tiempo, tipo y contenido de instrucción, o sea *oportunidad de aprender*. Es necesario también realizar observaciones de prácticas y conductas en el aula que tienen influencia directa en el proceso de aprendizaje y que podrían modificarse vía programas de entrenamiento, selección de maestros, textos escolares, incentivos, etc.

Sin información directa sobre los hechos y acciones que ocurren en el aula, es difícil imaginar cómo llegar a una teoría o modelo comprobable de aprendizaje escolar. Como se mencionó anteriormente, necesitamos también datos longitudinales para medir lo que aprenden los estudiantes y no sólo lo que saben. La combinación de observaciones en el aula con datos longitudinales (pre-test y post-test) es rarísima en América Latina, como lo es también

en otras partes del mundo. Sin estos dos elementos, posiblemente tendremos que abandonar la pretensión de realizar análisis de factores asociados que vayan mucho más allá que la simple presentación de correlaciones, sin mayores interpretaciones.

*Soluciones.* Adoptar una teoría integral sobre el aprendizaje y hacer un diseño de evaluación completamente nuevo.

- Determinar cuáles son las variables críticas que afectan la enseñanza y el aprendizaje, especialmente para medir oportunidades de aprender e inversión de tiempo.
- Hacer lo que sea necesario para medir dichas variables.
- Definir variables críticas sobre el contenido de la instrucción (por ejemplo, tiempo dedicado a diferentes materias, proporción de material nuevo y de revisión, presentación al grupo y al individuo, corrección de tareas) y sobre los conocimientos, habilidades y conductas de los maestros.
- Hacer lo que sea necesario para medir estas variables.
- Utilizar mediciones longitudinales donde se mida el aprendizaje

mismo, o sea el *cambio* en conocimientos.

## 7. Análisis e interpretación de logros

*Problema.* Después de una inversión enorme en implementar un sistema de evaluación de logros, los resultados presentados pueden ser percibidos como demasiado simples, por un lado, y difíciles de interpretar, por el otro.

Las definiciones de los constructos para las pruebas se realizan a nivel global (lenguaje, matemáticas, etc.) con referencia a ponderaciones que corresponden a los programas de estudio o a alguna definición de desarrollo intelectual. Por ejemplo, en matemáticas, puede haber una tabla de especificaciones que contempla X contenidos por Y niveles cognoscitivos, con N ítems por celda. Este plan asegura una representación adecuada del constructo global en el puntaje total.

Con el método de la TRI, se genera una escala para reportar el rendimiento de estudiantes individuales, de distribuciones de grupos, y promedios y cambios en el tiempo. Pero los números de la escala son inicialmente arbitrarios, sin significación o interpretación evidente. Puede tener una media de 500 y una desviación estándar de 100. Puede tener media de 0 y desviación estándar de 1. No son porcentajes de ningún conjunto de contenidos.

La escala puede adquirir alguna significación mediante el uso de comparaciones normativas. Podemos decir, por ejemplo, que un puntaje de 628 corresponde al nivel mínimo de rendimiento del 10% mejor de la población y que el puntaje 433 corresponde al nivel máximo de los 25% peores de la población -- estos serían los resultados con una escala normal con media de 500 y desviación estándar de 100. Pero una interpretación por normas no equivale a una evaluación por criterios. No podemos establecer o determinar estándares. La distribución de puntajes por percentiles no dice nada; toda distribución tiene todos sus percentiles, así que éstos resultan tan arbitrarios como la escala inicial.

Puede ofrecerse una interpretación mas sustantiva si se relaciona los puntajes de la escala con ítems ejemplificadores, algo que se explicará más adelante, pero el interés de muchos usuarios de la información suele dirigirse muy rápidamente a subconstructos tales como expresión escrita, comprensión lectora, vocabulario, decodificación de palabras; aritmética, geometría, resolución de problemas, comunicación, etc.

Aunque es penoso admitirlo, es necesario reconocer que nuestra tecnología de evaluación de logros a gran escala no permite medir muchos subconstructos con facilidad. Simplemente, no habrá números adecuados de ítems para obtener muchos subpuntajes precisos y comparables en el tiempo. No existe una regla exacta, pero

podemos decir que unos 20 a 30 ítems son necesarios para medir un subpuntaje con precisión suficiente para interpretarlo, y habrá que agregar otros 10 a 15 para mantener una equiparación precisa entre distintos años. Por lo tanto, de una prueba de 100 ítems, es posible que pueda extraerse dos o tres subpuntajes, pero no más, lo cual es muy poco, comparado con el número de categorías o áreas que suele contener un currículo.

Esta situación implica que siempre faltará información diagnóstica que permita relacionar el rendimiento diferencial en distintas áreas con factores asociados, características de los estudiantes, aulas, estratos, etc. Asimismo, es difícil relacionar rendimientos específicos con esfuerzos correspondientes de enseñanza.

*Soluciones.* Hay que tener expectativas realísticamente limitadas sobre lo que puede ofrecer esta clase de evaluaciones de logros. Específicamente, no puede haber muchos subpuntajes (por subáreas del currículo) sin aumentar enormemente el número de ítems y los costos correspondientes. Hay que concentrar esfuerzos en realizar interpretaciones válidas y sustantivas de la escala general y relacionarlas con el contenido de los ítems.

- Reportar los resultados a través de un *mapa de ítems*. Esto consiste en presentar un número razonable de ítems reales con sus grados de dificultad—porcentajes de aciertos y posiciones en la escala de TRI. Se puede ordenar la presentación de estos ítems según su grado de dificultad y según su contenido y añadir una discusión de los resultados. Puede producirse mapas con más o menos detalle para diferentes audiencias.
- Por ejemplo, se puede mostrar un ítem específico, A, que corresponde al puntaje 500 en el sentido de que para un estudiante a este nivel de puntaje general, la probabilidad de responder correctamente el ítem de acuerdo con el análisis TRI, sería de 80%. Así, estudiantes con puntajes menores tendrían una probabilidad menor a 80% de responder correctamente ese ítem y mejores estudiantes tendrían una probabilidad mayor. El ítem se identifica con el punto de la escala donde comienza a ser muy probable (80%) que los estudiantes den una respuesta correcta.
- Luego, se puede mostrar otro ítem, B, diciendo que un estudiante con un puntaje de 400 tendría 80% de probabilidad de responderlo correctamente, y un tercero, C, que sólo alumnos con puntajes mayores a 650 tendrían una probabilidad de 80% de responder correctamente. Éstos serían ítems de mediana, baja o alta dificultad y, en conjunto con otros ejemplos, nos ayudarían a entender qué pueden hacer los estudiantes que obtienen puntajes de 400, 500, 650 y puntos intermedios. Es decir, el mapa concretiza la correspondencia entre la escala de la prueba y el ordenamiento por grado de dificultad del contenido del currículo.
- El mapa facilita que los usuarios (educadores, padres de familia, público, curriculistas, etc.) comprendan la escala global y las interpretaciones en cuanto al rendimiento en diferentes componentes de la materia. Se anticipa que ítems de diferentes componentes se concentren en puntos altos o bajos de la escala, lo que demostrará cuáles contenidos o cuáles elementos de cada contenido son especialmente fáciles o difíciles.
- Para hacer un mapa adecuado, hay que tener un número suficiente de

ítems y resultados que puedan ser divulgados. Hay que asegurarse de que existan ítems ejemplificadores disponibles para su divulgación en todas las áreas y subáreas de contenidos y con diferentes niveles de dificultad. Esto requiere mucha atención en los momentos previos de planificación del diseño del conjunto de ítems y formas.

- El procedimiento general para hacer comparaciones entre grupos de estudiantes es, en primer lugar, presentar promedios, análisis de varianza, histogramas, etc., utilizando el puntaje en la escala como variable dependiente. Segundo, utilizar el mapa de ítems para interpretar las diferencias en términos de los puntos de la escala y sus correspondencias con los ítems.
- Por lo general, no es factible relacionar de manera rigurosa resultados por ítem con características de los estudiantes, aulas, estratos, etc. Debería evitarse comparaciones directas entre grupos en términos de ítems específicos o subpuntajes. Éstas serían muy difíciles de justificar desde el punto de vista de las estadísticas de validez.
- Sin embargo, hay que reconocer que la definición de subescalas y la relación de éstas con características de los procesos pedagógicos en el aula son objetivos importantes. El punto es que las pruebas que actualmente aplican los sistemas de evaluación a gran escala no permiten realizar este tipo de análisis. Siempre se observa una relación entre el número de ítems y la precisión de la medición. Puede aumentarse el número de ítems, pero esto es costoso. Una posible salida a este problema es adoptar en años subsecuentes diferentes focos para la evaluación, aumentando en cada caso el número de ítems específicos a una subárea. Así, en un periodo de varios años, se dará cobertura a las diferentes subáreas de una materia.

## 8. Conclusiones

El propósito de este texto ha sido revisar algunas cuestiones básicas teóricas y técnicas que subyacen las evaluaciones educativas a gran escala en América Latina. Se ha prestado especial atención a problemas y soluciones vinculadas a los puntos siguientes, en los cuales resalto lo que considero son mis sugerencias más importantes:

**Validez.** Puede incrementarse la validez de las evaluaciones si los diseños de las pruebas se mejoran incrementando el número, calidad y variedad de los ítems que contienen. Esto requiere la administración matricial de formas rotadas.

**Comparabilidad.** Se necesita prestar atención más rigurosa a la calibración y equiparación de las pruebas, de manera que se pueda tener mayor seguridad respecto a los indicios que dan respecto al cambio educativo. Calibrar se refiere a la construcción de escalas numéricas para servir como base constante y rigurosa para resumir y comunicar los resultados de las respuestas a los ítems y las pruebas, mientras que la equiparación se refiere al alineamiento de dichas escalas entre pruebas diferentes. Es necesario saber cuán grandes podrían ser los errores de equiparación. Estos son procesos técnicamente difíciles pero indispensables.

**Análisis del contexto.** La interpretabilidad y utilidad de la evaluación depende del análisis de variables de contexto y asociadas, pero las posibilidades de realizar inferencias válidas son mucho más altas si el diseño es longitudinal. Los diseños transversales implican serios riesgos de inferencias equivocadas.

**Análisis de logros.** No es fácil utilizar evaluaciones a gran escala para obtener información válida y confiable sobre niveles de logro en sub-dominios de contenidos. Puede mejorarse la interpretación y uso de la escala de

logros principal presentando mapas de ítems y relacionando los ítems liberados a diversos puntajes de la escala.

**Uso de las evaluaciones.** El efecto de la clase social es importantísimo, pero es ya suficientemente bien conocido. Sería muy útil que los evaluadores procedieran a utilizar teorías, modelos, datos y análisis referidos a características de la enseñanza y de la organización del aula que puedan modificarse para obtener mejores resultados.

¿Cómo determinar el impacto del reforma educacional si no se puede confiar en la validez de los resultados de las pruebas ni en su comparabilidad en el tiempo? Lo que se quiere enfatizar aquí es que, en realidad, con pruebas que hasta el momento son bastante cuestionables en su validez y comparabilidad intertemporal, muchos países latinoamericanos se encuentran en una posición inadecuada para evaluar de manera eficaz los impactos de las reformas y que, por lo tanto, deberían ejercer mayor cautela al respecto, ya que los cambios aparentemente observados podrían muchas veces representar meras fluctuaciones de error y sesgo.

El progreso real en educación será gradual y no milagroso. En los términos de la teoría informática: “el *ruido* en nuestras evaluaciones debería ser menor que la *señal* del progreso”.