



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

DOCUMENTOS

Los Próximos Pasos:

**¿Cómo Avanzar en la
Evaluación de Aprendizajes
en América Latina?**



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

Nº 20

Los Próximos Pasos:

**¿Cómo Avanzar en la
Evaluación de Aprendizajes
en América Latina?**

**Pedro Ravela (Editor), Richard Wolfe,
Gilbert Valverde y Juan Manuel Esquivel**

AGOSTO, 2001

Este trabajo se desarrolló colaborativamente en un taller realizado en Lima en agosto de 1999, por iniciativa del Grupo de Trabajo sobre Estándares y Evaluación de PREAL, coordinado por, Patricia McLauchlan de Arregui. Investigadora Principal de GRADE. E-mail: arregui@grade.org.pe

INDICE

Introducción	5
<i>Pedro Ravela</i>	
Capítulo I	10
El dilema de la “granularidad” en el diseño del sistema de evaluación: cobertura curricular vs. cobertura poblacional. <i>Richard Wolfe.</i>	
• Introducción	10
• ¿Quién es evaluado?	10
• ¿Qué es lo que se evalúa?	11
• Tipología de sistemas de evaluación	12
• El dilema de la granularidad	13
Capítulo II	14
La interpretación justificada y el uso apropiado de los resultados de las mediciones de logros. <i>Gilbert Valverde</i>	
• Introducción	14
• La validación de pruebas en América Latina. Ejemplos	15
• Opciones para la validación de mediciones en educación	16
• Consideraciones finales	18
Capítulo III	20
El diseño de las pruebas para medir logro académico: ¿referencia a normas o a criterios? <i>Juan Manuel Esquivel</i>	
• Introducción	20
• Las pruebas referidas a normas	20
• Las pruebas referidas a criterios	23
• Diferencias entre pruebas referidas a normas y pruebas referidas a criterios	25
• Nuevas perspectivas en la medición del logro: la evaluación del desempeño y la evaluación auténtica	28
• Conclusiones	28
Capítulo IV	30
La información sobre factores sociales e institucionales asociados a los resultados de las pruebas de rendimiento. <i>Pedro Ravela</i>	
• Introducción	30
• Problemas relacionados con la falta de contextualización sociocultural de la información sobre resultados de las pruebas	30
• Problemas relacionados con la información sobre las características de las escuelas y la enseñanza	31
• Problemas relacionados con la falta de investigación sistemática acerca de factores escolares asociados con el aprendizaje	32
Capítulo V	34
Alternativas Técnicas en Relación a las Escalas de Reporte de los Resultados de las Pruebas de Rendimiento. <i>Richard Wolfe</i>	
• Introducción	34
• Generalizabilidad de los resultados reportados	37
• Comparabilidad de los resultados reportados	37
• La elaboración de escalas de reporte	37
• El dilema respecto a las escalas de reporte	38
Capítulo VI	39
Conclusiones y recomendaciones. <i>Pedro Ravela</i>	39

INTRODUCCIÓN

Pedro Ravela

Durante la década de los '90, especialmente a partir de 1992, al menos 20 países de América Latina y el Caribe pusieron en funcionamiento algún tipo de sistema nacional de evaluación de aprendizajes. Este esfuerzo se apoya en algunas premisas generales que son ampliamente compartidas por académicos y responsables de la implementación de políticas educativas, tales como:

- La educación es por naturaleza propia una actividad "opaca" en cuanto a sus resultados, comparada con otras actividades humanas donde es más sencillo para la sociedad "ver" los resultados de lo que se hace. El hecho de que los niños estén o no aprendiendo lo que se espera no puede ser percibido directamente por la sociedad y por las familias.
- Los juicios que las familias pueden hacerse acerca de la calidad de la escuela a la que asisten sus hijos, normalmente se basan en aspectos como el orden existente, el trato que reciben los niños o la proposición de tareas para realizar en el hogar, pero difícilmente tienen una visión clara de los conocimientos y competencias que sus hijos están adquiriendo.
- Los resultados suelen ser "opacos" también para el propio maestro que, si bien puede tener una visión cabal acerca de lo que sus alumnos son capaces de hacer, por lo general no cuenta con una referencia externa acerca de los conocimientos y competencias que logran adquirir los niños en otras escuelas del país o de la región.
- Para las autoridades y otros tomadores de decisiones en materia de política educativa, ya no son suficientes los indicadores tradicionales sobre matrícula, cobertura, repetición y deserción. En un contexto en el que el desafío principal para la política educativa es cómo garantizar el acceso equitativo a los conocimientos y competencias fundamentales para el desempeño social, un sistema de evaluación que produzca información relevante sobre este aspecto adquiere hoy importancia estratégica para la gestión educativa.

En este sentido, **existe un consenso amplio respecto a la necesidad de contar con mecanismos que permitan producir información sobre lo que efectivamente se enseña y se aprende en las escuelas, de modo de dotar de mayor transparencia a**

los sistemas educativos y hacerlos más responsables ante la sociedad. Se asume y espera que esto contribuirá a mejorar la calidad de la educación.

Variedad de experiencias

Las experiencias impulsadas en este ámbito en la Región han tenido distintos propósitos y enfoques.

En una buena parte de los casos, la creación de sistemas nacionales de evaluación de aprendizajes ha sido impulsada por los organismos internacionales de crédito, como parte de sus convenios de préstamo con los países. Sin embargo, las características concretas que en cada país asume el sistema de evaluación parecen depender más de las capacidades técnicas y las decisiones políticas locales que de directivas específicas de dichos organismos.

En algunos países, las pruebas se realizan a nivel censal en ciertos grados, en tanto en otros se trabaja con muestras de escuelas y/o grupos determinados.

Entre quienes trabajan a nivel censal, unos han optado por publicar los resultados en la prensa, atribuyendo al sistema de evaluación la función principal de entregar información a las familias con el fin de que exista un control del usuario sobre la gestión escolar. Otros han optado por devolver la información a cada escuela con carácter confidencial, apostando al uso de la información como instrumento de aprendizaje profesional para los educadores. Algunos países han desarrollado experiencias de utilización de la información de resultados a nivel grupal, como elemento de evaluación de la labor del maestro y como parte del sistema de incentivos económicos.

Los países que trabajan sobre la base de muestras suelen presentar información de resultados generales a nivel nacional, con algunos niveles de desagregación por área geográfica y tipo de escuela. En algunos casos se realizan importantes esfuerzos por acompañar la devolución de resultados con materiales de orientación didáctica para los docentes, explicitando las áreas de menor logro y los problemas de aprendizaje y de enseñanza que pueden estar involucrados. En otros, después de varios operativos de evaluación y varios años de trabajo, nunca se han dado a conocer los resultados en forma pública.

Prácticamente todos los países evalúan logros en Lenguaje y Matemática, y existe una importante variedad de situaciones en cuanto a la evaluación de otras áreas del aprendizaje (ciencias naturales, ciencias sociales, autoestima, etc.), así como en cuanto a

los grados y niveles evaluados y la periodicidad de las evaluaciones.

La UNESCO, por su parte, ha impulsado el Laboratorio Latinoamericano de Medición de la Calidad de la Educación, que ha permitido desarrollar una primera experiencia de evaluación internacional en varios países de la Región.

Sólo como ejemplo de esta diversidad de enfoques, se puede mencionar que Chile ha optado por un esquema de evaluaciones de carácter censal con publicación de resultados en los medios de prensa, con el objetivo principal de informar a los usuarios acerca de la calidad del servicio que brinda cada establecimiento; mientras que México ha optado por un sistema masivo pero voluntario para los maestros, con el objetivo principal de usar la información como indicador de la calidad del trabajo del maestro y de allí derivar incentivos de carácter económico. En ambos casos las pruebas que se aplican son absolutamente confidenciales. Uruguay, por su parte, ha adoptado un enfoque centrado en el uso de la información como instrumento de aprendizaje al interior del sistema educativo, priorizando la devolución de resultados a cada establecimiento en forma confidencial y haciendo completamente públicas las pruebas luego de los operativos. Argentina ha desarrollado evaluaciones anuales de carácter muestral, con un fuerte énfasis en la producción de "cuadernos" de recomendaciones metodológicas para los docentes.

El panorama sucintamente presentado permite afirmar que **se ha dado un primer paso de enorme trascendencia para la Región: los sistemas se han instalado en los países, se ha generado una cierta capacidad para la implementación de operativos nacionales de evaluación a gran escala, y la sociedad y los cuerpos docentes comienzan a valorar y comprender la necesidad de este tipo de evaluaciones.**

Revisión de la experiencia y búsqueda de nuevos caminos

No obstante lo anterior, al cabo de una primera etapa de implantación de su sistema nacional de evaluación, muchos países se encuentran ingresando en una etapa de evaluación de su propia experiencia, y de consideración y estudio de nuevas alternativas para el desarrollo y rediseño de sus sistemas de evaluación hacia el futuro.

Ello obedece a la constatación de tres grandes tipos de insuficiencias en lo realizado hasta el momento:

- Insuficiente aprovechamiento de la información producida por los sistemas de evaluación, lo que tiene como consecuencia el insuficiente impacto de las evaluaciones en el conjunto del sistema educativo.
- Insuficiente calidad y capacidad de evaluación de aprendizajes complejos en las pruebas que están siendo aplicadas.
- Debilidades técnicas en los procesos de desarrollo y validación de los distintos instrumentos de medición.

De hecho, la mayoría de los sistemas de evaluación de la Región fueron construidos sobre un marco limitado de saberes respecto al diseño de pruebas, así como respecto al universo de opciones técnicas posibles, en un contexto caracterizado por la inexperiencia que había a principios de los '90 tanto en los Ministerios de Educación como en otras instituciones o centros especializados. A ello es preciso agregar el hecho de que normalmente en las comunidades académicas del mundo educativo en la Región existía un fuerte rechazo, de carácter más ideológico que técnico, a cualquier intento de medición en el área educativa.

En muchos países de la Región ha sido insuficiente la reflexión acerca de los fines específicos que se espera cumplan los sistemas de evaluación dentro de un país, así como de las definiciones técnicas más adecuadas para cada fin. Como se puede observar en el Recuadro 1, los sistemas de evaluación pueden perseguir distintos fines y es imposible que un mismo diseño sirva para satisfacerlos todos.

Recuadro 1
Finalidades de los sistemas de evaluación de aprendizajes

Un sistema nacional de evaluación de aprendizajes puede proponerse alguna o varias de las siguientes finalidades:

- 1. Evaluar la productividad de los maestros para establecer un sistema de incentivos.** En este caso se parte del supuesto de que un modo de lograr que los maestros enseñen mejor y los niños aprendan más, es estableciendo incentivos monetarios o de otro tipo para los maestros cuyos alumnos exhiben mejores niveles de logro. El papel principal del sistema de evaluación es producir información sobre los aprendizajes logrados en cada grupo escolar.
- 2. Brindar a los padres de familia información que les permita evaluar la calidad de las escuelas.** En este caso, se espera que el sistema de evaluación proporcione una medida de la calidad de la enseñanza que ofrece cada una de las escuelas, de modo que los padres estén en mejores condiciones para controlar su labor y para elegir la escuela que consideren mejor para sus hijos. Se supone que esto obligará a las escuelas a esforzarse por mejorar su trabajo.
- 3. Devolver información a las escuelas y maestros para que éstos examinen los resultados de su trabajo.** En este caso el sistema de evaluación debe producir información útil y detallada sobre lo que los alumnos están aprendiendo, para enriquecer la discusión técnica de los docentes y la búsqueda de nuevos caminos para mejorar la práctica pedagógica.
- 4. Establecer la acreditación de los alumnos que finalizan un determinado nivel de enseñanza.** En este caso, la evaluación debe permitir decidir si un estudiante individualmente considerado ha logrado los conocimientos y competencias indispensables para completar cierto nivel de enseñanza y obtener el certificado correspondiente.
- 5. Seleccionar u ordenar a los estudiantes.** El sistema de evaluación puede tener como objetivo ya no constatar si los alumnos dominan ciertos conocimientos y competencias, sino simplemente ordenar o jerarquizar a un conjunto dado de alumnos de acuerdo a sus niveles de dominio, por ejemplo, con vistas a un proceso de selección para el ingreso a distintas modalidades de educación superior o de formación para el trabajo.
- 6. Informar a la opinión pública y generar una cultura de la evaluación.** Aquí el propósito principal de la evaluación es producir información adecuada para rendir cuentas periódicamente ante la opinión pública de un país acerca de la marcha del sistema educativo, en términos de los niveles de aprendizaje que alcanzan los estudiantes en diferentes áreas disciplinarias y niveles del sistema, y su evolución a lo largo del tiempo. Esto ayuda a generar una cultura de evaluación del sistema educativo en la sociedad en general.
- 7. Contribuir a establecer estándares de calidad para el sistema educativo.** El sistema de evaluación puede tener como propósito explícito o implícito dar una señal a las escuelas y maestros acerca de qué conocimientos y competencias se espera que los alumnos dominen al finalizar un grado o nivel de la enseñanza o, en caso de que los mismos estén explícitamente definidos bajo la forma de estándares o indicadores de logro, evaluar el grado en que los mismos se alcanzan, a modo de mecanismo de control de la calidad del sistema educativo.
- 8. Construir un "mapa de situación" del sistema educativo con el fin de identificar áreas prioritarias de intervención y tipos de intervenciones necesarias.** La evaluación nacional de aprendizajes puede servir para detectar las regiones, distritos o establecimientos en que las dificultades para lograr los aprendizajes esperados son mayores, con el fin de facilitar el diseño de estrategias de intervención focalizadas y apropiadas. Del mismo modo, puede servir para identificar regiones, distritos o establecimientos con resultados especialmente buenos, con el fin de conocer y difundir sus modos de trabajar.
- 9. Evaluar el impacto de políticas, innovaciones o programas específicos.** En el marco de los procesos de reforma y cambio educativo en curso en todo el mundo, los Ministerios de Educación desean contar con información sobre los resultados de un nuevo curriculum que ha sido implementado en un conjunto de escuelas, un plan de capacitación de maestros o una inversión en nuevos materiales didácticos.
- 10. Realizar estudios de tipo costo-beneficio.** Una expectativa que muchas veces existe es que los sistemas de evaluación proporcionen información útil para evaluar los costos y beneficios en términos de inversión económica y resultados educativos de distinto tipo de intervenciones. Se busca por este camino apoyar los procesos de toma de decisiones, con el fin de que los recursos disponibles sean utilizados de manera efectiva y eficiente.
- 11. Contribuir a la generación de conocimiento.** Los sistemas de evaluación de aprendizaje generan importantes bases de información útiles para investigaciones que contribuyan a la acumulación de conocimiento sobre el funcionamiento de los sistemas educativos, las prácticas de enseñanza, el impacto de las variables sociales sobre el aprendizaje de los niños y los tipos de intervenciones más efectivos para mejorar los aprendizajes.

Muchos países han trabajado a partir de un propósito general de informar sobre los resultados del sistema educativo para contribuir a su mejoramiento, pero sin diseñar una estrategia más específica. Por otra parte, es bastante común que las autoridades ministeriales comiencen a demandar, sobre la marcha, que las evaluaciones sirvan para nuevos propósitos o que aporten información para fines para los que no fueron diseñadas.

Asimismo, normalmente no se cuenta con un plan de trabajo detallado de largo plazo respecto al desarrollo del sistema de evaluación y sus objetivos, que permita diseñar las estrategias adecuadas a los diferentes tipos de fines y ordenar las decisiones técnicas sobre la conformación de las bases de datos, la conformación de los bancos de ítemes y la comparabilidad de las evaluaciones, entre otras cosas.

En el presente, sin embargo, **parecen estar dadas las condiciones para realizar un “salto cualitativo” en materia de evaluación, una vez que se ha transitado por las primeras experiencias, que se ha superado la preocupación inicial por las enormes exigencias de la implementación de los operativos de medición a gran escala, y que existen sistemas funcionando que permiten debatir el tema ya no en abstracto, sino a partir de experiencias en marcha.**

Dar un “salto cualitativo” exige desarrollar un proceso de estudio y análisis en dos grandes planos, íntimamente relacionados entre sí:

- La discusión sobre las **opciones de política** en materia de evaluación nacional de aprendizajes: qué impactos específicos se espera que tengan los sistemas de evaluación en el sistema educativo, más allá de la definición genérica de la responsabilidad ante la sociedad y la mejora de la calidad.
- La discusión sobre las **opciones técnicas**: qué abanico de tipos de pruebas, instrumentos complementarios, procesamientos y análisis de información existen y cuáles son los más adecuados para los fines propuestos.

Propósitos de este documento

El presente documento pretende aportar a la reflexión sobre la relación entre las finalidades de política educativa que los sistemas de evaluación pueden proponerse y sus implicaciones técnicas. A través del mismo, se espera enriquecer el repertorio de alternativas

disponibles a la hora de reflexionar sobre los rumbos a seguir por los sistemas de evaluación de aprendizajes de la Región en el futuro.

El documento fue elaborado en el marco de un Taller de Trabajo convocado por GRADE, el cual se desarrolló en Lima (Perú) entre el 17 y el 20 de agosto de 1999. Se convocó a un conjunto de especialistas en el tema con conocimiento directo de distintos sistemas nacionales de evaluación de la Región, así como de los dilemas, opciones y revisiones a las que muchos de los países se han visto enfrentados.

Cuatro grandes temas o problemas fueron seleccionados en el Taller para organizar el documento, a cada uno de los cuales corresponde un capítulo del mismo:

- 1. El problema del diseño global del sistema nacional de evaluación.** Se intenta ofrecer una visión sistemática acerca de la relación entre los diversos fines del sistema de evaluación y las decisiones técnicas relacionadas con la cobertura de las mediciones, tanto en términos poblacionales (tamaño de las muestras y/o censo) como en términos de contenidos (nivel de detalle de los conocimientos y competencias a evaluar dentro de un área o disciplina). Junto con abordar este tema en el primer capítulo, se analiza en el capítulo quinto la relación entre los fines del sistema de evaluación y los modos de reportar los resultados, y se presenta un conjunto de alternativas técnicas para esos efectos.
- 2. El problema de la validez de los instrumentos de medición,** un tema hasta ahora insuficientemente atendido por los sistemas nacionales de evaluación de la Región, que está directamente vinculado con su valor como insumo para la toma de decisiones.
- 3. El problema de los paradigmas de construcción de pruebas de medición de aprendizajes o logros.** Se examinan las características, exigencias, posibilidades y limitaciones que caracterizan a los paradigmas de pruebas referidas a «normas» y pruebas referidas a «criterios», y se reflexiona sobre el problema de la validez en cada uno de los paradigmas.
- 4. El problema de los factores “asociados” a los resultados escolares.** La mayor parte de los sistemas de evaluación de la Región, junto con la aplicación de las pruebas, recogen un importante volumen de información de carácter contextual, tanto sobre las características socioculturales de los alumnos como sobre las propias escuelas y maes-

tros. Sin embargo, este tipo de información es escasamente difundida y poco utilizada en el análisis de los resultados de aprendizaje.

El documento fue elaborado a través de aproximaciones sucesivas de discusión colectiva, redacción individual por parte de los participantes, lectura y discusión de lo producido, y un nuevo proceso de redacción individual. Se realizaron tres ciclos de este tipo durante el Taller y luego se continuó con ajustes y correcciones durante un mes.

Si bien todos los contenidos del documento fueron discutidos en forma colectiva, la redacción de cada uno de los capítulos estuvo a cargo de una persona que se indica junto al título de cada capítulo. Participaron también Patricia Arregui y Santiago Cueto en las instancias de discusión colectiva del trabajo y en la revisión de los sucesivos borradores.

Es el deseo de todo el grupo de trabajo que el documento sea útil para enriquecer la discusión sobre los sistemas de evaluación de aprendizajes en la Región y para ampliar la mirada hacia el futuro en un área estratégicamente central para el mejoramiento de los sistemas educativos.

Capítulo I

EL DILEMA DE LA “GRANULARIDAD” EN EL DISEÑO DEL SISTEMA DE EVALUACIÓN: COBERTURA CURRICULAR VS. COBERTURA POBLACIONAL

Richard Wolfe

En las evaluaciones nacionales, ¿es preferible trabajar con muestras o hacerlo a nivel censal?

¿Es preferible emplear una única prueba o diferentes formas con distintos ítemes?

¿Con qué grado de desagregación es posible y deseable reportar los resultados?

¿Con qué grado de profundidad es posible y deseable medir los conocimientos y competencias adquiridas por los alumnos?

¿Es adecuado el modo en que los diseños de los sistemas de evaluación toman en cuenta todos estos aspectos?

Introducción

Los diseños de los sistemas nacionales de evaluación educacional en América Latina son muy variados y sus características dependen de las filosofías, estructuras, costumbres burocráticas e historias de la educación específicas de cada país, de las etapas de reforma educativa en que se encuentren y de los estados de desarrollo de la investigación y planificación educativas.

Pero al mismo tiempo, cuando se examinan en detalle los diferentes sistemas de evaluación, se observan algunas características comunes derivadas del hecho de tener objetivos fundamentales y requerimientos técnicos similares.

El propósito de este capítulo es examinar un asunto crítico en el diseño de los sistemas nacionales de evaluación: las cuestiones sobre la denominada “granularidad”, es decir, la cantidad de detalle con que el sistema recoge y luego reporta los datos. Por ejemplo, puede haber enormes diferencias en el costo y en el modo de utilización entre sistemas de evaluación que sólo proporcionan resultados nacionales y aquéllos que suministran resultados de to-

dos los estudiantes o escuelas individualmente. De igual manera, puede haber grandes diferencias entre las evaluaciones que dan información general sobre temas amplios —tales como los logros en matemáticas o lenguaje— y aquéllas que brindan información detallada sobre lo que los estudiantes pueden y no pueden hacer en esas asignaturas.

Para analizar el tema de la granularidad es necesario abordar sus dos dimensiones: quién es evaluado y qué es evaluado.

¿Quién es evaluado?

Si bien los estudiantes —y a menudo los padres, los profesores, los directores escolares y otros— son las fuentes primarias de datos en un sistema de evaluación educacional, no suelen ser la principal unidad para la cual se calculan los resultados y se hacen los reportes, salvo en los casos en que se trata de exámenes de certificación o graduación. La granularidad de los reportes, es decir, la unidad más pequeña respecto a la cual se brinda información sobre sus resultados, suele establecerse en niveles superiores de la estructura educativa.

En los sistemas de evaluación de América Latina, comúnmente se encuentran los siguientes niveles de análisis y de reporte:

- Poblaciones nacionales (o internacionales), tales como la población de escolares matriculados en tercer grado de educación primaria.
- Los principales estratos definidos educacional, política y socialmente, tales como estudiantes en escuelas públicas, en las escuelas rurales, o estudiantes en programas bilingües.
- Principales divisiones regionales, tales como regiones geográficas, provincias, o estados.
- Jurisdicciones menores, tales como ciudades o municipalidades.
- Escuelas.
- Salones de clase (o profesores).
- Estudiantes.

Además de la selección entre estos niveles de reporte, en todos los niveles se diferencia según el grado escolar.

La elección del nivel o los niveles de reporte debería depender del objetivo y de los usos del sistema de evaluación. También debería determinar la metodología general para llevar a cabo la evaluación. Por ejemplo, si se van a devolver los resultados a cada estudiante o si cada profesor o escuela va a ser calificado individualmente, es obvio que se necesita una aplicación censal. Por otro lado, si sólo son necesari-

Richard Wolfe. Especialista en medición, evaluación y estadística educacional, graduado de la Universidad de Chicago. Actualmente es Profesor Asociado del Departamento de Currículo, Enseñanza y Aprendizaje del Ontario Institute for Studies in Education de la Universidad de Toronto, Canadá. Ha asesorado en las áreas de su especialidad a varias instituciones gubernamentales e internacionales en las Américas.

rios los resultados nacionales o los principales resultados sub-nacionales para rastrear la productividad general y el cambio en el sistema educacional, se puede usar una encuesta por muestreo, lo cual es mucho más económico.

En efecto, **existe una relación sumamente importante entre la granularidad del reporte y su costo. Cuanto más detallada información se requiera, más costoso es suministrarla.** Es un hecho básico del muestreo estadístico que el tamaño requerido de una muestra para un nivel dado de precisión es principalmente una función del tamaño de la muestra y no, como intuitivamente muchos pensarían, del tamaño de la población. Por ejemplo, si es necesario obtener información igualmente precisa para cada provincia de un país, entonces los requerimientos de tamaño de la muestra serán igualmente altos para las provincias con pocos estudiantes como para las provincias con muchos estudiantes. El tamaño de la muestra agregada será muy grande en comparación con lo que sería necesario si el único requerimiento fueran estadísticas nacionales precisas. Por otra parte, un muestreo proporcional o una muestra simple al azar rinde buenos resultados generales y resultados razonables para amplias subpoblaciones (e.g., grandes provincias), pero obtener una alta precisión para pequeñas subpoblaciones requiere un sobremuestreo costoso.

¿Qué es lo que se evalúa?

¿Con qué tipo de detalle se calculan y presentan los resultados del potencial dominio de contenido? Este es otro aspecto de la granularidad en el diseño de los sistemas de evaluación, que también tiene conexiones importantes con los objetivos de un sistema de evaluación e implicancias para la metodología.

En los diferentes tipos de sistemas de evaluación, nos encontramos con estos niveles de reporte de la información:

- Resultados globales, incluyendo matrícula, participación en la evaluación, sin una verdadera evaluación de contenidos.
- Éxito o fracaso en general, culminación del plan de estudios, graduación, certificación, tal vez basados en evaluaciones de diferentes asignaturas y otra información.
- Puntajes en asignaturas, tales como el logro general en matemáticas o en lenguaje.
- Puntajes en áreas de asignaturas, tales como solución de algoritmos, álgebra, o geometría en matemáticas y comprensión de lectura, expresión escrita o convenciones gramaticales en lenguaje.
- Logro de niveles particulares de desempeño en

diferentes estándares en un área o asignatura, tales como la competencia para aplicar métodos geométricos en la solución de problemas de distancia, o interpretaciones a nivel de principiante en lecturas literarias.

- Estadísticas de respuestas para ítemes específicos, tales como el porcentaje correcto en un ítem de opción múltiple o el porcentaje de calificaciones que se sitúan en cada nivel de una tarea de desempeño.
- Registro detallado de las respuestas a una prueba, incluyendo patrones de distribución de las respuestas a los ítemes, transcripciones de desempeños o resultados cognitivos en laboratorio.

La granularidad del contenido de una evaluación determina fuertemente nuestra capacidad de interpretar y comprender la calidad del logro educacional y de tomar medidas para mejorarla. Por ejemplo, los resultados generales de logro respecto a un plan de estudios pueden ayudar a localizar áreas de éxito general relativamente alto o bajo (por ej., mejores escuelas o tipos de escuelas). Pero un conocimiento más detallado de la substancia y el contenido de esos logros permite evaluar la importancia y las consecuencias de tales diferencias, y legitimará, a su vez, medidas tales como la selección o el establecimiento de incentivos.

La granularidad del contenido también determina nuestra capacidad para usar información de la evaluación para diseñar ajustes al currículum y a la enseñanza. Con información detallada y en profundidad acerca de los contenidos, se puede llegar a comprender cuáles aspectos de un currículum son aprendidos exitosamente y se pueden hacer recomendaciones específicas sobre la secuencia curricular y las prácticas de enseñanza. Ello implica un desplazamiento de preguntar *cuánto* saben los estudiantes a preguntar *qué* saben y *qué* son capaces de hacer.

Las evaluaciones de grano más fino son generalmente más costosas, porque el número de ítemes requerido para cubrir en detalle un área de contenido es alto. Primero, habrá un número relativamente grande de sub-contenidos dentro de un área de contenido. Por ejemplo, en matemáticas tenemos áreas generales como aritmética, geometría o álgebra, y áreas más específicas tales como fracciones, adición de fracciones, formas geométricas, congruencia, series y secuencias. Segundo, cada subcontenido requiere un número suficiente de ítemes, tal vez cinco o diez, para suministrar una muestra adecuada de los posibles desempeños o niveles de desempeño. Asimismo, en la evaluación será necesario contar con una muestra adecuada de respuestas de los escolares a cada uno de los ítemes.

El requerimiento de un número mínimo de ítems tiene un fundamento sustantivo y otro estadístico. En cuanto a lo sustantivo, se necesita ver un número suficiente de ejemplos de lo que es difícil y de lo que es fácil para comprender los tipos de conocimiento y destrezas que poseen los estudiantes. Estadísticamente, se necesita obtener una buena medida del desempeño promedio de los estudiantes, de la variación entre los ítems y de la interacción entre los ítems y los estudiantes (algunos estudiantes hacen algunas cosas bien, otros estudiantes hacen otras cosas bien).

Por otro lado, una muestra relativamente pequeña de contenidos y de ítems puede ser suficiente si el propósito del reporte es suministrar unos promedios simples que resumen la situación en un dominio de contenidos, aunque ello no constituiría un diagnóstico del currículum.

Otra cuestión es que la evaluación de estándares educacionales importantes requerirá a menudo determinar si los estudiantes pueden llevar a cabo tareas complejas e integradas. Por ejemplo, en matemáticas no basta con interesarse sólo si los estudiantes pueden sumar o restar, sino más bien si pueden usar la aritmética en contextos novedosos y realistas para resolver problemas. Desde la perspectiva de la evaluación, este aspecto es simultáneamente de “grano grueso”, porque se refiere a un estándar general de la educación y atraviesa diferentes contenidos y áreas; y de “grano fino”, porque requiere la definición de tareas de desempeño particulares y la recopilación, calificación y análisis de registros de desempeño.

Tipología de sistemas de evaluación

A partir de estas dos dimensiones de la granularidad —quién es evaluado y qué es evaluado— se puede definir una tipología de sistemas de evaluación cuya granularidad se grafica en la Figura 1.

A. Estadísticas educacionales. Representan una forma simple de sistema de evaluación, común a todos los países. La mayoría de las estadísticas educacionales contemplan cantidades de estudiantes (tal vez diferenciados por grado, sexo, iniciación de *repetidor* o *promovido*); información respecto a dónde se ubican los estudiantes (en escuelas, distritos y niveles más altos del sistema educativo). Estas estadísticas suelen ser recopiladas mediante una metodología de censo, puesto

que todos los estudiantes y escuelas deben ser contabilizados y se requiere una desagregación muy fina de la ubicación. Sin embargo, no hay ningún contenido que diferenciar.

B. Programas de evaluación y certificación. Desde la perspectiva de la granularidad, un programa de este tipo está muy cercano a un sistema de estadística educativa, pero refinado en ambas dimensiones (el qué y a quién). El reporte de los datos se extiende hasta el nivel del individuo. Hay, además, una medición apropiada de contenidos, otorgada por una prueba de graduación o salida. Una mayor desagregación de los contenidos puede ser deseable, pero no es esencial. Puede también realizarse un resumen estadístico a distintos niveles de la jerarquía educacional por requerimiento administrativo. Los programas de evaluación y certificación son realizados de manera censal en las poblaciones afectadas. A menos que haya un propósito secundario de diagnóstico del éxito o fracaso estudiantil, o criterios múltiples para la certificación, hay poco interés en la diferenciación del contenido.

C. Programas de evaluaciones diagnósticas. Tienen el propósito específico de diagnosticar algunas características educacionales y psicológicas de los individuos (por ejemplo, dislexia). Suelen buscar identificar a los estudiantes con dificultades en el aprendizaje, para asignarlos a programas especiales o remediales. Los datos que se reportan son estrictamente individuales. Existe cierta diferenciación de contenido relacionada con los tipos de dificultades de aprendizaje que están siendo diagnosticados. Estos programas generalmente no son considerados evaluaciones nacionales de educación, salvo tal vez en pruebas de despistaje a gran escala y suelen ser implementados por personal local especializado.

D. Evaluación nacional por muestreo. Es un tipo de evaluación realizada en diversos países mediante relevamientos muestrales de ítems y de individuos. Lo usual es dividir un conjunto muy grande de ítems en múltiples formas de prueba para su administración, y se aplica cada forma a muestras paralelas de estudiantes, a través de algún sistema de rotación en la aplicación. En la Figura 1, la recolección y reporte de datos de este tipo de evaluación se muestra como un triángulo, porque suele ser posible una mayor diferenciación de los contenidos en los niveles más altos de agregación. Esto es simplemente una consecuencia de la precisión de las muestras. Como el número de estudiantes que responden un ítem particular es relativamente pequeño, no se tendrá una precisión ade-

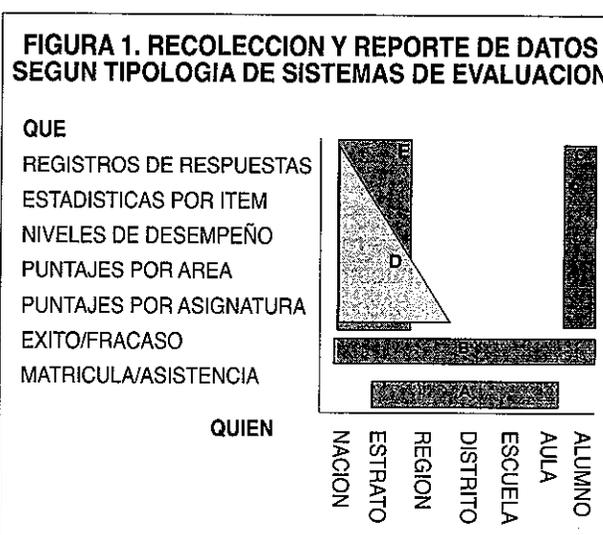
cuada para reportar estadísticas acerca de las respuestas dadas a ese ítem por diferentes poblaciones. En cambio, si se agregan los puntajes correspondientes a varios ítems, es posible trabajar conjuntamente los resultados de los estudiantes para proporcionar información más detallada sobre unidades más finas. Se podría incluso calcular puntajes individuales razonables en base a todo el conjunto de contenidos.

E. Estudio de investigación curricular mediante muestreo. Busca establecer los tipos de aprendizaje y enseñanza que se dan en una asignatura y estudiar las relaciones entre enseñanza y aprendizaje, así como los efectos de los contextos educativos y sociales. En la Figura 1, el área de la recopilación y reporte de datos está representada como un rectángulo en el cual todo tipo de diferenciación de contenidos es importante, pero no hay gran interés por informar sobre logros que no sean a nivel nacional o de los estratos más altos del sistema. Esto no quiere decir que las variables referidas a los estudiantes y las escuelas no sean importantes, sino que estas unidades (estudiantes, profesores, escuelas) son anónimas. Puede ser muy importante, por ejemplo, estudiar el impacto de las características de la escuela, las prácticas del profesorado y los antecedentes de los estudiantes en los resultados educativos, pero no interesa informar individualmente sobre las escuelas, clases o estudiantes.

El dilema de la granularidad

El propósito de detallar el tema de la granularidad y de suministrar una tipología de los sistemas de evaluación, ha sido exponer un dilema crucial que debe ser confrontado en el diseño de un sistema de evaluación educativa.

El dilema es que las dos dimensiones de la granularidad están en conflicto. Ello significa que, para un costo y esfuerzo fijos, un incremento en la granularidad de un tipo debe correr de la mano con una disminución de la granularidad del otro tipo. Por ejemplo, si se necesita obtener medidas detalladas sobre cada municipalidad del país, probablemente habrá que usar una prueba muy breve y simple, y la cantidad de detalle en los contenidos será mínima. Si, por otro lado, se quiere tener una gran profundidad en la medición de un dominio de contenidos, habrá que usar rotaciones de ítems y múltiples formas de prueba, que harán que el detalle de información quede muy disperso en los niveles más finos de reporte, en especial para los estudiantes individuales y posiblemente para las aulas, escuelas y niveles intermedios.



Un desafío central para el diseño de una evaluación es desarrollar metodologías que permitan combinar diferentes propósitos de la mejor manera. Por ejemplo, para el TIMSS se desarrolló un sistema muy intrincado y cuidadosamente diseñado de formas de prueba. Cada forma contiene una muestra estratificada de ítems que provienen de todo el campo de contenidos, además de algunos ítems constantes que sirven para la calibración. Muchas de las formas contienen tareas de desempeño que insumen la mayor parte del tiempo de administración, mientras que otras formas incluyen casi exclusivamente ítems de opción múltiple o de respuesta corta. La administración de las formas fue llevada a cabo con una rotación y balance cuidadosos dentro de cada aula y escuela de la muestra. Si bien el tamaño de la muestra total de TIMSS no era grande (alrededor de 200 aulas y 7.000 estudiantes por grado en cada país), hay una enorme cantidad de información disponible para un análisis y examen detallados de los aprendizajes en matemáticas y ciencias y sus relaciones con los factores asociados.

A manera de contraste, se puede decir que hay varios sistemas nacionales de evaluación en América Latina que trabajan con muestras de mucho mayor tamaño o que son llevados a cabo como operaciones censales, que tienen mucho menor detalle de los contenidos y que, simultáneamente, no reportan información más allá de los estratos de agregación más altos.

En este sentido, parece recomendable que se otorgue mayor atención al análisis cuidadoso de estos aspectos de la granularidad en el diseño de las evaluaciones de rendimiento escolar en la Región.

Capítulo II

LA INTERPRETACIÓN JUSTIFICADA Y EL USO APROPIADO DE LOS RESULTADOS DE LAS MEDICIONES

Gilbert Valverde

¿Qué significan los resultados que obtienen los estudiantes en nuestras pruebas nacionales de rendimiento?

¿Estamos realmente sacando conclusiones apropiadas, significativas y útiles a partir de los resultados de las evaluaciones?

¿En qué medida podemos justificar la manera en que interpretamos estos resultados?

¿Se usan los resultados de las evaluaciones de manera apropiada en la toma de decisiones?

Introducción

Cuando los sistemas de evaluación conducen sus actividades, su interés es descubrir, describir e interpretar facetas del sistema educativo. Un propósito que comparten todos los sistemas de evaluación en América Latina es el de comprender qué capacidades académicas adquieren los niños y las niñas como resultado de su asistencia y participación en las escuelas y colegios del país. En el lenguaje curricular y evaluativo, a esas capacidades adquiridas como resultado de la escolarización comúnmente se las denomina *logro*.

Es posible que la estrategia óptima para comprender cómo se da el logro sea registrar el tipo de éxito que el estudiante experimenta al enfrentar cada una de las situaciones que aprendizaje en los que participa al año. Esto es lo que los docentes comúnmente intentan hacer como parte de su labor de evaluación en el aula.

Pero también hay diversos actores que esperan obtener otros tipos de información de las evaluaciones, tales como:

- Las autoridades políticas y la sociedad civil tienen interés por obtener información acerca del sistema educativo. Por ejemplo, cómo es la calidad de la educación en comparación con los estudiantes de otros países o con grupos de estudiantes en

generaciones anteriores en su propio país, o en relación con los propósitos académicos que el sistema mismo se ha fijado para sí mismo.

- Hay quienes están preocupados por los aspectos de equidad y necesitan descubrir si el sistema educativo favorece en forma desigual a distintos grupos económicos, culturales o lingüísticos.
- Otros desean información útil para juzgar la eficacia de distintos tipos de inversiones o intervenciones que se proponen hacer en el ámbito nacional en la educación.

Resulta casi evidente que la estrategia "óptima" mencionada anteriormente no se ajustaría a sus requerimientos, ya que sería imposible realizar un seguimiento a todos los estudiantes de un país o a un número representativo de ellos.

Hasta la fecha, la estrategia que se sigue en todos los sistemas de evaluación en América Latina es la de plantear una situación relativamente novedosa a los estudiantes, que dura uno o dos periodos lectivos, donde los éstos deben demostrar que han adquirido un número significativo de las capacidades esperadas. En todos los países de la Región, el tipo de situación que plantea el sistema de evaluación es una *prueba escrita* con preguntas que, a criterio de los autores de la prueba, exigen que los estudiantes utilicen lo que aprenden en la escuela para contestarlas correctamente. Las representaciones que más típicamente arrojan las pruebas en América Latina, son números llamados *promedios* o *notas*, cuyo significado debe ser bien entendido por las personas encargadas de interpretar estos números.

Ahora bien, ¿cuán fieles son estas representaciones?

Interpretar correctamente y usar apropiadamente la información que dan las pruebas, significa que hay que preocuparse por entender el tipo de representación del logro que éstas permiten. Esto significa preocuparse por lo que, en medición, se llama *validez*.

La validez no es una propiedad intrínseca de las pruebas o encuestas, sino una propiedad de las interpretaciones y los usos que se propone dar a los datos que se obtienen de ellas. Es así que actualmente se define la validez como *el grado en que la evidencia empírica y la teoría dan sustento a las interpretaciones de los resultados de una medición*. Asimismo, la validez se refiere al *ámbito del legítimo uso de esas interpretaciones* y también al *grado en que el uso de la prueba no produce un impacto negativo no deseado sobre el sistema educativo*. En otras palabras, la

Gilbert A. Valverde, Ph.D., Department of Educational Administration and Policy Studies, School of Education, University at Albany, State University of New York, Education 313 A, Albany, New York 12222. Fono: (518) 4425089, Fax: (518) 4425084
E-mail: valverde@uemail.albany.edu

validez se refiere a la calidad de las conclusiones que se toman a partir de las mediciones y a las consecuencias que las mediciones generan en los procesos que se proponen medir¹.

La validación de pruebas en América Latina. Ejemplos.

Algunas situaciones que se dan en América Latina sirven para ejemplificar algunos tipos de preocupación por la validez de las evaluaciones que se realizan en la actualidad:

Caso 1.

¿Habilidades de resolución de problemas o simple memorización?

El Ministerio de Educación se encuentra implementando un nuevo currículum nacional de Matemáticas, cuyo enfoque principal es que los estudiantes aprendan a resolver problemas novedosos de la vida real utilizando elementos de razonamiento matemático. Sin embargo, para medir el logro, se administra una prueba escrita cuya mayoría de preguntas o reactivos exigen a los estudiantes que recuerden términos y principios matemáticos, o sólo requieren que ellos apliquen procedimientos rutinarios para resolver problemas o ejercicios muy parecidos a los que aparecen en sus libros de texto.

En este caso, el Ministerio de Educación claramente no cuenta con un instrumento apropiado para descubrir si los estudiantes han logrado dominar las capacidades que persigue el nuevo currículum nacional. Sería injustificado concluir que los estudiantes que obtienen un alto promedio en esta prueba poseen la capacidad de resolver problemas novedosos de la vida real, porque las preguntas no exigen que recurran a este tipo de habilidades.

Caso 2.

¿Medición de conocimientos o de habilidad lectora?

Se escribe una prueba para descubrir si los estudiantes de 7 años de edad están adquiriendo conocimientos acerca de ciencias naturales. En las aulas se enseñan estos contenidos sin texto escolar, usando elementos del entorno natural de la escuela. La prueba contiene muchas preguntas cuya comprensión exigiría que los niños y las niñas posean gran habilidad para comprender textos escritos y un vocabulario altamente desarrollado.

En una prueba de esta naturaleza, el significado de los promedios es sumamente difícil de descubrir. ¿Acaso un bajo promedio indica la no-adquisición de los conocimientos que se pretendía medir, o más bien mide la habilidad lectora de los niños? En el caso de

niños pequeños, ¿en qué medida son las supuestas pruebas de ciencias (o de matemáticas, ciencias sociales, etc.) en realidad pruebas de lectura?

Caso 3.

¿Un problema de eficacia educativa o de acceso a recursos?

Se administra una prueba de logros a todos los estudiantes de octavo grado en un país. El Ministerio de Educación utiliza los resultados obtenidos por los estudiantes en cada escuela para calcular el promedio de logro para cada establecimiento. Comparando los promedios de los establecimientos según éstos sean privados o públicos se descubre que los promedios de las escuelas privadas son más altos que los de las públicas. Se concluye que las primeras son más eficaces que las segundas, aun cuando ocurre que éstas no cuentan con textos que aborden uno de los temas más importantes de la prueba.

Aquí es muy problemática la interpretación que se propone para los resultados, ya que un recurso esencial para el aprendizaje de un área de contenido o competencia específico (libro de texto que cubra temas medidos en la prueba) no se encuentra repartido equitativamente en los establecimientos. ¿Se justifica la interpretación de un bajo promedio como indicador de falta de eficacia del establecimiento? ¿No habría que interpretarlo como indicador de una falta de equidad en la distribución de los recursos?

Caso 4.

¿Dominio general de conocimientos o la suerte de dominar el área de contenidos que más se midió?

En un país se utiliza una prueba a final de la educación secundaria o media para avalar un diploma que se otorga al egreso de ese nivel. Se interpreta que pasar esta prueba indica que un estudiante ha logrado dominar todos los objetivos del currículum propuestos para cada año en ese nivel. En la prueba se miden algunos aspectos del currículum con una variedad de preguntas, otros con muy pocas. Se otorga el diploma correspondiente a todos los estudiantes que aprueban.

Preocupa en este caso si la conclusión de que un estudiante domina los objetivos del nivel se puede defender si no se mide con igual rigor los distintos componentes del currículum.

Caso 5.

Pertinencia de la evaluación educativa y el peligro de las comparaciones generales.

En un país no existe un currículum nacional, sino que cada provincia tiene el suyo propio. La Secretaría de

Educación administra una prueba en todas las provincias. Para garantizar que la prueba es justa para todas, se decide poner sólo preguntas sobre aquellos temas que se enseñan en todas ellas, lo que significa que se evalúa un subconjunto de las cosas que en cada provincia se pretende enseñar. Comparando los promedios de cada provincia, se encuentra que en algunas se obtienen resultados muy superiores que en las demás. Se concluye que es mayor la eficacia de los establecimientos en aquellas provincias. Sin embargo, ocurre que en las provincias de alto rendimiento, se pretende enseñar muy pocos temas que no están en la prueba nacional. En las provincias de más bajo rendimiento, los temas que se evalúan en la prueba nacional representan sólo una pequeña parte de los temas que se proponen enseñar, y no se les dedica mucho tiempo lectivo ni espacio en los libros de texto.

¿Es pertinente hacer una comparación entre los resultados de las provincias cuando en algunas de ellas se está enseñando una mayor proporción de los temas evaluados que en otras? ¿Acaso los promedios diferentes obtenidos de esta manera indican diferencias en eficacia educativa? ¿No será más bien que estos distintos promedios indican diferencias en la pertinencia de la prueba para cada una de las provincias?

Caso 6.

Decisiones de inversión a partir de limitadas evidencias.

Se diseña una prueba de lenguaje que entre sus preguntas contiene una sola en la cual los estudiantes escriben un texto propio. Al revisar este texto, se califican aspectos de ortografía, gramática y otras características de la escritura. El Ministerio de Educación desea distribuir material de apoyo pedagógico para docentes de lenguaje, pero para usar mejor su presupuesto, pretende descubrir los aspectos más débiles de los logros de los estudiantes, para lo cual se fija en los resultados de la prueba. Observa que la mayor parte de los estudiantes tuvieron mal rendimiento en la pregunta donde se pedía que escribieran su propio texto. En consecuencia, se escriben módulos de apoyo pedagógico y se proporciona capacitación a los docentes para ayudarlos a enseñar mejor gramática y expresión escrita.

¿Acaso la falta de éxito en contestar una sola pregunta es suficiente para concluir que los estudiantes no dominan esas capacidades? Si el Ministerio cuenta con recursos limitados para esfuerzos de refuerzo pedagógico y trata de utilizar los resultados de la evaluación para sacar provecho máximo de su inversión en ella, ¿ha utilizado en forma apropiada los resultados de la evaluación? Por otro lado, si los docentes adquieren, mediante los módulos y capacitaciones,

la convicción de que deben dedicar mucho más esfuerzo a enseñar gramática y expresión escrita, ¿ha sido apropiada la información para ocasionar ese cambio en las prioridades de los docentes?

Opciones para la validación de mediciones en educación

Las situaciones anteriores ejemplifican los problemas que existen en torno a las interpretaciones justificadas y al uso apropiado de la información que arrojan las mediciones. A menudo se distorsionan los significados reales, lo que afecta su validez y, en consecuencia, su pertinencia como insumo para la toma de decisiones. Dado que éste es un riesgo ineludible en la medición, **es importante explicitar de antemano los tipos de uso para los cuales los resultados podrán ser empleados legítimamente, así como los fines para los cuales los resultados NO podrán utilizarse de manera justificada.**

El proceso de acumulación de evidencias que dan sustento a las interpretaciones que se proponen para una medición se denomina *validación*. Existe una gran cantidad de opciones en cuanto al tipo de evidencia que se puede acumular y reportar. Cada tipo de evidencia ilumina o da apoyo a distintas facetas de la validez, pero no representa un tipo distinto de validez. La validez es un concepto unitario que obliga a los diseñadores y usuarios a evaluar de manera integral toda la evidencia disponible sobre cuán bien están justificadas las interpretaciones de los datos y las maneras de utilizar la información recogida durante la aplicación de la medición.

En el caso de las pruebas de logro, sean éstas referidas a normas o a criterios, se pretende derivar conclusiones que van más allá de las preguntas que componen las pruebas. Es decir, en ambos casos se reconoce que las preguntas que contiene la prueba representan solamente una pequeña muestra de todas las preguntas posibles que se podrían formular para conocer si los y las estudiantes poseen ciertas capacidades. De los análisis de cualquiera de los dos tipos de pruebas mencionadas se concluye que si los estudiantes contestan con éxito 80 por ciento de las preguntas formuladas en la prueba, serían también capaces de contestar con éxito 80 por ciento de todas las preguntas posibles que se podrían formular para medir esa capacidad.

Una forma obvia de proceder para sustentar esta conclusión es mediante una definición clara de lo que se quiere medir. Una vez que se cuenta con esa definición, es posible comparar cada pregunta que se propone para la prueba y juzgar su concordancia con la

definición. Si las preguntas de la prueba se han escrito de acuerdo a una definición precisa de lo que se pretende medir, las inferencias que se realicen con respecto al desempeño de los y las estudiantes en esas preguntas serán más válidas que en el caso contrario. Desde este punto de vista, la validación es un proceso inherente al procedimiento que se sigue para diseñar pruebas referidas a criterios (ver el capítulo al respecto en este mismo volumen), puesto que la definición del dominio (en términos de campo de conocimientos o habilidades) y el esfuerzo por asegurar la concordancia de las preguntas con el dominio definido son dos de sus preocupaciones centrales. Cuando se desarrolla y aplica este tipo de pruebas, la documentación de las definiciones de los dominios, los juicios acerca de la concordancia de las preguntas con los dominios y los pasos seguidos para asegurar que los dominios representen con justicia el currículo o los estándares, sirven a dos propósitos: guían el desarrollo de la prueba y documentan la evidencia de la validación de la medición propuesta.

Frecuentemente se propone también que las pruebas sean interpretadas con relación a un criterio externo. Esto es típico, por ejemplo, de las pruebas de admisión a la educación superior. En esos casos se establece (con mayor o menor grado de fundamento) que un promedio determinado *predice* una exitosa carrera universitaria. En algunos países, se pretende establecer que un diploma de educación secundaria –avalado por una prueba de bachillerato– certifica que el diplomado posee ciertas capacidades básicas como posible empleado, de modo tal que se supone que el éxito en la prueba predice una exitosa carrera como trabajador.

Aun en los casos en que no existe un criterio externo propuesto explícitamente para la prueba, la utilización de referentes externos puede reforzar la validación de las pruebas. Por ejemplo, cuando se compara dos formas de medir la misma competencia y ambas formas arrojan resultados semejantes, esto puede dar evidencia para la validación.

Para la validación de los resultados que generan las pruebas, es de suma importancia que los servicios nacionales de evaluación educativa publiquen informes técnicos que contesten las siguientes preguntas con claridad:

- **¿Acerca de cuáles capacidades o destrezas se derivarán conclusiones?** En esos informes se debe incluir una definición explícita de las capacidades que interesan, así como de aquéllas que se pretende evitar que debiliten la validez de la medición de las primeras. Por ejemplo, debe explicarse cómo se ha procurado que la habilidad para

leer no obstaculice la oportunidad que tienen niños de corta edad de demostrar lo que saben en una prueba de ciencias naturales.

- **¿Cómo se aseguró concordancia entre las preguntas y las capacidades o destrezas que se propuso medir?** Es necesario documentar los procedimientos del caso y describir en detalle el resultado de su uso. Por ejemplo: ¿cómo se utilizaron las definiciones a la hora de escribir preguntas o cómo procedieron los jueces para asegurar la concordancia entre las preguntas y los dominios a medir? ¿De qué manera se recogieron y analizaron sus juicios?
- **¿Qué tipos de preguntas permiten comprobar que se dominan las capacidades?** Por ejemplo, si se quiere comprobar si los estudiantes pueden resolver problemas novedosos de la vida real en matemáticas o producir textos propios legibles, coherentes y persuasivos, ¿se puede usar preguntas en las cuales los estudiantes escogen la opción correcta entre cuatro o cinco posibilidades? ¿Acaso la habilidad de reconocer la respuesta correcta entre distintas opciones es idéntica a la generación de una respuesta propia? ¿O se necesitan más bien preguntas que les pidan demostrar los pasos que siguen para resolver problemas o para escribir textos? ¿Por qué?. Algunas destrezas o capacidades, para ser medidas, requieren del uso de más de un tipo de preguntas, en cuyo caso habrá que documentar cuáles tipos, cuántos de cada tipo y justificar el peso que se asignará a cada tipo a la hora de calcular promedios, entre otras cosas.
- **¿Cómo se evidencia que lo que predice la prueba ocurre en realidad?** Cuando el propósito de una prueba es predecir el éxito académico o el éxito en la vida laboral, se debe acumular y reportar evidencias acerca de la relación entre puntajes o promedios obtenidos por los estudiantes en las pruebas con lo que ocurre de hecho durante su carrera académica o laboral.
- **¿En qué medida son compatibles los resultados obtenidos con un instrumento y los obtenidos con otro?** A menudo existen distintos instrumentos que pretenden medir cosas semejantes. Por ejemplo, pueden existir provincias que desean medir el logro de sus estudiantes con el propósito de reportarlo a cada familia. Si existiera simultáneamente una prueba nacional que se usa con el fin de evaluar logros promedio en el ámbito nacional en las mismas áreas, se puede comparar los resultados de los mismos estudiantes en las dos pruebas para acumular evidencia acerca de la con-

vergencia de los resultados. De la misma forma, las pruebas internacionales pueden servir para propósitos técnicos de validación de las mediciones nacionales. Otra estrategia de validación es contrastar los resultados de una prueba con una observación directa a estudiantes o con el análisis de sus tareas y proyectos realizados en clase.

- **¿Cómo se aseguró que las posibilidades que tienen los estudiantes de demostrar lo que saben no está mediada por factores ajenos a su control?** Es importante describir cómo se asegura que todos los estudiantes estén en igualdad de condiciones para demostrar lo que saben. Por ejemplo, tener evidencia de que las preguntas son interpretadas de la misma forma en distintas partes del país o entre distintos grupos lingüísticos, culturales y socioeconómicos. Por otro lado, si se pretende utilizar los resultados de las pruebas para evaluar programas de estudio, opciones pedagógicas o currículum, es importante describir cómo se hará para discriminar entre las ocasiones en que los estudiantes no pueden contestar preguntas que versan sobre cosas que les fueron enseñadas en clase, de aquellas ocasiones en que no pueden contestarlas porque no les fueron enseñadas. Esto es importante, puesto que existen serios problemas éticos cuando se responsabiliza a los estudiantes por contenidos que no han tenido la oportunidad de aprender, o cuando se responsabiliza a los docentes por el logro de sus estudiantes, no habiéndoles proporcionado materiales o capacitación para enseñar esos contenidos.

- **¿Cómo se aseguró una relación óptima entre los contenidos que se pretende enseñar en el grado evaluado y los contenidos evaluados?** Es importante documentar la relación entre el currículum o los estándares y el contenido de las pruebas. ¿Cómo se aseguró congruencia entre ambos? ¿Hubo participación o consulta de las unidades responsables de elaborar el currículum o planes de estudio durante el proceso de construcción de la prueba? ¿Cómo se procedió?

Estas son solamente algunas de las evidencias de validez que los sistemas de medición en América Latina deben considerar en sus estrategias de validación y que en la actualidad raramente se reportan.

Consideraciones finales

Como se estableció anteriormente en la definición formal, la validez es cuestión de *grado*. **No existen mediciones perfectamente válidas o que reproduz-**

can fielmente todas aquellas facetas de la realidad educacional que pretenden medir. Lo que existen son mediciones más o menos válidas, dependiendo de las conclusiones que se pretende tomar a partir de ellas o del uso que se pretende hacer de la información que arrojan.

Es importante recordar que las responsabilidades con respecto a la validación de las mediciones corresponden tanto a los diseñadores de las mismas como a sus usuarios. Quienes diseñan mediciones tienen la responsabilidad de reportar con claridad para qué sirven y para qué no sirven. Deben reportar toda la información pertinente para que los usuarios tengan elementos de juicio para evaluar su validez. Por su parte, los usuarios tienen la responsabilidad de usar los resultados de acuerdo a los criterios de validez que tienen o, si proponen un uso nuevo para las mediciones, les corresponde validarlas para ese nuevo uso.

Debe señalarse también que en América Latina se pretende a menudo que una misma evaluación sirva para más de un propósito. Frecuentemente se espera que una misma prueba permita, por ejemplo, distinguir entre estudiantes que logran o no logran los objetivos académicos de un nivel y que, al mismo tiempo, sirva para juzgar la eficacia de distintas escuelas y la eficacia de diversos programas en las cuales participan dichas escuelas. La validación es específica de acuerdo al uso, es decir, validar un propósito de una prueba no equivale a validarla para otro.

También es cierto que la validez es específica a las poblaciones: una prueba validada para un país o una provincia determinada, no puede ser considerada como validada para otras poblaciones.

Adicionalmente, hay que tener en cuenta que el tiempo cambia las características de los fenómenos y que, por lo tanto, la validación es una tarea continua y una forma de asegurar que nuevos factores que puedan aparecer con el tiempo no atenúen la validez de las mediciones.

Finalmente, dado que su objetivo es asegurar la congruencia de la medición con la realidad educacional que se supone se está midiendo, la validación constituye una actividad *científica*. También se trata de una *actividad técnica de desarrollo*, porque la tarea de acumular evidencia de validez a menudo trae como consecuencia el rediseño o el afinamiento de los instrumentos o de sus sustentos teóricos.

Es necesario reconocer que en América Latina puede no ser posible diseñar evaluaciones específicas para cada propósito para el cual se necesita contar con

información para tomar decisiones. Esto genera un dilema importante que deben confrontar los países. Pongamos un ejemplo: si no existiera actualmente una prueba que se haya validado específicamente para ser usada para distinguir entre la eficacia de centros educativos que utilizan un programa de estudios y la de centros que utilizan otro, y es necesario decidir cuál de los programas debe ser difundido y promovido por el Ministerio, ¿significa acaso que no debemos utilizar las pruebas existentes para ese propósito? No hay respuesta simple. Para decidir sobre este asunto, será necesario determinar en qué medida es mejor la decisión que tomaríamos utilizando los resultados de la prueba, en comparación con la decisión que tomaríamos sin usarla. Si el posible mayor valor de una decisión tomada sobre la base de la prueba se juzga suficiente, sería sin duda un insumo que se debe usar. Pero es necesario tener presente que esto no significa que la hemos validado para este propósito. El valor de los resultados de las pruebas como insumos para la toma de decisiones sólo puede optimizarse cuando se asume la responsabilidad de validarlos para ese propósito. Tomar una decisión basada en una inferencia inválida equivale a tomar una decisión sin fundamento.²

Notas:

(1) Durante mucho tiempo, la concepción de validez más vigente y extendida, y que dominó el mundo académico y de las prácticas de medición evaluativa tanto en América Latina como en gran parte del mundo, fue la propuesta en 1949 por L.J.Cronbach en su libro *Essentials of Psychological Testing* (New York: Harper and Row), cuya versión quizás más conocida fue la ofrecida por A.Anastasi en su *Psychological Testing*, publicado en 1954. Desde entonces, la evolución de la teoría y métodos de las mediciones psicológicas y educacionales ha llevado a una nueva conceptualización y a la estandarización de la misma entre los profesionales de esas disciplinas. Así, en la tercera edición del texto *Educational Measurement* de R.L. Linn, publicada también por Macmillan en 1989, apareció la propuesta de Samuel Messick. Revisiones de esa propuesta llevaron a la acepción de validez actualmente establecida y que está documentada en los *Standards for Psychological and Educational Measurement*, publicados conjuntamente por la Asociación Americana de Investigación Educativa, la Asociación Psicológica Americana y el Consejo Nacional de Medición Educativa de los Estados Unidos en 1999. Es esta concepción, que se refiere a las acciones, decisiones e inferencias que se toman a partir de las

mediciones – es decir, a cómo se usan – la que se ha utilizado en este capítulo.

Fuentes bibliográficas adicionales recomendables para este tema son Campbell, L. J. y Fiske, D. W. (1959) "Convergent and Discriminant Validity in the Multitrait-Multimethod Matrix" en *Psychological Bulletin*, 56: 81-105; Cronbach, L.J. (1989) "Construct Validation after Thirty Years" en R. L. Linn (Ed.) *Intelligence: Measurement Theory and Public Policy*. Urbana: University of Illinois Press; Messick, S. (1989b) "Meaning and Values in Test Validation: The science and ethics of Assessment" en *Educational Researcher*, 18 (2), 5–11; Messick, S. (1994) "The Interplay of Evidence and Consequences in the Validation of Performance Assessments" en *Educational Researcher*, 23 (2), 13-23; Moss, P.A. (1995) "Themes and Variations in Validity Theory" en *Educational Measurement: Issues and Practice*, 14 (2). 5-12.

(2) En este capítulo se ha abordado solamente el tema de la validez. Otra cuestión técnica asociada a la validez es el tema de la consistencia de las mediciones, denominada confiabilidad. Esta se refiere a tres cosas interrelacionadas: (1) A la noción de la estabilidad de la medición. En este sentido, nos preguntamos si las pruebas o encuestas arrojan resultados similares siempre que se aplican a sujetos similares en condiciones similares. (2) A su nivel de precisión, esto es, la relación de los resultados de la medición con la "realidad" que mide. (3) A la confiabilidad, es decir, a la cantidad de error (llamada varianza sistemática) que contiene la medición. Si una prueba no mide con confiabilidad lo que se propone medir, no es válida. Es importante señalar, sin embargo, que la confiabilidad es una condición necesaria, mas no suficiente, para la validez. Sin confiabilidad no hay validez, pero la confiabilidad no es garantía de validez.

Capítulo III

EL DISEÑO DE LAS PRUEBAS PARA MEDIR LOGRO ACADÉMICO: ¿REFERENCIA A NORMAS O A CRITERIOS?

Juan Manuel Esquivel

¿Qué opciones existen para medir, a escala nacional, los conocimientos que los estudiantes adquieren en las escuelas?

¿Se hacen las mediciones en Latinoamérica con el propósito de comparar los logros de grupos de estudiantes con otros grupos de estudiantes?

¿Se realizan, en cambio, para medir si éstos han logrado los aprendizajes que el sistema educativo pretende que ellos logren?

¿Cuáles son las diferencias y similitudes conceptuales y metodológicas entre esas dos formas de realizar la medición?

¿Cuál es el papel de la evaluación del desempeño y evaluación auténtica en los sistemas de medición de la región?

Introducción

A los funcionarios de ministerios de educación y entidades encargadas de los sistemas de medición de logro de los países de América Latina se les presenta la siguiente disyuntiva al desarrollar las evaluaciones educacionales:

- elaborar pruebas que permitan comparar el logro de grupos de estudiantes con otros grupos; o
- elaborar pruebas que permitan descubrir qué aspectos, conocimientos u objetivos específicos logran los estudiantes.

Responder a este dilema implica desarrollar pruebas en base a paradigmas con fundamentaciones teóricas diferentes y, en ciertos aspectos, contradictorios. En el primer caso, se puede trabajar dentro del paradigma de medición referida a *normas*, mientras que en el segundo se debe recurrir al paradigma de medición referida a *criterios*.

En la mayoría de los países de la Región, se ha recurrido a las pruebas referidas a normas, en las que se privilegia la función de ordenamiento o “discrimina-

ción” entre grupos o individuos. Este enfoque está fuertemente marcado por su función principal, que históricamente ha sido la de seleccionar individuos para el ingreso al ejército o a las universidades. En esos casos no importaba tanto si el individuo dominaba o no ciertos campos del conocimiento, sino distinguir a los individuos más aptos de los menos aptos. El enfoque de pruebas referidas a criterios, en cambio, se propone comprobar principalmente si los individuos dominan un cierto campo de contenidos y/o destrezas.

Respecto a la disyuntiva planteada, en este capítulo se analiza un ejemplo de lo que típicamente se encuentra en la Región en cuanto a desarrollo y validación de pruebas de conocimiento dentro del paradigma referido a normas. Luego, se revisan algunas de las diferencias entre los paradigmas señalados y se da un ejemplo de desarrollo de una prueba de acuerdo con el paradigma referido a criterios. Finalmente, se hace una referencia al impulso que, en algunos países desarrollados, se ha comenzado a dar al empleo de las denominadas *pruebas de desempeño* y a la *evaluación auténtica*.

Las pruebas referidas a normas

Como se mencionó, **la mayoría de los países latinoamericanos ha desarrollado pruebas para sus sistemas de medición del logro dentro del modelo psicométrico de las pruebas referidas a normas.** La información producida por estas pruebas generalmente se ofrece en forma de promedios basados en el número de preguntas correctas obtenidas por los estudiantes o como una escala derivada de esta información básica, por ejemplo, el porcentaje de respuestas correctas o la nota en términos de la escala de calificación empleada en cada país. Estos promedios, aunque tienen una utilidad innegable para realizar comparaciones entre los diferentes niveles de desagregación de las variables de interés en las muestras (por ejemplo: urbano-rural, público-privado), tienen escaso sentido pedagógico, pues no entregan información real sobre el logro de conocimientos, destrezas o habilidades específicas de parte de los estudiantes. ¿Qué información de utilidad le comunican los promedios a un maestro de aula que le permita mejorar su trabajo con los niños? ¿Qué utilidad tienen para un curriculista en el Ministerio de Educación?

Entre las razones por las cuales se ha recurrido al enfoque referido a normas están:

- La abundancia de experiencia e información internacionalmente disponible sobre los procedimientos que tradicionalmente se han seguido en

Juan Manuel Esquivel. Profesor de Educación Media en Química. B.Sc. en Química M. Sc. Medición e Investigación Educativa. Ph.D. en Currículo. Catedrático de la Facultad de Educación de la Universidad de Costa Rica. Investigador retirado del IMEC-UCR. Consultor en Evaluación y Medición. Actualmente Director de dos proyectos educativos en la Coordinación Educativa y Cultural Centroamericana.
E-mail: sgcecco@racsa.co.cr

la elaboración y validación de pruebas de este tipo.

cipios de la teoría de pruebas referidas a normas.

- La limitada formación y capacitación académica en el área de la medición a la cual han podido acceder los funcionarios ministeriales encargados de desarrollar las pruebas, se ha dado sobre los prin-
- La disponibilidad de paquetes estadísticos de computo que permiten realizar análisis de ítemes tradicionales o novedosos y otros análisis técnicos.

Recuadro 2

La medición del logro en América Latina: Un ejemplo típico de prueba bajo el paradigma referido a normas.

Etapa 1.

Durante la preparación de un proyecto de préstamo o de donación con un organismo internacional, se detectó la necesidad de tener un sistema de medición del logro académico de los estudiantes. Una vez financiado el proyecto, se inició la preparación de pruebas para tercero y sexto grados de primaria en matemáticas y lenguaje, utilizando como base el curriculum prescrito. Como primer paso del proceso, se definió una tabla de especificaciones, en la cual los contenidos del curriculum se dividieron por áreas y éstas, a su vez, se subdividieron en contenidos más específicos. La tabla se balanceó de acuerdo con la complejidad cognitiva (nivel taxonómico) que se pretende tengan los ítemes y con el número de ítemes con los que se quería medir esos conocimientos.

Etapa 2.

Se procede a la elaboración de los ítemes para llenar las expectativas establecidas en la tabla de especificaciones. Para ello, se ha acostumbrado realizar grandes operativos, como talleres con maestros en diferentes lugares del país, a quienes se les pide redactar ítemes de preguntas de selección múltiples para diferentes contenidos y en distintos niveles taxonómicos. Así se recogen enormes cantidades de ítemes, lo que se justifica bajo el argumento de que le da validez curricular a la prueba. En otros lugares, la elaboración de ítemes está a cargo de un número reducido de personas que tiene una vasta experiencia y/o capacitación en estas tareas.

Etapa 3.

Se someten los ítemes a una revisión, tarea cuya complejidad y sistematización varía desde una revisión relativamente informal para detectar defectos gruesos en la estructura hasta revisiones con hojas de calificación preparadas con ese propósito expreso y, en unos pocos casos, revisiones de relación ítem-contenido llevadas a cabo por jueces independientes.

Etapa 4.

Se prepara una prueba piloto de ítemes. Se selecciona una muestra de escuelas o aulas de acuerdo con un plan de muestreo previamente establecido. En la mayoría de los casos el muestreo es intencional, procurando que incluya escuelas representativas de los diversos estratos de interés del sistema de medición y que cada ítem sea respondido por entre 150 y 300 estudiantes. Los ítemes disponibles se agrupan en diversos formularios que se aplican simultáneamente. El propósito fundamental de la aplicación piloto es poder hacer análisis estadísticos para conocer índices que caracterizarán a los ítemes. Generalmente estos análisis se ejecutan con paquetes estadísticos comerciales. Con los parámetros obtenidos (dificultad, correlación ítem-puntaje total de la prueba —discriminación—, frecuencia de respuesta según opción, etc.), se realiza una selección de los ítemes que constituirán las pruebas definitivas. Esto generalmente se rige por los principios de la teoría de pruebas referidas a normas, que establecen como preguntas ideales aquéllas que tienen una dificultad cercana al 50% y una discriminación por encima de 0.40 o correlaciones ítem-puntaje total positivas y significativamente mayores que cero.

Etapa 5.

Una vez aplicada la prueba, se reportan resultados en términos de los promedios obtenidos en cada una de las áreas en que se dividió la tabla de especificaciones. Estos se reportan como puntajes de logro y se interpretan en términos de dominio de cada una de esas áreas, cuando se supera cierto puntaje de corte. Además, se reportan resultados de promedios totales de las pruebas, que resultan muy convenientes para establecer comparaciones entre los diversos niveles de desagregación de las variables de la muestra (grupos de alumnos o escuelas).

En el Recuadro 2, se presenta un caso que se podría encontrar típicamente en los sistemas de medición del logro en la Región, sobre el cual se harán los siguientes comentarios y observaciones puntuales referidas a las diferentes etapas de su desarrollo.

Etapa 1.

Preparación de las especificaciones de los contenidos a medir.

Cabe destacar tres hechos relevantes respecto a este tema:

- La decisión que generalmente se hace de tomar el currículo prescrito como base de las pruebas. Al hacerlo, se asume que éste es conocido y comprendido por todos los maestros, que ellos han recibido la capacitación adecuada para ejecutarlo, que tienen acceso a los mismos materiales de enseñanza y, por lo tanto, los niños han tenido alguna oportunidad para aprenderlo. Dado que asumir estas condiciones no es siempre justificable, al menos se debería planear la ejecución de un estudio paralelo a las pruebas para conocer en qué medida esas condiciones realmente están dadas en cada caso.
- La decisión, que normalmente se toma, de basar el diseño de la prueba en una tabla de especificaciones dividida por áreas y éstas, a su vez, en contenidos más específicos y de emplear alguna taxonomía para catalogar la complejidad cognitiva con que se quiere medir los contenidos. La complejidad cognitiva de los contenidos, en general, incluye desde los conocimientos memorísticos hasta el empleo del razonamiento lógico. Se presentan diferentes maneras de denominar estos niveles de complejidad cognitiva de acuerdo con la taxonomía que se emplee, aunque la más popular es la taxonomía de Bloom y sus colaboradores. Los niveles taxonómicos señalan la complejidad cognitiva que se pretende que tengan los ítemes con que se medirán los contenidos determinados en la tabla de especificaciones. Debe tenerse en mente que la tabla de especificaciones es un instrumento que se emplea con el propósito de tener alguna seguridad de que la prueba será una muestra representativa de los contenidos considerados para ser medidos con la prueba y los niveles taxonómicos con que se quiere medir esos contenidos. Dentro del paradigma de pruebas referidas a normas, la evidencia de que la prueba es una muestra representativa de la totalidad de los contenidos considerados para ser medidos en una prueba es una información fundamental para establecer la validez de la interpretación de los re-

sultados. La intención de dividir el contenido por áreas, en la tabla de especificaciones, se hace evidente cuando se leen los informes de resultados, pues en ellos se reportan los promedios de estas áreas como puntajes de logro. Esto significa que se hace una interpretación propia de la medición referida a criterios (ver más adelante) con una prueba que se ha diseñado como referida a normas.

- La enorme debilidad que tiene cualquier sistema taxonómico como medio de balancear una tabla de especificaciones. Esta debilidad deriva del hecho que el nivel taxonómico que se le atribuye a un ítem determinado está definido por la experiencia de enseñanza que tiene la persona que lo juzga. Así, un mismo ítem recibirá una variada gama de niveles taxonómicos cuando se somete a juicio de varias personas.

Etapa 2.

Elaboración de ítemes

- La escritura de preguntas de selección u opción múltiple de buena calidad es un trabajo especializado que requiere para su ejecución de personal con experiencia y capacitación en esta labor. El típico constructor de ítemes debe tener dos características: (i) un excelente dominio de la materia sobre la que escribirá ítemes, y (ii) una amplia y probada experiencia en labores de escritura y revisión de ítemes. Las preguntas de opción múltiple, si están adecuadamente elaboradas, permiten medir habilidades complejas y, por lo tanto, la crítica de que con ellas solamente se mide memoria y habilidades simples es bastante infundada, o es válida únicamente para pruebas mal diseñadas.
- Se debe reconocer que los operativos para que los maestros escriban ítemes de las pruebas, como ocurre con frecuencia, pueden tener beneficios políticos para la aceptación del sistema de medición y para capacitar maestros. Sin embargo, hay que tener claro que estos operativos no necesariamente contribuyen a la calidad de la prueba. El argumento de que así se le da validez curricular a la prueba, no tiene sentido, dado que a los maestros se les pide redactar ítemes sobre contenidos y niveles taxonómicos que han sido decididos en el nivel central. Estos contenidos en esos niveles taxonómicos pueden nunca haber sido enseñados por los maestros que han construido ítemes para medirlos. Por otra parte, se ha comprobado que la inmensa mayoría de los ítemes escritos por maestros en estos operativos, son desechados en la primera revisión.

Etapa 3.

Revisión de ítemes que se incluirán en las pruebas

- En general, en la Región el desarrollo y validación de las pruebas empleadas han sido limitados, pues no se ha puesto la debida atención a los procesos de comprobación de la calidad de la estructura de los ítemes y a establecer claramente la relación de cada ítem con el contenido que pretende medir. Esto es particularmente serio para efectos de establecer la evidencia de validez necesaria para la interpretación y uso del resultado de la medición. La revisión de los ítemes debería dividirse en dos aspectos: (i) una revisión estructural sistemática, hecha por jueces especialistas que dominen la disciplina para la que se escribieron los ítemes y tengan amplia experiencia en la escritura de ítemes de selección múltiple; y (ii) una comprobación de la relación entre el ítem y el contenido que se supone que éste mide, también a través de jueces que tengan una experiencia reciente en la enseñanza en el nivel en el que se aplicará la prueba y dominio de los conocimientos de la disciplina. Ellos trabajarán independientemente, juzgando si cada ítem mide o no el contenido que se supone que mide. Una mayoría calificada de los jueces (alrededor de un 75%) tendrá que mostrar acuerdo en la relación de cada ítem con un contenido.

Etapa 4.

Aplicación de prueba piloto

En esta etapa se observan dos debilidades básicas:

- Casi nunca se establece como objetivo de la aplicación de la prueba piloto la obtención de retroalimentación sobre la prueba por parte de los estudiantes y de los docentes. La recolección de información cualitativa acerca del contenido que cubren las preguntas de la prueba y sobre la claridad y comprensión de los ítemes sería de vital importancia para el esfuerzo por acumular evidencia de validez. Esto podría lograrse mediante una discusión con los estudiantes y una conversación con el maestro.
- La selección de los ítemes que formarán las pruebas definitivas sobre la base de las preguntas que muestran una dificultad cercana al 50% y una discriminación de 0.4, como se suele hacer, tiene algunas consecuencias, pues debido a ello comúnmente los resultados reportados como promedios de los puntajes totales de las pruebas están alrededor del 50%. En otras palabras, sería absolutamente imposible obtener resultados que no estuvieran en el rango de alrededor de la mitad del

puntaje posible. Por otra parte, este procedimiento implica descartar los ítemes que resultan muy difíciles o muy fáciles, aunque los mismos sean buenos desde el punto de vista pedagógico y midan competencias relevantes, lo que implica perder la posibilidad de recoger información valiosa sobre las capacidades y conocimientos de los estudiantes.

Etapa 5.

Reporte de los resultados una vez aplicadas las pruebas

- Dados los procedimientos seguidos en su desarrollo, las interpretaciones de logro que suelen darse a los resultados por área carecen de sustento teórico y empírico. Esto significa que se está realizando una interpretación referida a criterios para una prueba referida a normas.

Resumiendo algunos aspectos que se derivan de lo analizado, cabe destacar lo siguiente:

1. Las pruebas referidas a normas tienen un espacio en los sistemas de medición del logro si su desarrollo es congruente con el objetivo que se pretende alcanzar al aplicar las pruebas, como sería la comparación del rendimiento general de los estudiantes de acuerdo a variables tales como sexo, rural- urbano, sostenimiento de las escuelas, regiones geográficas, etc.
2. Existen usos apropiados para las pruebas referidas a normas y para las pruebas referidas a criterios, dependiendo del grado de la "granularidad" (ver capítulo I de este documento) de lo que se mide y a quién se mide.
3. De acuerdo con el análisis aquí realizado, existiría aún mucho espacio para mejorar el desarrollo y validación de las pruebas referidas a normas que actualmente se emplean en la Región.

Las pruebas referidas a criterios

Unos pocos sistemas de medición del logro en la Región han desarrollado pruebas referidas a criterios y las han sometido a un proceso de validación. Esta alternativa de medición tiene la gran ventaja de que permite obtener información con mucho significado pedagógico, pues evalúa los conocimientos, destrezas y habilidades específicas que un grupo de estudiantes logra dominar.

Con el ejemplo del Recuadro 3, se ilustra el tipo de procedimientos que emplea este enfoque.

Recuadro 3

Ejemplo de una prueba referida a criterios desarrollada en América Latina

El Ministerio de Educación define como objetivo de las pruebas de final de la Educación Primaria brindar información específica sobre el logro de los objetivos fundamentales del currículo por parte de los estudiantes. Para ello, el Departamento de Pruebas Nacionales (DPN) del Ministerio decide elaborar y someter a un proceso de validación pruebas referidas a criterios en las cuatro asignaturas básicas.

La primera pregunta que plantea el DPN a las autoridades políticas es: ¿Se quiere medir el currículo prescrito o el currículo enseñado? La respuesta de dichas autoridades es el currículo prescrito o sea el “deber ser” curricular. Las siguientes preguntas clave que se hace el DPN son: ¿cuáles son los objetivos fundamentales del currículo? y ¿quiénes son las personas más indicadas para realizar la selección de los objetivos fundamentales? Para responder la primera pregunta se realizó una lectura e interpretación de los programas de estudio (fundamentados en el humanismo-constructivista) de cada una de las asignaturas. Esto dio origen a listados de entre 50 y 70 objetivos de aprendizaje por asignatura. Esta interpretación hecha por pares de especialistas en cada asignatura, miembros del equipo de la DPN, se sometió al juicio de grupos de 10 especialistas en cada asignatura del Departamento de Currículo del Ministerio. A estos jueces se les solicitó, en primer lugar, manifestar su acuerdo o desacuerdo con la interpretación hecha de los programas de estudio y participar en una discusión para, mediante el consenso, llegar a una interpretación única de los programas de estudio. En segundo lugar, se les solicitó que trabajando independientemente señalaran los 30 objetivos que consideran más importantes de lograr por un estudiante que termina la Educación Primaria. Luego, se les solicitó realizar la priorización de 20 objetivos de los 30 seleccionados. Mediante un procedimiento estadístico, se definieron los 20 objetivos que tuvieron las más altas prioridades promedio, los cuales constituyeron los objetivos fundamentales de cada una de las asignaturas, y para cada uno de estos objetivos se desarrollaron especificaciones de contenido.

A continuación, el DPN procedió a seleccionar la técnica mediante la cual se definirían las especificaciones de contenido. En este caso se escogió la técnica de los objetivos amplificados. Los pares de técnicos por asignatura fueron capacitados en el empleo de la técnica en un taller de una semana. Durante ese período y las dos semanas siguientes, estos técnicos desarrollaron las veinte especificaciones de su asignatura. El paso siguiente consistió en la validación de los objetivos amplificados escritos. Para ello, se solicitó a cinco especialistas en la asignatura que escribieran un ítem para cada objetivo amplificado, de acuerdo con las condiciones establecidas en ellos. Si los ítems producidos por los cinco especialistas para cada uno de los objetivos amplificados resultaban muy similares, se comprobaría que los objetivos amplificados cumplen adecuadamente la doble función de limitar el contenido y establecer y comunicar las reglas de estructuración de los ítems.

La fase siguiente fue la escritura de los ítems. Se decidió que se escribirían 12 ítems para cada objetivo amplificado. La escritura de los ítems estuvo a cargo de los pares de especialistas por asignatura y de dos personas más seleccionadas entre el grupo de escritores de ítems, ya capacitados, que tiene el DPN. Estas dos personas adicionales por asignatura fueron profesores de educación primaria que en ocasiones anteriores habían probado su habilidad para escribir ítems de selección múltiple.

A continuación se llevaron a cabo dos procesos de vital importancia:

1. La revisión estructural de los ítems. Para ello se contrataron dos personas en cada asignatura. Estos profesionales tenían la doble característica de poseer una gran experiencia en la preparación y revisión de ítems de selección múltiple, y un reconocido dominio de la materia. Para que sirviera de base al trabajo de estos pares de jueces, se preparó una hoja de cotejo que resumía las principales características de la estructura de los ítems que ellos debieron examinar. Como producto de esta revisión se modificaron algunos ítems y se desecharon unos pocos.

2. El establecimiento del índice de la congruencia de cada uno de los ítemes con su objetivo amplificado. Para esto se contrataron 10 jueces por asignatura, quienes tenían un probado dominio de la materia y alguna experiencia de enseñanza en la Educación Primaria. Se prepararon formularios en los cuales los jueces vertieron su juicio independiente y se acondicionó un local para facilitar su trabajo y el control que sobre ese trabajo tuvieron que realizar los especialistas del equipo del DPN. Mediante la lectura óptica de los formularios empleados por los jueces, se capturó la información producida en esta fase de juicio. Con un programa de cómputo apropiado se calculó el índice de congruencia para cada ítem. Los ítemes con índices de congruencia menores a 0,75 fueron desechados.

En cada asignatura, con los ítemes que presentaban índices de congruencia iguales o mayores a 0,75 se constituyeron folletos o cuadernillos de prueba con 40 ítemes cada uno; cuatro ítemes por cada objetivo amplificado. Se constituyeron seis cuadernillos diferentes, puesto que se tenían 20 objetivos amplificados y hasta 12 ítemes por objetivo. En los casos en que se hubiesen desechado ítemes por los procesos antes descritos, y no se tuvieran los 12 ítemes por objetivo, se repetían ítemes en algunos de los diferentes cuadernillos. Se seleccionó una muestra no aleatoria de escuelas de diferentes características geográficas, sociales, de tamaño y de financiamiento (públicas-privadas-subsuencionadas). La meta era que un mínimo de 200 estudiantes respondieran cada formulario de cada asignatura. Además, se planificó la recolección de información cualitativa sobre los ítemes, mediante las discusiones que se tuvieron con los estudiantes acerca de la claridad y comprensión de los ítemes. A los maestros de los grupos de estu-

diantes a los que se les aplicaron los cuadernillos de prueba, también se les solicitó y se registró su opinión acerca de los ítemes. Esta información cualitativa sirvió para detectar deficiencias de lenguaje en los ítemes y se utilizó para modificarlos. Con las respuestas a los ítemes y empleando los programas de cómputo apropiados, se calcularon la dificultad y la discriminación de Brenann_ para cada ítem.

Para cada objetivo amplificado se hizo un banco de ítemes con aquellos ítemes que presentaban los índices de congruencia de más alto valor y que, además, tuvieran un valor de discriminación cercano a cero y una dificultad mayor al 50% (el índice de dificultad señala el porcentaje de alumnos que contesta correctamente, lo que significa que si es mayor al 50% es un ítem relativamente fácil). De este banco de ítemes se seleccionaron aleatoriamente cuatro ítemes para medir cada objetivo amplificado. De acuerdo con la definición de prueba en la medición referida a criterios, estos cuatro ítemes constituían una prueba. En cada asignatura, se constituyeron dos folletos o cuadernillos, y en cada uno de ellos se reunieron 40 ítemes pertenecientes a 10 objetivos amplificados. De esta manera se cubrieron los 20 objetivos amplificados para los que se desarrollaron ítemes. Los folletos o cuadernillos se aplicaron a muestras de estudiantes seleccionados aleatoriamente o estratificadas por región educativa, tamaño de escuela y zona geográfica.

Una vez analizada la información, se escribieron varios tipos de informes. A las escuelas y autoridades técnicas regionales y centrales se les hizo llegar uno en que se ofrecía el porcentaje de estudiantes que había dominado cada uno de los objetivos fundamentales, de acuerdo con los diferentes niveles de las variables en que se estratificó la muestra.

Diferencias entre pruebas referidas a normas y pruebas referidas a criterios

Es común escuchar y leer la errónea aseveración de que lo único que distingue a las pruebas referidas a criterios de las de normas es la interpretación de los resultados, interpretación que puede ser relativa (o sea con respecto a la media aritmética y la variabilidad) o referida al logro. Esta confusión se deriva de

una concepción errada de las características técnicas de las pruebas referidas a criterios y de un desconocimiento de la teoría que sustenta este paradigma.

Existen, de hecho, importantes diferencias conceptuales y metodológicas entre ambos tipos de pruebas, como se resumen en el Recuadro 4.

Recuadro 4

Diferencias conceptuales y metodológicas de las pruebas referidas a normas y las referidas a criterios

Pruebas referidas a criterios	Pruebas referidas a normas	Pruebas referidas a criterios
DIFERENCIAS CONCEPTUALES		
1. Paradigma de base	Se basa en el paradigma psicométrico, cuya premisa es que los resultados de la medición de cualquier característica humana en una población se comportarán de acuerdo con la curva normal. Como consecuencia, privilegia maximizar la variabilidad y así asegurarse que los resultados de la aplicación de las pruebas se comportan normalmente.	Se basa en el paradigma educométrico, cuyo principio es que la educación persigue que todos los niños aprendan; por consiguiente, se espera una distribución de resultados sesgada hacia los valores más altos de la escala de puntajes. La variabilidad no es una característica que importe, por lo cual no preocupa su valor.
2. Tipos de información que busca	Privilegia la comparación entre estudiantes o entre grupos de estudiantes	Privilegia la comparación de los logros de los estudiantes con respecto a las metas de aprendizaje o a las competencias que el sistema educativo persigue que éstos alcancen.
DIFERENCIAS METODOLOGIAS		
1. La definición de lo que se va medir	Suele ser una definición general y vaga. Generalmente consiste en listados de conocimientos a manera de temarios o en listados de objetivos más o menos definidos.	La definición tiene que ser clara y específica, detallando el "dominio del conocimiento" que abarca el contenido por medir y las reglas básicas de estructuración de los ítemes con que se medirá ese dominio. A estas definiciones se las conoce como "especificaciones de contenido".
2. La definición de prueba	La prueba es el conjunto de ítemes que forman una muestra representativa de todos los conocimientos, destrezas y habilidades que se quiere medir. El criterio de representatividad puede ser muy variado.	La prueba es el conjunto de "n" ítemes, aleatoriamente seleccionados de una población infinita de ítemes, que se emplea para medir únicamente una especificación de contenido. En otras palabras, se mide un solo conocimiento, habilidad o destreza.
3. La evidencia de validez	Para esta evidencia será muy importante: (a) la certeza de que los ítemes tienen las características adecuadas en su construcción; (b) el proceso de juicio por el cual se recoge información acerca de la relación del ítem con el contenido que pretende medir; y (c) la evidencia que se necesita obtener sobre el ajuste entre los ítemes seleccionados para constituir la prueba definitiva y las especificaciones contenidas en la tabla correspondiente.	Aquí es importante: (a) la certeza de que los ítemes están contruidos de acuerdo con las características propias de las preguntas de selección múltiple; y (b) un valor aceptable en el índice de congruencia entre cada ítem y su especificación de contenido.

4. Los procedimientos de cálculo de la confiabilidad y de análisis de ítems.

- a) Los modelos empleados para estos efectos se fundamentan en la maximización de la variabilidad de dos variables que están correlacionadas, lo que permite asegurarse de que la magnitud de la correlación entre esas dos variables será más alta. La correlación es la técnica fundamental empleada en el cálculo de la confiabilidad.
 - b) Los parámetros resultantes del análisis de ítems, principalmente la dificultad y la discriminación o el índice de correlación ítem-puntaje total de la prueba, se usan como indicadores fundamentales para la selección de los ítems posibles de incluir en la prueba definitiva.
- a) Se da más importancia a la consistencia entre los resultados obtenidos por un grupo de alumnos a los cuales se les aplica la misma prueba o una prueba paralela en dos oportunidades distintas. El índice de confiabilidad resulta de la proporción de estudiantes cuyas respuestas en ambas oportunidades demuestren que sí (o no) dominan una competencia o especificación de contenidos.
 - b) La selección de ítems que constituirán la prueba definitiva se realiza considerando en primer lugar el índice de congruencia entre el ítem y la especificación de contenido y, luego, considerando la discriminación y la dificultad.

5. La interpretación de los resultados de la medición.

- La interpretación es relativa. El puntaje tiene significado al ser comparado con la media aritmética y la desviación estándar (o con las normas, si la prueba ha sido estandarizada o normalizada)
- La interpretación es absoluta. El resultado se interpreta en términos del logro o no logro de la especificación del contenido medido o sea, en términos del dominio del conocimiento, habilidad o destreza medida.

6. Otras diferencias.

- a) No necesita establecer un puntaje de corte, entendido esto como el puntaje que define si un estudiante domina o no la especificación de contenido, como ocurre en las pruebas referida a criterios.
 - b) El número de preguntas que constituyen una prueba estará determinado, entre otros factores, por el objetivo de la prueba, por el nivel de escolaridad de los niños que tomarán la prueba, la asignatura que se mide, la tabla de especificaciones y el tiempo del que se dispone para su aplicación.
- a) Necesita establecer un puntaje de corte.
 - b) El número de preguntas dependerá fundamentalmente del tipo de decisión que se va a tomar con la prueba (formativa o sumativa) y con respecto a quién se va a tomar esa decisión (un individuo o una muestra de individuos). En general se determina que cuando las decisiones son formativas y para muestras de individuos, el número de ítems varía entre 3 y 5, mientras que decisiones sumativas e individuales requieren entre 8 y 10 ítems.

Nuevas perspectivas en la medición del logro: la evaluación del desempeño y la evaluación auténtica.

Una nueva alternativa en materia de evaluación, que ha cobrado fuerza durante la última década en los países desarrollados, es la llamada “evaluación del desempeño”, un enfoque de medición según el cual los estudiantes deben producir sus respuestas o ejecutar tareas, en lugar de simplemente seleccionar la respuesta correcta entre varias alternativas.

El desempeño de los estudiantes se juzga con criterios pre-establecidos, basados en el discernimiento humano. Los medios que emplea la evaluación del desempeño han sido en el pasado usados por los educadores en las aulas, y requieren un tiempo sustancial de parte del estudiante:

- Preguntas de respuesta o final abierto.
- Producción de un ensayo.
- Resolución de problemas.
- Producción de materiales o discursos para exhibición pública.
- Elaboración de artefactos y documentos.
- Producción de un portafolios o muestras de trabajos realizados a lo largo de un período.

Se dice que la evaluación del desempeño es *auténtica* cuando las tareas que el estudiante ejecuta tienen como contexto situaciones propias del mundo real o recrean un contexto del mundo real. **Lo novedoso de este enfoque radica en el énfasis en la medición de conocimientos y habilidades complejas y de alto nivel de pensamiento, preferiblemente en un contexto de mundo real en el que se emplean esos conocimientos y habilidades.**

Es necesario, sin embargo, tomar en cuenta algunas características de este enfoque cuando se aplica a muestras masivas de estudiantes, y que pueden ser consideradas como una limitación:

- **El tiempo que requiere un estudiante para completar una tarea específica** (algunas veces todo un curso, como es el caso de los portafolios)
- **La limitación técnica de depender del juicio humano para juzgar la calidad de la ejecución de la tarea.** Este factor se ha señalado como una debilidad fundamental del enfoque, pues es muy difícil cumplir con principios básicos que permitan considerar confiable la calificación. Esta realidad, a su vez, influye en la calidad de la evidencia empírica que permite interpretar los resultados de estas mediciones de la forma que se pretenden interpretar.

- **Bajo poder de generalización.** En la mayoría de los casos que se observan en la práctica de la evaluación del desempeño, se mide un conocimiento o habilidad compleja con una sola tarea, dado lo extenso y complejo de la misma y el tiempo que consume su ejecución. Esto hace evidente una limitación de esta opción, dado el poco poder de generalización del resultado. La decisión del logro que se hace sobre la base de una respuesta a una única tarea tiene poca validez.
- **Costo.** Esto es una seria limitante para el uso de este tipo de mediciones en la Región, pues no sólo se requerirá emplear más tiempo en el proceso de medición, con el costo que esto implica, sino que se requerirá más materiales para realizarla. El costo sube aún más al agregar el pago del personal que califica las tareas, su capacitación y el necesario control para que se mantengan los niveles mínimos de la confiabilidad de la calificación.

Sin embargo, resulta indudable que es necesario medir el conocimiento y las habilidades complejas, y la idea de aplicar algunos de los medios de la evaluación del desempeño en las pruebas de los sistemas de medición del logro la Región resulta atractiva. Una posibilidad es hacerlo en pequeñas submuestras de estudiantes para, de esta manera, mantener los costos bajos y poder tener información sobre el logro de conocimientos y habilidades complejas, las cuales no se pueden medir con las pruebas tradicionales con preguntas de selección múltiple.

Conclusiones

Con lo expuesto en este capítulo se ha pretendido dar respuesta a las preguntas con las que se abrió el mismo. Se puede resumir lo escrito de la siguiente manera.

1. En la región latinoamericana todos los países tienen alguna forma de sistema de medición de logro. La mayoría de las pruebas que se emplean se desarrollan bajo los principios de la medición referida a normas. En muchos casos se ha descuidado la recolección de evidencia empírica que sustente la interpretación válida de los resultados. Con la medición basada en normas no es posible tener información específica y válida sobre el logro de conocimientos, habilidades y destrezas. Sólo es apropiada cuando el objetivo de la medición es realizar comparaciones acerca del rendimiento académico general entre diversos estratos de la población sometida a medición.
2. Unos pocos países han ensayado la elaboración de pruebas referidas a criterios. En estos casos,

se ha tenido una mayor preocupación por sustentar la validez de la interpretación de los resultados. La medición basada en criterios permite llegar a conclusiones sobre el logro específico de ciertos conocimientos, habilidades y destrezas, lo cual entrega información valiosa para evaluar el cumplimiento de los objetivos curriculares.

3. Existen diferencias conceptuales profundas entre la medición referida a criterios y la medición referida a normas. Estas diferencias conceptuales, a su vez, dan lugar a diferencias en los procedimientos metodológicos del desarrollo y validación de las pruebas.
4. La *evaluación del desempeño* y la *evaluación auténtica* ofrecen medios para medir el logro de conocimientos y habilidades complejas. En muchos países de Latinoamérica, se están llevando a cabo reformas curriculares profundas que dan mayor importancia al logro de habilidades complejas. Por ello, resultaría conveniente que los sistemas de medición incluyan alguno de los medios de la evaluación del desempeño en los instrumentos de medición que empleen. Esto daría validez curricular a los resultados de la medición y permitiría un mejor y mayor alineamiento entre la medición del logro y las reformas curriculares. Sin embargo, se debe tener en cuenta que, por sus características, la evaluación del desempeño presenta limitaciones de otra índole en cuanto a la confiabilidad y validez de sus resultados. Asimismo, el costo de su empleo es un factor que deberá considerarse cuando se planifica su uso en los sistemas de medición del logro en la Región.

Capítulo IV

LA INFORMACIÓN SOBRE FACTORES SOCIALES E INSTITUCIONALES ASOCIADOS A LOS RESULTADOS

Pedro Ravela

¿Es necesario incluir cuestionarios de familia y encuestas a maestros en las mediciones de aprendizaje o alcanza con la aplicación de pruebas?

¿Para qué puede resultar útil la información sobre los contextos sociales e institucionales?

¿Es adecuadamente aprovechada la información que en el presente muchos países recogen junto con la aplicación de pruebas?

¿Es posible mejorar la calidad de los instrumentos de recolección de este tipo de información?

Introducción

En la mayoría de los países latinoamericanos se aplica, junto con las pruebas de logro, cuestionarios dirigidos a recoger información de una enorme gama de variables relacionadas con las características de las familias y hogares en que viven los alumnos, así como acerca de las características de las escuelas a las que asisten y los maestros que los atienden. Sin embargo, generalmente esta información está siendo muy poco aprovechada y no forma parte de los reportes nacionales.

La mayor parte de los países suele limitarse a informar los resultados bajo la forma de porcentajes de respuestas correctas para las pruebas en su conjunto o para partes de ellas, por lo general desagregados por jurisdicción político/geográfica (provincia, región, estado, departamento) y tipo de escuela (urbano/rural, público/privada).

En casos excepcionales, se ofrece información sobre las variables relativas a las familias, es decir, externas a los sistemas educativos, más que acerca de las variables escolares sobre las cuales los Ministerios pueden tomar decisiones.

Adicionalmente, pocos países han desarrollado trabajos sistemáticos de investigación acerca de los fac-

tores escolares asociados a las diferencias de resultados de aprendizaje entre las escuelas.

A continuación se analizarán los problemas relacionados con la subutilización de la información “de contexto”, así como con la falta de investigación sobre la relación entre esta información y los resultados educativos.

Problemas relacionados con la falta de contextualización sociocultural de la información sobre resultados de las pruebas

La ausencia de caracterización sociocultural de las poblaciones a las que “enseñan” los distintos sectores del sistema educativo impide extraer conclusiones válidas acerca de la eficacia de dicha enseñanza. Normalmente aparecen como menos eficaces los sectores del sistema educativo que atienden a la población con mayores carencias, al tiempo que aparecen como “mejores” los sistemas educativos de las provincias o regiones cuya población está más alfabetizada y vive en mejores condiciones. Del mismo modo, normalmente se reportan mejores resultados en la educación privada en relación a la educación pública, pero no se analiza el tipo de selección social que uno y otro sector hacen del alumnado que atienden.

Al respecto, cabe tener presente las diferencias que se presentan por diversas categorías.

- **Educación pública y educación privada.** Graficaremos esto con un ejemplo: En un país, la comparación de resultados entre el conjunto de las escuelas públicas y el conjunto de las privadas muestra diferencias de 25 puntos porcentuales en la proporción de alumnos que logra un nivel satisfactorio en la prueba de Matemática. Sin embargo, el análisis de los datos socioculturales indica que diferencias similares existen entre ambos tipos de escuelas, en variables tales como los niveles de educación alcanzados por los padres y madres de los niños, la existencia de libros en los hogares, el nivel general de equipamiento de los mismos y las condiciones de las viviendas. Cuando las diferencias de logro entre escuelas se analizan controlando las variables socioculturales, las diferencias en la proporción de alumnos que logran un nivel satisfactorio en el conjunto de la prueba se reducen a unos 5 puntos porcentuales y, en algunos sectores sociales, son favorables a las escuelas públicas.
- **Diferencias territoriales.** Algo similar ocurre con la presentación de los datos en función de agregaciones político/geográficas. Cuando se pre-

Pedro Ravela. Profesor de Educación Media en Filosofía, Magíster en Ciencias Sociales con Especialización en Educación. Gerente de Investigación y Evaluación en la Administración Nacional de Educación Pública (ANEP) de la República Oriental del Uruguay. Profesor de “Evaluación de la Gestión Curricular” en la Maestría en Educación de la Universidad Católica del Uruguay. E-mail: pravela@adinet.com.uy.

sentan los resultados desagregados por provincia, estado, región o departamento, sin información adicional, la conclusión inmediata para el lector no especializado es que las escuelas y los maestros deben estar trabajando mejor en aquellas regiones en que los resultados son más “altos”. Sin embargo, normalmente éstas serán las regiones con mayores tasas de alfabetización y con mejores indicadores de desarrollo en general.

- **Lo urbano y lo rural.** Los reportes nacionales suelen entregar la información desagregada en función del carácter urbano o rural de la escuela. Sin embargo, es preciso señalar que esta opción desconoce la enorme disparidad y heterogeneidad que normalmente existe al interior del mundo urbano. En dicha categoría quedan incluidas las escuelas pertenecientes a pequeños poblados del interior —probablemente muy similares a las rurales—, las escuelas ubicadas en zonas marginales de la periferia de las grandes ciudades y las de los barrios acomodados y altamente educados de esas mismas ciudades. Es discutible pues, la relevancia de comparar al conjunto de escuelas urbanas en relación a las rurales. Del mismo modo, es discutible en muchos países tratar a las escuelas rurales como una categoría homogénea, ignorando las diferencias culturales y lingüísticas que en algunos casos existen en su interior.

Recomendaciones

A partir de lo mencionado, parece necesario reflexionar sobre cómo establecer formas relevantes de caracterización sociocultural de los niveles de desagregación de la información, de modo de poder hacer comparaciones entre establecimientos, departamentos o provincias que atienden poblaciones con algún grado de similitud. En este sentido, es recomendable:

- Utilizar la información social recogida en los operativos de evaluación, u otra información sociocultural disponible a partir de los Censos o Encuestas Nacionales de Hogares, para caracterizar a los niveles de desagregación elegidos. Ello permitiría, junto con la comparación global, ofrecer comparaciones y generar “competencia” al interior de ciertos segmentos del sistema educativo que atiende a sectores de la población en cierto modo similares. En los casos en que la información se entrega desagregada a nivel de escuela, esto es aun más importante.
- Adoptar una metodología de “valor agregado” cuando se desea emitir un juicio sobre la calidad de una escuela o de una jurisdicción del sistema

educativo. Se denomina enfoque de “valor agregado” a aquellas evaluaciones en las que se intenta medir la calidad de una escuela o jurisdicción no sólo en función de sus “resultados absolutos”, sino principalmente en función de sus “resultados ajustados” por el tipo de alumnado que la escuela atiende: lo que logra por encima o por debajo de lo anticipable de acuerdo a la población con la que trabaja.

- Mejorar el diseño de instrumentos que recojan información sociocultural de base relevante, dado que en muchos países los cuestionarios están dirigidos principalmente a recoger opiniones de las familias sobre el sistema educativo y la escuela, pero no relevan datos “duros” que permitan caracterizar a esas familias.

Problemas relacionados con la información sobre las características de las escuelas y la enseñanza

Por lo general, todos los sistemas nacionales aplican en sus operativos encuestas a maestros y directores en las que se recoge información sobre materiales didácticos empleados, clima escolar y años de experiencia del maestro, entre otros. Sin embargo, en los Informes Nacionales se reporta muy poco sobre estos aspectos y su relación con los resultados de aprendizaje.

La ausencia de difusión de información respecto a las variables estrictamente escolares que están asociadas con los resultados de las pruebas, implica desaprovechar información sumamente valiosa para la adopción de decisiones de intervención y mejoramiento. Los aspectos estrictamente escolares y pedagógicos son los únicos susceptibles de ser modificados, en el corto plazo, desde la política educativa y desde las decisiones que cotidianamente maestros y directivos toman al interior de las escuelas. Su ausencia en los reportes nacionales puede contribuir a generar la imagen de que el sistema educativo nada puede hacer ante la fatalidad de las diferencias sociales. Sin embargo, una vez que los niños ingresan a la escuela, lo que allí ocurre cuenta en términos de aprendizaje. De hecho, se ha mostrado que al interior de una misma categoría social existen diferencias en los niveles de logro de las escuelas, las que son atribuibles a lo que éstas hacen o dejan de hacer.

Dos dificultades principales parecen plantearse en relación a la información sobre factores escolares:

- **La sobreabundancia de información que al respecto se recoge,** por lo general sin un plan de análisis y difusión previos, que luego hace suma-

mente difícil decidir cómo organizar esa información y cómo vincularla con los datos sobre aprendizaje.

- **La asistematicidad a través del cual se construyen los instrumentos de recolección de la información de contexto.** El diseño de este tipo de instrumentos se realiza en parte en base a la intuición, en parte para satisfacer requerimientos de información de distintas unidades de los Ministerios, en parte en base a la acumulación de conocimiento respecto a factores de efectividad y en parte en base a los modelos utilizados por otros países, pero sin que se desarrolle un proceso de pilotaje y análisis de lo que los instrumentos pueden o no rendir.

Recomendaciones

- Es necesario avanzar en el desarrollo de una metodología para el diseño de los instrumentos de relevamiento de “factores escolares” que incluya:
 - un marco conceptual explícito respecto al papel de los factores escolares, que sistematice y organice la investigación existente sobre escuelas y prácticas de enseñanza eficaces;
 - la identificación más precisa de las variables escolares que es *relevante y posible* medir en el marco de un operativo nacional de evaluación, teniendo en cuenta, aquellos aspectos sobre los cuales es posible la toma de decisiones tanto desde la política educativa como desde el interior de los establecimientos, así como las limitaciones propias de los cuestionarios autoadministrados;
 - mejores modos de formular las preguntas, así como el desarrollo *ex ante* de escalas dirigidas a medir aspectos específicos tales como el clima institucional, el empleo del tiempo en el aula, el currículum implementado, los enfoques didácticos, los tipos de actividades realizadas por los niños, la utilización de los materiales y textos además de su mera existencia, entre otros; y
 - el pilotaje y validación previa de los instrumentos.
- Se requiere avanzar, en particular, en la recolección de información acerca de lo que realmente se enseña en las escuelas. Muchas veces los niveles de logro insatisfactorios en ciertas áreas no reflejan una enseñanza “no efectiva” sino, sencillamente, ausencia de enseñanza de ciertos temas y dominios.
- Es preciso publicar la información existente en cada país sobre factores escolares, incluso simplemente bajo la forma de tablas descriptivas de la distribución de las diferentes variables. Ello con-

tribuiría al conocimiento sobre lo que está ocurriendo al interior de los sistemas educativos, permitiría comenzar a realizar comparaciones entre sistemas educativos y, hacia el futuro, podría constituirse en una forma de preparar el terreno para la construcción regional de indicadores educacionales comparables. Asimismo, permitiría ir acumulando conocimiento para afinar el tipo de preguntas que es útil formular, aligerar los cuestionarios o ir pasando en evaluaciones sucesivas a indagar nuevos aspectos.

- Se necesita diseñar mejores formas de reportar los resultados de las pruebas, junto con la información sobre contextos sociales y factores escolares. Esto implica preguntarse: ¿Cómo ofrecer a los diferentes destinatarios –autoridades, otras unidades ministeriales, maestros, opinión pública– información que permita una lectura más compleja de los datos, sin abrumar a los eventuales lectores? ¿Cómo diversificar los tipos de informes que se producen? ¿Es posible avanzar hacia ciertos “formatos tipo” más sofisticados que los existentes hasta el momento (porcentaje de respuestas correctas por jurisdicción político/geográfica y por tipo de escuela)?

Problemas relacionados con la falta de investigación sistemática acerca de factores escolares asociados con el aprendizaje

Además de introducir mejoras en las formas de reportar los resultados de las evaluaciones nacionales, **parece necesario mejorar el aprovechamiento de la información generada por los sistemas nacionales de medición con fines de investigación sobre el modo en que los diversos factores inciden sobre los aprendizajes.** Al respecto hay que tener presentes algunos problemas centrales que se enfrentan en este terreno:

- **Las investigaciones suelen utilizar evaluaciones que realizan los países sobre niveles de logro al final de ciertos grados o niveles de enseñanza, pero no necesariamente aprendizaje en un cierto período de tiempo.** El nivel de logro de un estudiante al final de cierto grado escolar depende de múltiples factores ajenos a lo que ocurrió durante ese año en su aula. Tiene relación, por ejemplo, con la historia escolar anterior de los integrantes del grupo y con la acumulación de conocimiento con la que llegaron. En rigor, una evaluación de aprendizaje cuyo objetivo es investigar acerca de los factores que explican esos aprendizajes, exige contar con mediciones de conocimientos y competencias al inicio y al final del año es-

colar. Sólo así se podrán establecer relaciones entre lo que ocurrió en la escuela ese año y el avance de los alumnos en términos de aprendizaje. Probablemente, la realización de operativos de evaluación al inicio y al final de un mismo año esté fuera del alcance de las posibilidades logísticas y económicas de los países de la Región. Sin embargo existen caminos intermedios a explorar, como realizar la medición inicial en muestras más pequeñas o, en países que evalúan grados sucesivos (por ejemplo, 5° y 6° grados de primaria), considerar como medida de aprendizaje de los alumnos del grado superior a la diferencia de logro con relación a los alumnos del grado inferior.

- **La investigación sobre factores asociados implica la utilización de técnicas estadísticas sofisticadas de carácter multivariado y plantea severas exigencias en cuanto a la conformación y calidad de las bases de datos.** Normalmente será necesario contar con información completa sobre todas las variables incluidas en el modelo para todos los alumnos, lo que no siempre es posible cuando los relevamientos se efectúan a través de cuestionarios autoadministrados y se trabaja con muestras muy grandes. Una de las principales limitaciones de los modelos estadísticos multivariados es que la posibilidad de que una variable “ingrese” al modelo depende del grado en que la misma varía en la realidad. Aquellas variables con menor variabilidad difícilmente ingresan, lo que no implica que no sean relevantes en la producción de los resultados.
- **A través de la investigación de corte estadístico, aun la más sofisticada, se puede construir ciertos tipos de conocimiento y de información, pero no es suficiente para la comprensión de los fenómenos.** Es necesario desarrollar también estrategias de investigación de carácter cualitativo o “estudios de casos”, que permitan una mirada distinta sobre aspectos que los instrumentos usualmente utilizados no pueden captar. El hecho de contar con una medición de logros educativos y contextos sociales e institucionales, constituye un formidable “mapa” sobre el cual efectuar la selección de casos relevantes para un estudio en profundidad. Asimismo, la acumulación de conocimiento en el área parece requerir también de investigaciones de corte “cuasi-experimental”, que permitan medir y controlar un conjunto de variables, como las relativas a las prácticas de enseñanza. Por otra parte, es necesario preguntarse acerca de qué tipo de decisiones de política educativa es posible tomar a partir de los resultados de un análisis estadístico.

co. Difícilmente podrá o deberá establecerse una relación directa entre los resultados de un trabajo de investigación y la toma de decisiones de política educativa. Para decirlo en forma caricaturizada, normalmente un Ministro no está esperando los resultados del análisis multivariado para decidir si compra libros o dicta una resolución para que los maestros dediquen más tiempo a enseñar quebrados. Es necesario un proceso de acumulación de conocimiento previo a la toma de decisiones, más allá de que ésta está regida además por otro tipo de consideraciones y restricciones.

Recomendaciones

- Es necesario propiciar el establecimiento de asociaciones y convenios de colaboración entre las Unidades de Medición y centros de investigación especializados, de modo que estas últimas permitan potenciar el aprovechamiento de las bases de datos existentes mediante la realización de trabajos que las Unidades no logran llevar adelante y, simultáneamente, colaborar en el mejoramiento de los instrumentos de medición y de la calidad de las bases de datos. Ello requiere, en primer término, voluntad política de parte de los Estados para facilitar el acceso a las bases de datos y, en segundo término, apoyar el desarrollo de las capacidades de investigación en estos temas. Esto incluye apoyar la capacitación de recursos humanos y la acumulación de conocimiento y experiencia en materia de investigación educativa no sólo al interior de los Ministerios de Educación sino también en las universidades y centros no estatales.

Capítulo V

ALTERNATIVAS TÉCNICAS EN RELACIÓN A LAS ESCALAS DE REPORTE DE LOS RESULTADOS DE LAS PRUEBAS DE RENDIMIENTO

Richard Wolfe

¿Cuál es el mejor modo de informar acerca de los resultados de una evaluación nacional: puntajes promedio para un conjunto de ítems, porcentaje de respuestas correctas a ítems individuales, porcentajes de alumnos que alcanzan cierto puntaje en una prueba, otros?

¿Cuándo es adecuado emplear unos u otros?

¿Cuál es el significado real de las cifras a través de las cuales se reportan los resultados?

¿Bajo qué condiciones es técnicamente válido realizar comparaciones entre las cifras obtenidas a partir de mediciones efectuadas en distintas áreas de contenidos o en distintos momentos en el tiempo?

Introducción

El propósito de este capítulo es examinar las alternativas técnicas existentes con respecto a las escalas de reporte de los resultados de las evaluaciones nacionales.

Por "escala de reporte" se entiende, en primera instancia, a qué tipos de números se usan para presentar resultados (números o frecuencias simples, porcentajes, percentiles, puntajes en escalas, etc.), pero involucra también las maneras en que se registran, procesan, transforman, agregan y presentan los datos cualitativos y cuantitativos sobre las respuestas dadas por los alumnos. Las escalas de reporte están asociadas a:

- **La granularidad** (ver primer capítulo). Por ejemplo, si las estadísticas sobre los ítems de las pruebas que miden un contenido específico se informan por separado o se suman para conformar una medida a nivel de un área de conocimientos u otro tipo de escalas.
- **La representación cualitativa y numérica de los resultados**, tales como la presentación de ítems que realmente fueron aplicados en las pruebas, la provisión de modelos ejemplares del trabajo de los alumnos, porcentajes simples para categorías de respuesta o cualquiera de las diversas maneras posibles de resumir, escalar y mostrar las distribuciones del logro escolar.

- **La consistencia y comparabilidad** de distintos contenidos y a lo largo del tiempo.

Históricamente, los reportes sobre logros educacionales se han derivado de las estadísticas educativas, con refinamientos sucesivos que progresaron desde las matrículas hacia las tasas de egreso y, luego, hacia algún tipo de "porcentaje de logro".

Cuando se piensa más detenidamente en los reportes de logros, surge la pregunta respecto a qué significan esos porcentajes. Existe una desafortunada confusión en los modos de encarar las discusiones públicas y profesionales sobre los resultados de logro. Por ejemplo, es posible ver informes de evaluación en los cuales un logro de 50% es considerado bajo (o alto), cuando en realidad ello es simplemente una consecuencia arbitraria de un proceso de desarrollo y selección de ítems mediante el cual se eligió aquellos que tenían aproximadamente 50% de dificultad (ver tercer capítulo).

Asimismo, es de crucial importancia determinar en qué medida los resultados reportados de las evaluaciones son consistentes y comparables. Por ejemplo, ¿bajo qué condiciones tiene sentido decir que los logros en matemáticas son mayores o menores que los logros en lenguaje? ¿Cómo podemos decir que el desempeño en matemáticas es mayor o mejor en sexto grado que en tercer grado? ¿Cómo podemos producir reportes que demuestren que el desempeño ha mejorado de un año al otro?

Las investigaciones educacionales y psicométricas han proporcionado varios tipos de soluciones a estos problemas:

- El análisis de los dominios de los contenidos y de las pruebas, y la determinación de la "generalizabilidad" de los puntajes y porcentajes.
- El uso de métodos estadísticos para igualar y calibrar pruebas diferentes.
- El desarrollo de teorías de respuesta al ítem (IRT), que intentan colocar a los ítems y a los estudiantes en dimensiones o escalas latentes. (Se trata de un caso especial del método estadístico).

Todos estos enfoques tienen ventajas sustanciales pero, al mismo tiempo, presentan peligros sustanciales. Antes de abordar las técnicas, es necesario considerar las maneras en que se reportan los datos de las evaluaciones y cómo ello depende del nivel de agregación, o granularidad. (ver recuadros 5, 6 y 7)

Recuadro 5

Enfoques básicos de los reportes (1)

Reporte sobre ítemes individuales

Cuando se quiere comprender en detalle qué tipos de cosas son capaces de hacer los estudiantes, se puede reportar:

1. Registros del desempeño real de los individuos, tales como su éxito o fracaso en ítemes o tareas específicos de las pruebas, calificaciones de los productos de sus tareas de desempeño, o los mismos productos de la prueba, tales como un ensayo escrito.
2. Ejemplos de desempeños individuales para ilustrar niveles de respuesta típicos o extremos dentro del conjunto de respuestas de los estudiantes, tales como ejemplos de desempeño *novato*, *competente* y *experto*.
3. Estadísticas sobre logros de grupos o poblaciones de estudiantes, tales como estadísticas de ítemes (porcentaje de respuestas correctas) o distribuciones de desempeño (porcentaje de individuos con desempeño satisfactorio) o parámetros de los rasgos latentes derivados de los modelos de IRT.

Reporte áreas de contenidos más amplias

Para generalizar los hallazgos en aspectos o maneras diversas de considerar el aprendizaje, así como para resumir los logros por áreas de contenido más amplias, es necesario realizar agregaciones de la información que se reporta. Algunas alternativas son:

1. Dar a conocer las estadísticas de ítemes, incluidas las distribuciones de las respuestas correctas y de las incorrectas. En el caso de ítemes complejos, cuyas respuestas han sido calificadas por jueces, se puede reportar las distribuciones de sus puntajes. La validez de contenido depende de la interpretabilidad del ítem como representativo de un aspecto importante del aprendizaje de la asignatura, o del conjunto de ítemes como representativo de una muestra amplia de una sub-área, así como de la interpretabilidad de la calificación misma.
2. Informar sobre promedios entre ítemes dentro de pequeñas áreas de contenidos (tópicos), tales como la multiplicación de números enteros o la extracción de información objetiva a partir de la lectura de un texto. La precisión y la interpretabilidad dependen de la calidad y el tamaño de la muestra de ítemes.
3. Llegar a medidas compuestas de todo un dominio de contenido o incluso abarcando varios dominios.

Recuadro 6

Enfoques básicos de los reportes (2)

Alternativas de métricas (escalas numéricas) para los reportes de resultados de las pruebas

1. Reportar porcentajes de estudiantes y de ítemes. El puntaje de un estudiante sería su porcentaje de respuestas correctas. La dificultad de un ítem sería el porcentaje de estudiantes que lo respondió correctamente.
2. Con un sistema de análisis a partir de la Teoría de Respuesta al Ítem (IRT), asociar las dificultades de los ítemes, así como las habilidades de los estudiantes, con valores de la escala IRT. Dicha escala tiene en su primera aplicación un rango arbitrario, que se puede fijar, por ejemplo, con una media de 500 y una desviación estándar de 150 (o de 50 y 15, o de 100 y 16). Pero en aplicaciones posteriores de acuerdo con el modelo IRT, la métrica puede mantenerse en el siguiente sentido: a través de un proceso de calibración, puede cambiarse la muestra de ítemes sin cambiar los valores de la escala para los estudiantes, o puede cambiarse la muestra de estudiantes sin cambiar los valores de la escala para los ítemes.
3. Reportar datos normativos, tales como el percentil en que se ubica el puntaje de un individuo con respecto a la distribución de puntajes de un universo de referencia de estudiantes, o el porcentaje de estudiantes de un grupo particular que está en o por encima de un porcentaje adoptado como criterio de referencia (e.g. 50%) para el universo.
4. Establecer estándares y clasificar a los individuos en categorías tales como *novicio*, *competente*, *proficiente* o *experto*, etc., de acuerdo al nivel de competencia que demuestran en las pruebas, así como informar qué proporción de un grupo (o de diversos grupos) de individuos alcanzan cada uno de esos niveles o categorías.

Por supuesto, cualquier resultado puede ser diferenciado según variables estratificadoras de los estudiantes (sexo, edad) o de las escuelas (públicas, privadas, región, etc.).

Recuadro 7

Enfoques básicos de los reportes (3)

Comparaciones

Además de lo señalado en los recuadros 5 y 6, hay otras comparaciones importantes que pueden y suelen aparecer en los reportes de las pruebas nacionales o que son inferidas por quienes los leen:

1. Comparaciones a lo largo de las áreas de contenido. Por ejemplo, si el desempeño en matemáticas es más alto que el desempeño en lenguaje.
2. Comparaciones a través de los grados. Por ejemplo, ¿es el desempeño en matemáticas de los alumnos de sexto grado relativamente más alto que el de los alumnos de tercer grado?
3. Comparaciones a lo largo del tiempo. El desempeño medido este año en un tópico y en un grado particulares, ¿es más alto que el desempeño medido el año anterior para ese mismo tópico y grado?

Muchos investigadores y psicometristas dirán que el primer tipo de comparación, entre contenidos, es casi imposible de hacer, y también es difícil en el caso de comparaciones entre grados escolares. Sin embargo, la comparación de un mismo contenido a lo largo del tiempo sí es susceptible de análisis riguroso y solución técnica. Si bien a menudo se realiza incorrectamente, con datos erróneos o análisis equivocados, cuando puede hacerse bien, suministra exactamente el tipo de información que el sistema educativo y el público necesitan para la evaluación y la corrección del progreso educacional.

Generalizabilidad de los resultados reportados

La teoría de la generalizabilidad (Cronbach, Gleser, Nanda, y Rajaratnam, 1972) constituye un modelo integral para analizar los puntajes alcanzados por los estudiantes en las pruebas y los reportes de los mismos. Es una expansión y extensión del modelo tradicional de confiabilidad, erigida sobre la noción de que los ítems de las pruebas y las condiciones de aplicación y calificación de las mismas constituyen muestras extraídas de universos más amplios de ítems y condiciones de aplicación y calificación. Así, por ejemplo, en una prueba de capacidad de expresión escrita, se puede pedir a los estudiantes que respondan por escrito a cada una de varias preguntas o reactivos. Éstos buscarían provocar la redacción de textos narrativos, informativos o de otra naturaleza. Esto significa que serían muestras de textos de una muestra de tipos de redacción. Más aun, la calificación de los textos redactados estaría a cargo de una muestra de docentes. El objetivo es generalizar el desempeño de los estudiantes a los universos correspondientes de preguntas o reactivos para la redacción, tipos de textos y docentes-calificadores.

En la teoría de la generalizabilidad, la confiabilidad de los puntajes de las pruebas es considerada desde la perspectiva de la generalización a partir de muestras a las poblaciones relevantes. El análisis de generalizabilidad implica la determinación de la variabilidad de la muestra en diferentes componentes medidos por el desempeño en las pruebas. En particular, el objetivo es determinar qué partes de la variación de los puntajes de las pruebas se deben a características estables de los individuos y grupos que son independientes de los ítems particulares y de las condiciones de aplicación, y qué parte de la variación obedece a características de los ítems y de las circunstancias de la aplicación de las pruebas. Esto último es considerado como "error de medición".

La aplicación de la teoría de la generalizabilidad al diseño de evaluaciones educacionales requiere la definición cuidadosa y formal de las poblaciones o universos de ítems y de las condiciones de medición. Las tablas de especificaciones convencionales de las pruebas son un buen punto de partida, pero es necesario añadirles una definición formal de las poblaciones o universos de ítems que corresponden a cada una de las celdas de la tabla, y de una especificación igualmente explícita acerca de las condiciones de medición, incluyendo los tipos de ítems, las maneras de administrar la prueba, los procedimientos de calificación o las condiciones de replicación (si es que la prueba en cuestión ha sido ya administrada en anterior oportunidad). Las rúbricas de calificación y las replicaciones se tornan especialmente complejas en

los casos de tareas de desempeño.

El método estadístico apropiado para el análisis de la generalizabilidad es el Análisis de Varianza. Con un conjunto detallado de información sobre las categorías de los ítems y de las condiciones de medición (lo que a veces se denomina un estudio "G"), se obtiene información rigurosa acerca de la magnitud de los diferentes componentes de la varianza, lo que permite calcular la precisión de las mediciones obtenidas y, especialmente, de las comparaciones que es posible hacer a lo largo del tiempo.

Comparabilidad de los resultados reportados

Por lo general, no es factible usar una misma prueba en dos momentos diferentes en el tiempo. Pero hay métodos que permiten poner dos pruebas paralelas o superpuestas en la misma escala y luego tratar sus mediciones como si fueran resultantes de la misma prueba.

Cuando las pruebas son paralelas, en el sentido de que tienen el mismo contenido y estructura, ítem por ítem, el alineamiento estadístico es considerado una "equiparación" (*equating*). Que el contenido sea el mismo se logra mejor asignando de manera aleatoria pares de ítems a dos formas de pruebas.

Cuando las formas de las pruebas no son estrictamente paralelas, sino que tienen una superposición sustancial —es decir, ítems en común— se aplican diferentes procedimientos estadísticos de regresión y el resultado se denomina "calibración".

Estos procedimientos para obtener comparabilidad entre las mediciones a lo largo del tiempo requieren obviamente una planificación de largo plazo para el diseño y la administración de las pruebas. No se puede desarrollar pruebas, *de novo*, cada año.

La tecnología para la comparabilidad mediante la equiparación y la calibración es estadística. Por lo tanto, está sujeta a error estadístico. Es importante calcular y dimensionar el error de equiparación o calibración y no realizar comparaciones que excedan la precisión alcanzada.

La elaboración de escalas de reporte

La Teoría de Respuesta al Ítem (IRT) está siendo propuesta como una suerte de panacea para los problemas de construcción y diseño de pruebas. Se basa en un modelo sofisticado que sugiere cómo se encuentran situados teóricamente los estudiantes

y los ítemes de las pruebas en una escala numérica común y cómo las respuestas de los estudiantes a los ítemes están estadísticamente determinadas por sus posiciones relativas en dicha escala. De ese modelo se ha derivado toda una tecnología que parece resolver varios problemas difíciles en la equiparación de formas de prueba, en la calibración de los resultados a lo largo del tiempo y en la construcción de mediciones paralelas por vías más “fuertes” que los métodos estadísticos clásicos. Un importante ejemplo de esa fortaleza es que es posible establecer una relación entre los estudiantes en un nivel de puntaje particular y los ítemes que ellos dominan. Esto se realiza de una manera referida a criterios, sin tener que tomar en cuenta poblaciones normativas de estudiantes o de ítemes.

El hecho de que esta referencia funcione o que los problemas de comparabilidad sean realmente resueltos, depende de la adecuación de los modelos con respecto a un sistema de pruebas específico, es decir, con respecto a poblaciones particulares de ítemes y de estudiantes. Un aspecto cuestionable del modelo, por lo menos en muchas aplicaciones educacionales, es que presupone que fundamentalmente existe una sola dimensión en la variación en la dificultad de los ítemes y en la capacidad de los estudiantes.

El uso de los métodos IRT en el análisis estadístico y la interpretación de los resultados de las evaluaciones educacionales requiere programaciones y análisis estadísticos sofisticados, aunque esto es también cierto para los métodos convencionales. Es muy importante recordar que los sistemas de análisis, sean el clásico o el IRT, son sólo instrumentos estadísticos para resolver algunos problemas técnicos de análisis de ítemes, de equiparación y calibración de pruebas y de evaluación de la precisión estadística de los resultados. Los problemas fundamentales tanto conceptuales como pedagógicos son, por un lado, la definición de los universos de ítemes y, por otro lado, la construcción y selección de una muestra de esos universos.

El dilema respecto a las escalas de reporte

El requerimiento clave para el desarrollo de escalas de reporte apropiadas para las evaluaciones educacionales es que la interpretación que se haga a partir de los resultados sea auténtica. Esto significa:

- que a las cifras reportadas se les atribuirá significados que estén justificados por las características de las poblaciones de contenidos y de estudiantes, y por el sistema empleado para la medición;

- que la precisión de los reportes será tomada en cuenta correctamente; y
- que las comparaciones al interior de y entre partes de la evaluación serán hechas sólo en la medida en que estén justificadas tanto desde el punto de vista sustantivo como del estadístico.

El dilema en el diseño de escalas de reporte en educación es que las cuestiones técnicas son complejas, mientras que los conocimientos e intuiciones sobre las mismas que tienen los usuarios están muy arraigados, aunque errados (por ejemplo, piensan que un puntaje de 50% siempre significa un fracaso o un logro mínimo). Esto hace muy difícil encontrar los medios apropiados para la comunicación de los resultados.

Capítulo VI

CONCLUSIONES Y RECOMENDACIONES

Pedro Ravela

El propósito de este capítulo es resumir las principales reflexiones formuladas a lo largo del documento y proponer algunas líneas de acción que los organismos regionales interesados en apoyar el desarrollo de los sistemas nacionales de evaluación de aprendizajes y PREAL, a través de su Grupo de Trabajo sobre Estándares y Evaluación, deberían impulsar en los próximos años.

Durante la década de los 90' se desarrolló en América Latina una primera fase de instalación de sistemas de evaluación de aprendizajes a nivel nacional. Tuvo lugar, además, una primera experiencia de evaluación a nivel regional conducida por UNESCO/OREALC. El desarrollo de estas experiencias constituye una clara manifestación de la preocupación de los gobiernos por producir información sobre los aprendizajes que se logran al interior de los sistemas educativos. En un contexto internacional en que el conocimiento y las capacidades de los individuos serán cada vez más importantes para el desarrollo y competitividad de los países, es previsible que en los próximos años este esfuerzo se mantenga. Asimismo, en la medida en que crece la conciencia respecto a la necesidad de incrementar los recursos destinados al sector educación, será imprescindible contar con información que permita evaluar el impacto de la inversión adicional de recursos y monitorear en forma permanente y adecuada los avances o retrocesos en los resultados del sistema educativo.

El panorama resultante de esta primera década muestra una importante diversidad de enfoques y experiencias en cuanto al tipo de conocimientos y competencias que son evaluados, a la periodicidad de las evaluaciones, a los grados y áreas curriculares que abordan, al tipo de variables contextuales sobre las que se recoge información, y a los análisis y formatos de devolución de información, entre otros. Detrás de esta heterogeneidad de experiencias y enfoques existe, en muchos casos, un esfuerzo de reflexión y construcción de un modelo a nivel nacional. En otros, se han adoptado modelos en forma menos reflexiva. En todos los casos, se aprecia la fragilidad propia de los primeros pasos dados en un terreno nuevo y desconocido.

La mayoría de estos sistemas de evaluación se encuentra aún en una fase de institucionalización, ya que su continuidad en el tiempo está fuertemente atada a los cambios políticos y/o a los recursos aporta-

dos por organismos internacionales de crédito. Por otra parte, cabe destacar que los sistemas de evaluación compiten por recursos con otras actividades y necesidades igualmente importantes. Por lo tanto, su sostenibilidad depende de que se aprovechen y maximicen los beneficios que en cierto modo "prometen" a la política educativa.

En este contexto, el presente documento busca aportar a la reflexión sobre los próximos pasos a dar para fortalecer los sistemas de evaluación de aprendizajes en América Latina, con la convicción de que los mismos pueden constituirse en una herramienta de política educativa eficaz para promover un mejoramiento de los aprendizajes a los que acceden los niños de todos los estratos sociales. Y también con la convicción de que, para que esto último ocurra, es necesario ingresar en una nueva etapa de revisión y consolidación de los sistemas nacionales de evaluación, para lo cual se torna imprescindible abordar en profundidad un conjunto de debates sobre opciones técnicas y políticas en esta materia.

Se han priorizado cuatro grandes temas centrales.

En primer término, se analiza un conjunto de alternativas técnicas relacionadas con el diseño global del sistema de evaluación y con los objetivos que se espera que el mismo cumpla. Por un lado, se plantea la existencia de una relación inversa entre cobertura curricular y cobertura poblacional. Cuanto más detalladamente se desee conocer qué aprenden los alumnos en cierto nivel del sistema educativo, menos factible será contar con información desagregada a nivel de distritos y establecimientos educativos. Por el contrario, si el propósito es generar información con estos últimos niveles de desagregación, sólo será posible obtener mediciones más globales y menos detalladas de lo que los alumnos aprenden. En última instancia, la opción depende del rol que se espera que el sistema de evaluación desempeñe en la política educativa. En el quinto capítulo se analiza la existencia de múltiples alternativas técnicas para reportar los resultados, que también dependen del enfoque dado al sistema de evaluación, y que debieran ser discutidas en profundidad para mejorar la calidad y pertinencia de la información que están reportando los sistemas de evaluación en América Latina.

En segundo lugar, se alerta sobre la necesidad de analizar más cuidadosamente la información que producen los sistemas de evaluación, en el sentido de validar las conclusiones e interpretaciones que de dicha información se realiza. Ello implica, por un lado, una labor de "formación permanente" de los usuarios —

autoridades, opinión pública, medios de comunicación, maestros— respecto a los usos válidos de los distintos tipos de información que aportan estas evaluaciones y respecto al tipo de interpretaciones y conclusiones que *NO* es posible extraer válidamente de ella. Para esto, el primer paso es que las propias unidades de evaluación mejoren el modo en que reportan los resultados e incluyan reportes técnicos completos y comprensibles acerca de las limitaciones y potencialidades de la información que producen, y acerca de los procedimientos de producción de la información.

En tercer lugar, se plantea la necesidad de profundizar la discusión respecto al enfoque de diseño de pruebas más adecuado. En la mayoría de los países de la Región, el diseño de pruebas ha estado fuertemente marcado por los principios y procedimientos propios de la elaboración de pruebas referidas a normas, en las que se privilegia la función de ordenamiento o “discriminación” entre grupos o individuos. Este enfoque está fuertemente marcado por su función principal, que históricamente ha sido la de seleccionar individuos para el ingreso al ejército o a las universidades. En esos casos no importaba tanto si el individuo dominaba o no ciertos campos del conocimiento, sino distinguir a los individuos más aptos de los menos aptos. El enfoque de pruebas referidas a criterios, en cambio, se propone como objetivo central comprobar si los individuos dominan un cierto campo de contenidos y/o destrezas, y se busca hacerlo del modo más exhaustivo posible. Ello implica que, en el diseño de las pruebas, no necesariamente deben ser descartados aquellos ítemes que resultan fáciles o difíciles, dado que, aún cuando no sirvan para discriminar entre malos y buenos estudiantes, pueden aportar información relevante acerca del grado en que los conocimientos y competencias definidos como fundamentales están siendo logrados por los estudiantes en el sistema educativo. Este enfoque parece ser el más adecuado para sistemas de evaluación de aprendizajes que buscan producir información relevante para la mejora del currículum y la enseñanza. Para ello es imprescindible enriquecer y mejorar los procedimientos de diseño de pruebas e ítemes. Asimismo, se requiere avanzar en el diseño y corrección estandarizada de pruebas de desempeño.

En cuarto lugar, se señala que es necesario mejorar sustancialmente no sólo la calidad de los instrumentos de medición de aprendizajes, sino también los instrumentos de medición de aspectos relevantes del contexto social y escolar en que ocurren los aprendizajes. La mayoría de los países recoge información sobre variables sociales y escolares, pero en muchos casos la calidad de la misma no es suficiente y, me-

nos aún, el aprovechamiento que de ella se hace para el análisis de los resultados de aprendizaje y la investigación. Es necesario, por un lado, mejorar la medición de variables de tipo sociofamiliar, con el fin de contextualizar socialmente el análisis y reporte de los resultados y evitar la falacia de atribuir a los establecimientos educativos el mérito o la culpa por resultados que en realidad obedecen a la selección social del alumnado. Por otro lado, es preciso mejorar la medición de variables institucionales y pedagógicas, para desarrollar investigaciones que permitan comprender mejor la compleja trama de factores que intervienen en el logro de los aprendizajes y, de este modo, enriquecer el horizonte conceptual y la base empírica de la toma de decisiones en materia de política educativa.

Las reflexiones y análisis anteriores permiten afirmar que, si bien al cabo de esta primera década de instalación se han dado pasos muy importantes, lo que se está haciendo no es suficiente. Es necesario inventar cosas nuevas. En este sentido, una posible agenda de trabajo para los pasos a dar en los próximos años, estaría centrada en tres ejes principales:

1. Analizar el papel de los sistemas de evaluación en la política educativa, es decir, la estrategia a través de la cuál se espera que un sistema de evaluación nacional de aprendizajes tenga algún impacto en la mejora de los aprendizajes que se logran en el sistema educativo.

Se requiere propiciar instancias de discusión sobre el rol que se espera del sistema de evaluación en el marco de la política educativa. Si bien existe un consenso genérico en cuanto a que el simple acto de evaluar —por el mero hecho de dar “visibilidad” a los resultados— puede tener un efecto positivo sobre el sistema educativo, es necesario delinear a nivel nacional una estrategia más específica al respecto. En algunos casos se ha buscado que el impacto se produzca a través del control de los padres sobre la calidad de las escuelas; en otros, por la vía de utilizar las pruebas para certificar la aprobación de cierto nivel de enseñanza; mientras que en otros se ha utilizado la información principalmente para promover la actualización docente y el aprendizaje profesional al interior de las escuelas.

En este documento se ha planteado una amplia gama de alternativas de política en materia de evaluación, destacando que las diversas opciones técnicas dependen de decisiones de política relativas a la finalidad de las evaluaciones. En la mayoría de los países de la Región ha existido la premisa básica general de que “evaluar ayuda a mejorar”, pero ha faltado una reflexión en profundidad sobre las alternativas de po-

lítica y estrategia en materia de evaluación de aprendizajes y sobre la necesidad de articular diferentes finalidades en un diseño coherente, técnicamente adecuado y pensado para el largo plazo. Esta carencia está relacionada, asimismo, con las urgencias que impone la puesta en marcha de un operativo nacional de evaluación, la necesidad de cumplir con plazos y compromisos asumidos con organismos internacionales.

La reflexión que se propone debería involucrar preguntas tales como:

- ¿Conviene que la evaluación tenga consecuencias “fuertes” para las escuelas y maestros —ya sea bajo la forma de incentivos explícitos o bajo la forma de la publicación de un ranking de resultados—, o es preferible que cumpla una función fundamentalmente informativa?
- ¿De qué modo articular los esfuerzos de evaluación con los esfuerzos de reforma y actualización de las currícula? ¿De qué modo pueden las evaluaciones contribuir a mejorar la definición de las metas e indicadores de logro curriculares?
- ¿Se desea contar con información exhaustiva acerca de las competencias y conocimientos de los alumnos a nivel nacional o se prefiere producir información menos detallada pero tenerla a nivel de cada establecimiento?
- ¿Es conveniente desarrollar pruebas nacionales de acreditación; es decir, que determinen la aprobación o reprobación de los alumnos al cabo de algún nivel de la enseñanza?
- ¿Se espera que el sistema de evaluación permita constatar avances o retrocesos a lo largo de los años? ¿En qué áreas curriculares y en qué niveles del sistema educativo?
- ¿Con qué frecuencia realizar operaciones nacionales de evaluación?

En relación con este eje, en el futuro inmediato PREAL y otras entidades regionales deberían facilitar la realización de eventos de debate y presentación de experiencias nacionales en materia de diseño del sistema nacional de evaluación y su papel en la política educativa. Se podría desarrollar un “observatorio” internacional sobre el tema, propiciando la construcción de estudios de casos de países de la Región y del mundo desarrollado en los que se describa: diversas modalidades de diseño de los sistemas nacionales de evaluación; las condicionantes históricas, sociales y políticas que constituyeron el contexto de su desarrollo; las características técnicas del diseño; y los impactos, costos, beneficios y efectos perversos constatables en cada caso. Esta información podría luego difundirse bajo la forma de una serie de publicaciones o en el marco de seminarios de discu-

sión. Un desafío importante en este ámbito será el de involucrar en estos debates a otros actores de la política educativa y no sólo a los técnicos directamente relacionados con las evaluaciones.

2. Mejorar la calidad técnica de los diversos aspectos constitutivos de los sistemas de evaluación, en especial el diseño de los instrumentos de recolección de información y los modos de procesar y reportar los resultados.

A lo largo del documento se insistió respecto a que las decisiones técnicas que son aptas para ciertos fines no lo son para otros. Por tanto, es necesario garantizar la congruencia entre las opciones de política y las decisiones técnicas que definen el diseño del sistema de evaluación. Asimismo, se señaló que es imprescindible mejorar el diseño de los instrumentos de medición y garantizar una interpretación apropiada —válida— de los resultados que se obtienen. Todo ello exige intensificar los esfuerzos de capacitación de cuadros técnicos y la acumulación de conocimiento y experiencia en una materia que aún es nueva en la región y sobre la que existe escasa “masa crítica”.

En este terreno PREAL y otras organizaciones de carácter regional podrían facilitar el contacto de los profesionales de la Región vinculados al área de la evaluación, con especialistas de la comunidad internacional. Esto se podría hacer a través de seminarios en los que participarían países que estén dispuestos a someter sus instrumentos y procedimientos de evaluación al escrutinio de otros, para analizar detalladamente las fortalezas y debilidades de los instrumentos y procedimientos empleados. Ello podría servir también para desarrollar procedimientos comunes a varios países para medir ciertos aspectos relevantes, tanto en el terreno de los aprendizajes como en el de las variables sociales y escolares.

También sería de gran utilidad impulsar la formulación de un conjunto de estándares técnicos que deberían cumplir las pruebas, los procedimientos de implementación de los operativos de evaluación, los procesos de conformación y procesamiento de las bases de datos, y los reportes de resultados.

Finalmente, cabe destacar que toda iniciativa dirigida a propiciar la formación sistemática de cuadros en materia de evaluación, tanto a nivel de agentes estatales como de organizaciones no gubernamentales, será bienvenida.

3. Discutir las estrategias de uso y difusión de los resultados de las evaluaciones, que si bien está estrechamente relacionado con los dos puntos anterior-

res, merece una consideración específica.

Todavía es muy poco lo que se sabe acerca del uso e impacto que la información producida por los sistemas de evaluación de aprendizaje tiene en sus potenciales usuarios. Aún no se ha construido evidencia empírica acerca de:

- el modo en que los resultados son analizados y utilizados en las escuelas;
- el grado en que las familias y la opinión pública reciben y comprenden la información;
- el modo en que la misma es empleada como insumo en la toma de decisiones de política educativa por parte de los Ministerios de Educación;
- y
- el grado en que las bases de datos son aprovechadas por académicos y centros de investigación para producir conocimiento.

En este terreno, PREAL y otros organismos regionales deberían propiciar el desarrollo de trabajos de investigación que permitan recoger evidencia empírica acerca de los efectos que los distintos tipos de reportes de resultados de las evaluaciones nacionales tienen en diversos públicos.

En segundo término, sería sumamente útil realizar algún tipo de evento que permita "escuchar a los destinatarios":

- ¿Qué tipo de información esperan recibir del sistema de evaluación de aprendizajes las diversas audiencias: periodistas, padres, maestros, políticos, autoridades y técnicos de los Ministerios de Educación?
- ¿Cómo perciben la información que actualmente se les está entregando? ¿Han podido comprenderla? ¿La han utilizado de algún modo?
- ¿Qué visión general tienen acerca de los sistemas de evaluación de aprendizajes a nivel nacional? ¿Cuáles son sus expectativas y prejuicios acerca de los mismos?

Ello permitiría analizar la demanda potencial de información, aprender acerca de los modos pertinentes de informar a los diversos tipos de usuarios potenciales y desarrollar diferentes tipos de formatos de informe adecuados a cada uno de ellos. Asimismo, dado que la mayoría de los potenciales destinatarios probablemente no tenga una noción cabal de lo que espera de un sistema de evaluación de aprendizajes, un evento como el descrito aportaría pistas para desarrollar estrategias de "formación de la demanda". Es decir, ayudaría a pensar qué tipo de acciones desarrollar para ir conformando en nuestros países una cultura con relación a la evaluación de aprendizajes

que incluya aspectos tales como la conciencia acerca de la necesidad de contar con esta información, los alcances y limitaciones de la misma, el tipo de interpretaciones válidas, los modos de utilizarla y los tipos de información que es posible demandar al sistema de evaluación para fines específicos.

PREAL estimula la participación de actores estratégicos del campo educacional y de los ámbitos de la producción, la política, la sociedad civil y la cultura, promoviendo un debate informado para la formulación y ejecución de políticas educativas consensuadas. Para ello, ha constituido una red regional compuesta por los siguientes centros nacionales de investigación y políticas públicas:

Fundación Getulio Vargas
Brasil

Instituto SER de Investigación
Colombia

Centro de Investigación y Desarrollo de la Educación (CIDE).
Chile

Corporación Participa
Chile

Instituto de Investigación para el Mejoramiento de la Educación Costarricense (I.I.M.E.C.).
Costa Rica

Fundación Empresarial para el Desarrollo Educativo (FEPADE).
El Salvador

Asociación de Investigación y Estudios Sociales (ASIES).
Guatemala

Centro de Investigaciones Económicas Nacionales (CIEN).
Guatemala

Consejo Nacional de Educación Maya (CNEM).
Guatemala

GRUPO BASICO, S.A.
Guatemala

Fundación para la Educación Ernesto Maduro Andreu (FEREMA).
Honduras

Universidad Centroamericana (UCA).
Nicaragua

Foro Educativo
Perú

FLACSO / Plan Educativo
República Dominicana

Instituto de Estudios Superiores de Administración (IESA).
Venezuela

13



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

El Programa de Promoción de la Reforma Educativa en América Latina y el Caribe, es un proyecto conjunto del Diálogo Interamericano, con sede en Washington, y la Corporación de Investigaciones para el Desarrollo, con sede en Santiago de Chile.

Los objetivos básicos del PREAL son promover el diálogo regional informado sobre política educacional, situar el tema de la reforma educativa como una prioridad en la agenda política de los países de la región, crear espacios para la búsqueda de consensos y difundir experiencias exitosas en materia educativa.

La ejecución de las actividades se realiza a través de Centros Asociados de Investigación y Políticas Públicas en diversos países de la región y comprenden la realización de estudios, la organización de debates y la promoción de diálogos públicos sobre temas de política educacional y reforma educativa.

Las actividades regionales del Programa, incluyendo esta publicación, son posibles gracias al apoyo que brinda el Banco Interamericano de Desarrollo (BID), la United States Agency for International Development (USAID), el Canadian International Development Research Centre (IDRC), la GE Fund y otros donantes.



Inter-American Dialogue • 1211 Connecticut Ave. N.W. Suite 510
Washington, D.C. 20036 U.S.A. • Tel:(202) 822-9002
Fax:(202) 822-9553 • E-mail: iad@thedialogue.org
Internet: www.thedialogue.org & www.preal.org

CINDE • Santa Magdalena 75, Piso 10 • Oficina 1002 • Providencia
Santiago, Chile • Tel: (56-2) 334-4302
Fax:(56-2) 334-4303 • E-mail: infopreal@preal.org
Internet: www.preal.org

