PREAL

Programa de Promoción de la Reforma Educativa en América Latina y el Caribe

GRUPO DE TRABAJO SOBRE ESTÁNDARES Y EVALUACIÓN

LOS PRÓXIMOS PASOS: ¿HACIA DÓNDE Y CÓMO AVANZAR EN LA EVALUACIÓN DE APRENDIZAJES EN AMÉRICA LATINA?

Pedro Ravela (editor)
Richard Wolfe
Gilbert Valverde
Juan Manuel Esquivel

grade

Grupo de Análisis para el Desarrollo

Grupo de Trabajo sobre Estándares y Evaluación

LOS PRÓXIMOS PASOS:

¿HACIA DÓNDE Y CÓMO AVANZAR EN LA EVALUACIÓN DE APRENDIZAJES EN AMÉRICA LATINA?

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

Marzo del 2000

Este trabajo se desarrolló colaborativamente en un taller realizado en GRADE, en Lima, en agosto de 1999, a iniciativa de la coordinadora del Grupo de Trabajo, Patricia Arregui. El Grupo de Trabajo sobre Estándares y Evaluación de GRADE/PREAL es una de las actividades del Programa de Promoción de la Reforma Educativa en América Latina que lideran el Diálogo Interamericano de Washington, D.C. y CINDE, de Santiago de Chile. Cuenta con apoyo financiero del BID, de USAID, de IDRC, del GEFund y obtiene recursos para actividades puntuales de una diversidad de fuentes.

LOS PRÓXIMOS PASOS: ¿HACIA DÓNDE Y CÓMO AVANZAR EN LA EVALUACIÓN DE APRENDIZAJES EN AMÉRICA LATINA?

Tabla de contenidos

INTRODUCCIÓN

Pedro Ravela

Durante la década de los 90' prácticamente la totalidad de los países de la región ha puesto en funcionamiento algún tipo de sistema nacional de evaluación de aprendizajes.

Un reciente documento elaborado por Carlos Rojas y Juan Manuel Esquivel para el Banco Mundial da cuenta de la existencia de sistemas nacionales de evaluación de aprendizajes en no menos de 20 países de la región de América Latina y el Caribe, incluyendo una rápida descripción de los objetivos y características más generales de cada uno de dichos sistemas. La mayoría de estos sistemas fue creada recién a partir de 1992.

El esfuerzo por desarrollar sistemas nacionales de evaluación de aprendizajes se apoya en algunas premisas generales que son ampliamente compartidas por académicos y responsables de la implementación de políticas educativas.

Para empezar, la educación es por naturaleza propia una actividad "opaca" en cuanto a sus resultados. En otras áreas de la actividad humana es más sencillo para la sociedad "ver" los resultados de lo que se hace. En educación ello no ocurre así. El hecho de que los niños estén o no aprendiendo lo que se espera no es algo que pueda ser directamente percibido por la sociedad y por las familias. Normalmente, los juicios que las familias pueden hacerse acerca de la calidad de la escuela a la que asisten sus hijos están basados en aspectos visibles tales como el orden existente, el trato que reciben los niños, la proposición de tareas para realizar en el hogar, etc. Pero difícilmente pueda el padre o madre del niño tener una visión apropiada de los conocimientos y competencias que su hijo está adquiriendo en la escuela. Varios trabajos de investigación muestran una importante asociación entre los juicios de las familias sobre la escuela y su propia trayectoria educacional: normalmente las familias con mayor trayectoria en la educación formal son las más exigentes y menos satisfechas con la escuela a la que asisten sus hijos. En cambio, las familias menos educadas suelen tener opiniones más positivas sobre las escuelas.

Del mismo modo, los resultados suelen ser "opacos" 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.

Asimismo, para las autoridades y otros tomadores de decisiones en materia de política educativa, ya no son suficientes los indicadores tradicionales relativos a la matrícula, cobertura, repetición y deserción. En un contexto en el que el desafío

principal para la política educativa no se reduce a la ampliación del acceso al sistema, sino que el eje de las preocupaciones ha pasado a ser cómo garantizar que el acceso al sistema se traduzca en acceso equitativo a los conocimientos y competencias fundamentales para el desempeño social, un sistema de evaluación que produzca información relevante sobre este último aspecto pasa a ser de 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 los sistemas educativos.

Sobre estos supuestos generales, los países de la región han puesto en marcha diversas experiencias con 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, que suelen incluir una demanda en este sentido en 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 otros países trabajan con muestras de escuelas y/o grupos. Entre quienes trabajan a nivel censal han existido opciones diferentes. Algunos países 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. En tanto, otros países han optado por la devolución de la información a cada escuela con carácter confidencial, apostando al uso de la información como instrumento de aprendizaje profesional para los educadores. Otros 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 por lo general presentan información de resultados generales a nivel nacional, con algunos niveles de desagregación por área geográfica y tipo de escuela. Algunos países realizan importantes esfuerzos por acompañar la devolución de resultados con materiales de orientación didáctica para los docentes, en los que se explicitan las áreas de menor logro y los problemas de aprendizaje y de enseñanza que pueden estar involucrados. En otros casos, también es preciso decirlo, después de varios operativos de evaluación y varios años de trabajo, nunca se han dado a conocer resultados en forma pública.

Prácticamente todos los países evalúan logros en Lenguaje y Matemática. Existe luego 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.

Más recientemente, la UNESCO ha llevado adelante la experiencia del Laboratorio Latinoamericano de Medición de la Calidad de la Educación, en el que ha participado un buen número de países y que ha permitido desarrollar una primera experiencia de evaluación internacional en la región. La existencia del Laboratorio es un fuerte indicador de la importancia que la evaluación sistemática de aprendizajes ha adquirido en la región y del interés por la investigación sobre el funcionamiento de los sistemas educativos y sobre la enseñanza y el aprendizaje en América Latina.

El panorama suscintamente 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, la sociedad y los cuerpos docentes comienzan a valorar y comprender la necesidad de este tipo de evaluaciones.

Sin embargo, simultáneamente, al cabo de una primera etapa de implantación de su sistema nacional de evaluación, muchos países se encuentran ingresando en una etapa de revisión de lo hecho hasta el momento, 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:

a. insuficiente aprovechamiento de la información producida por los sistemas de evaluación, lo que tiene como consecuencia el insuficiente impacto del sistema de evaluación en el conjunto del sistema educativo;

b. insuficiente calidad y capacidad de evaluación de aprendizajes complejos en las pruebas que están siendo aplicadas;

c. debilidades técnicas en los procesos de desarrollo y validación de los distintos instrumentos de medición.

Las insuficiencias señaladas se derivan de que, como era lógico que ocurriera en una primera etapa de constitución de los sistemas de evaluación, se trabajó con un abanico limitado de opciones técnicas y políticas. No obstante la variedad de experiencias reseñadas más arriba, 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 otras áreas. En la mayoría de los países de la región, con anterioridad a los años 90’ prácticamente no existía experiencia acumulada en materia de evaluaciones de aprendizaje a gran escala, ni personal preparado al interior de los Ministerios de Educación, y tampoco instituciones o centros especializadas fuera de los Estados. 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 tipo de intento de medición en el área educativa.

En este contexto, cada país fue dando sus primeros pasos y construyendo su propia capacidad institucional en materia de evaluación, en el marco de un conjunto limitado de alternativas técnicas y políticas que tuvo disponibles.

En el presente, parecen estar dadas las condiciones para realizar un "salto cualitativo" en materia de evaluación, una vez que se ha transitado la etapa de las primeras experiencias, que se ha superado la absorbente preocupación inicial por las enormes e ineludibles 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. Sin duda muchos de los debates sobre opciones técnicas y políticas en materia de evaluación que es posible plantearse ahora, no hubieran sido posibles algunos años atrás, cuando casi no existían referentes reales en la región ni los saberes básicos sobre los cuales encauzar esos debates.

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

a. 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;

b. la discusión sobre las **opciones técnicas**: qué abanico de tipos de pruebas, instrumentos complementarios, procesamientos y análisis de información existen hoy en el mundo y cuáles son los más adecuados para los fines propuestos.

La mayor parte de los países ha puesto en marcha su sistema de evaluación de aprendizajes sobre la base de definiciones muy genéricas acerca del rol e impactos que esperan que tenga en el sistema educativo y sin una clara conciencia en cuanto a que las decisiones técnicas que se adoptan en el proceso de implementación de las evaluaciones pueden servir para unos fines pero no para otros. Si el objetivo es evaluar el currículum, puede ser suficiente con muestras pequeñas pero es necesario desarrollar una amplia diversidad de pruebas y reactivos que abarquen los diversos conocimientos y competencias que se espera los alumnos dominen. Si, por el contrario, el objetivo es evaluar la gestión de las distintas jurisdicciones o de las propias escuelas, puede ser suficiente con una menor diversidad de reactivos en las pruebas pero es necesario trabajar con muestras de amplia cobertura o a nivel censal.

Sin pretensiones de exhaustividad, se puede afirmar que un sistema nacional de evaluación de aprendizajes puede proponerse alguna o varias de las siguientes finalidades:

1. ***Evaluar la productividad de los maestros a los efectos de 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 establecer incentivos de tipo monetario u otro para aquellos maestros cuyos niños 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 lo que se espera del sistema de evaluación es que sirva para proporcionar 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 lo que se espera del sistema de evaluación es que produzca información útil para realimentar la práctica pedagógica de los maestros. Lo que se busca es entregarles información detallada sobre lo que los alumnos están aprendiendo y lo que no, con el fin de enriquecer la discusión técnica de los docentes y la búsqueda de nuevos caminos para mejorar la enseñanza.

4. ***Establecer la acreditación de los alumnos que finalizan un determinado nivel de enseñanza.*** En este caso se espera de la evaluación que ofrezca información adecuada para decidir si un estudiante individualmente considerado ha logrado los conocimientos y competencias que se considera indispensable 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.*** En este caso el propósito principal de la evaluación es producir información que sea 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. Lo que se busca es generar una cultura de la evaluación en relación al sistema educativo, que implique que éste debe informar sistemáticamente acerca de los resultados de su labor y que contribuya a que la sociedad en general esté atenta y preocupada por lo que ocurre en el sector educativo.

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, normalmente 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. Uno de los propósitos de las evaluaciones nacionales es el de ofrecer información pertinente para la evaluación del impacto de este tipo de intervenciones específicas.

10. ***Realizar estudios de tipo costo-beneficio***. Otra de las expectativas que muchas veces existe en relación a los sistemas de evaluación de aprendizajes es que proporcionen información útil para la evaluación de 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.*** Finalmente, los sistemas de evaluación de aprendizaje generan importantes bases de información que pueden resultar útiles para el desarrollo de trabajos de investigación 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.

Como resulta obvio, es imposible que un mismo diseño del sistema de evaluación sirva para todos los fines señalados -- y otros que el lector podría incorporar--. Algunos pueden ser perseguidos con un mismo diseño, pero otros son incompatibles entre sí. Por otra parte, cada una de estas finalidades tiene sus propias exigencias técnicas. Para algunos de estos fines son adecuados ciertos tipos de pruebas que no lo son para otros. Para algunos de estos fines se requieren ciertos tipos de muestras que no son adecuadas para otros fines. Las definiciones técnicas que sirven para un fin no sirven para otro o, lo que es peor, pueden dar lugar a graves malentendidos cuando se las utiliza para ese otro fin.

La experiencia indica que 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 acerca de las definiciones técnicas más adecuadas para cada fin. 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 sobre la marcha las autoridades ministeriales comiencen a demandar 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 a asumir y ordenar las decisiones técnicas respecto a la conformación de las bases de datos, la conformación de los bancos de reactivos, la comparabilidad de las evaluaciones, etc.

Por todo lo antedicho, el momento actual de los sistemas de evaluación de la región parece propicio para una mirada más amplia a la diversidad de opciones disponibles y a la diversidad de experiencias y caminos transitados en diversas partes del mundo, para desde allí replantearse con mayor propiedad las opciones a asumir al interior de cada país.

En este contexto, el propósito del presente documento es realizar un aporte 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 pretende 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 que se desarrolló en la ciudad de Lima entre el 17 y el 20 de agosto de 1999. Se convocó a un conjunto de especialistas en el tema que tuvieran un 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. El propósito fue formular, a partir de dicho conocimiento, un conjunto de aportes y reflexiones en torno a la relación entre definiciones técnicas y finalidades de los sistemas de evaluación. El documento pretende arrojar luz sobre lo que es posible hacer y lo que no a partir de ciertas definiciones técnicas y, simultáneamente, ampliar el abanico de opciones a considerar.

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:

a. El problema del diseño global del sistema nacional de evaluación. Sobre este aspecto 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 --. Al tratamiento de este tema está dedicado el primer capítulo. Asimismo, en el capítulo quinto se analiza la relación entre los fines del sistema de evaluación y los modos de reportar los resultados, y se propone un conjunto de alternativas técnicas que al respecto existen.

b. 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.

c. El problema de los paradigmas de construcción de pruebas de medición de aprendizajes o logros. En este capítulo 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 plantea un conjunto de reflexiones en torno al problema de la validez en cada uno de los paradigmas.

d. 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 en relación a las características socioculturales de los alumnos, como en relación a las características propias de las escuelas y maestros. 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 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 a través del correo electrónico.

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. El capítulo relativo a los problemas de validez estuvo a cargo de Gilbert Valverde. El capítulo sobre pruebas referidas a normas y pruebas referidas a criterios fue responsabilidad de Juan Manuel Esquivel. Pedro Ravela tuvo a su cargo la redacción del capítulo sobre factores asociados, de esta introducción y de las conclusiones. Richard Wolfe redactó los capítulos primero y quinto, relativos al diseño de los sistemas de evaluación. Participaron además de las instancias de discusión colectiva del trabajo y de la revisión de los sucesivos borrados Patricia Arregui y Santiago Cueto.

Dada la metodología de trabajo seguida, se optó por respetar la diversidad de estilos en los capítulos, situación que seguramente percibirá el lector.

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?

Los diseños de los sistemas nacionales de evaluación educacional en América Latina son tan variados como cualquier otro aspecto de la educación y de la cultura. Ellos 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.

Al mismo tiempo, cuando examinamos en detalle los diferentes sistemas de evaluación, encontramos características comunes que se derivan 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". Por "granularidad" nos referimos a 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 todos los estudiantes o escuelas individualmente. De igual manera, puede haber enormes 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.

A. ¿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 encontramos los siguientes niveles de análisis y de reporte:

A1. Poblaciones nacionales (o internacionales), tales como la población de escolares matriculados en tercer grado de educación primaria.

A2. Los principales estratos definidos educacional, política y socialmente, tales como estudiantes en escuelas públicas, escuelas en las áreas rurales, o estudiantes en programas bilingües.

A3. Principales divisiones regionales, tales como regiones geográficas, provincias, o estados.

A4. Jurisdicciones menores, tales como ciudades o municipalidades.

A5. Escuelas.

A6. Salones de clase (o profesores).

A7. 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, por supuesto, 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, entonces es obvio que se necesita una aplicación censal. Por otro lado, si sólo son necesarios 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 resulta 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.

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 sobre-muestreo costoso.

B. ¿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:

B1. Resultados globales, incluyendo matrícula, participación en la evaluación, sin una verdadera evaluación de contenidos.

B2. Éxito o fracaso en general, culminación del plan de estudios, graduación, certificación, tal vez basados en evaluaciones de diferentes asignaturas y otra información.

B3. Puntajes en asignaturas, tales como el logro general en matemáticas o en lenguaje.

B4. 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.

B5. 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.

B6. Estadísticas de respuestas para ítems 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.

B7. Registro detallado de las respuestas a una prueba, incluyendo patrones de distribución de las respuestas a los ítems, 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 todo un plan de estudios pueden ayudar a localizar áreas de éxito general relativamente alto o bajo (e.g.. mejores escuelas o tipos de escuelas). Pero un conocimiento más detallado de la substancia y el contenido de esos logros puede conducir a una evaluación de la importancia y las consecuencias de tales diferencias. Éstas, a su vez, legitimarán medidas tales como la selección, el establecimiento de incentivos u otras intervenciones.

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, podemos llegar a comprender cuáles aspectos de un curriculum son aprendidos exitosamente y podemos 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 ítems requerido para cubrir en detalle un área de contenido es alto. Primero, habrá un número relativamente grande de áreas (sub-contenidos) dentro de un área de contenido. Por ejemplo, en matemáticas tenemos áreas generales como aritmética, geometría, álgebra, etc., y áreas más específicas tales como fracciones, adición de fracciones, formas geométricas, congruencia, series y secuencias, etc. Segundo, cada área o subárea requiere un número suficiente de ítems, 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 ítemes tiene un fundamento sustantivo y otro estadístico. En cuanto a lo sustantivo, necesitamos 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, necesitamos obtener una buena medida del desempeño promedio de los estudiantes, de la variación entre los ítemes y de la interacción entre los ítemes y los estudiantes (algunos estudiantes hacen algunas cosas bien, otros estudiantes hacen otras cosas bien). Las magnitudes de estas variaciones y los tamaños de las muestras de ítemes determinan la "generalizabilidad" de nuestros resultados. Otros aspectos de las habilidades de los estudiantes también deberían ser considerados e introducidos en el diseño y el análisis, tales como las habilidades de los estudiantes para mostrar logros en una variedad de contextos (e.g., problemas matemáticos verbales o cálculos sencillos), modalidades de respuesta (abiertas, opción múltiple), etc.

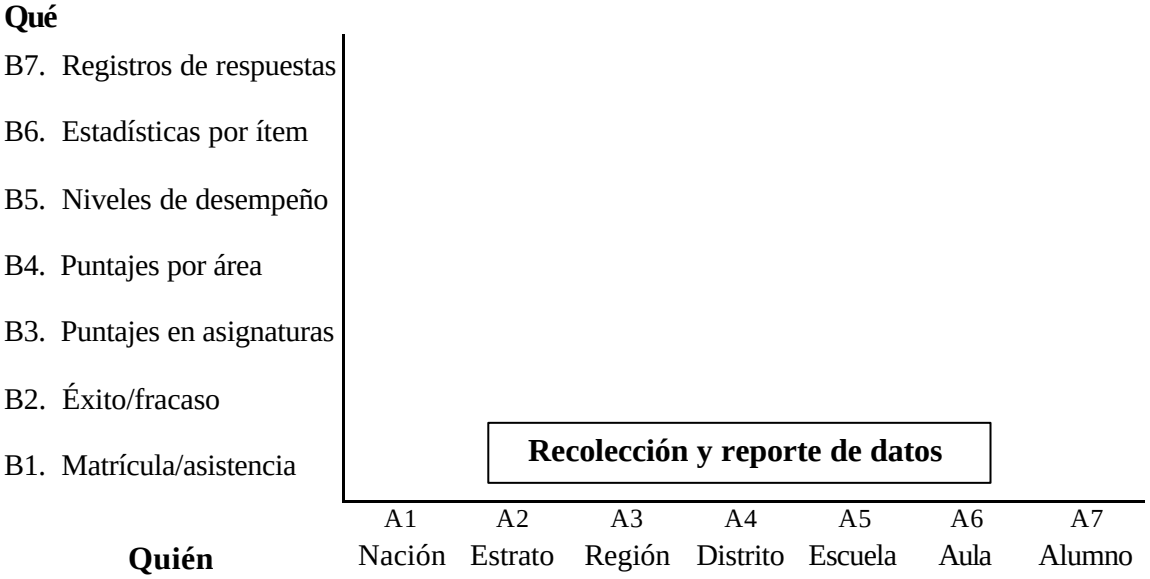
Por otro lado, una muestra relativamente pequeña de contenidos y de ítemes puede ser suficiente si el propósito del reporte es suministrar unos promedios simples que resumen la situación en un dominio de contenidos, aunque ello no constituiría un diagnóstico del curriculum.

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. Esa es la manera como se definen los estándares y como adquieren su significado. Por ejemplo, en matemáticas no deberíamos interesarnos sólo en si los estudiantes pueden sumar o restar, sino más bien queremos saber 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.

C. Tipología de sistemas de evaluación

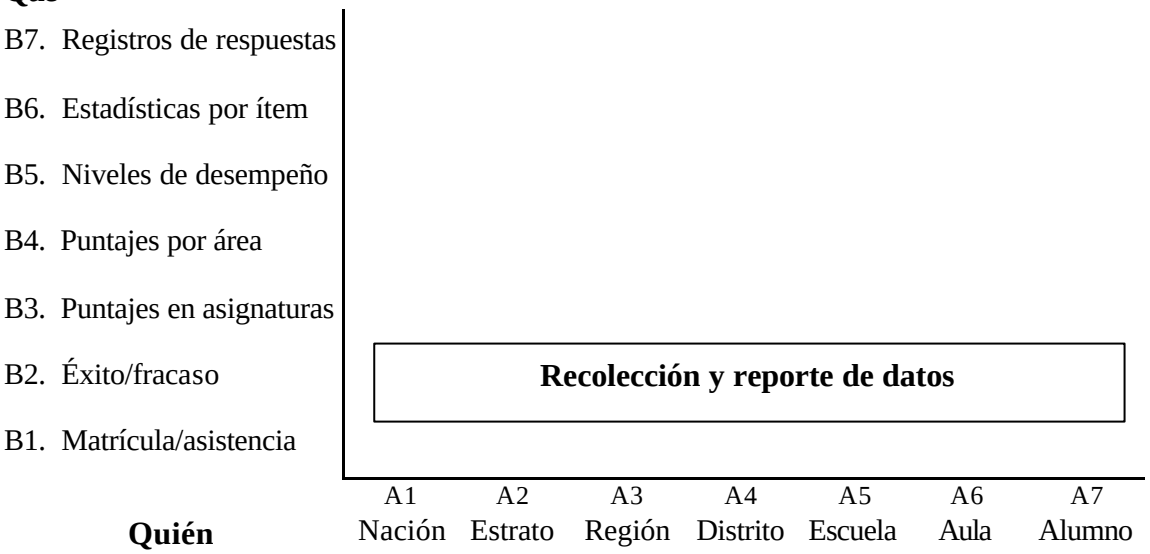
A partir de estas dos dimensiones de la granularidad -- quién es evaluado y qué es evaluado -- podemos definir una tipología de sistemas de evaluación.

C1. Estadísticas educacionales. Una forma simple de sistema de evaluación, común a todos los países, es la recopilación y presentación de estadísticas educacionales. En este caso hay muy poco contenido o diferenciación de contenidos. La mayoría de las estadísticas educacionales trabajan con cantidades de estudiantes, tal vez diferenciados por grado, sexo, repetidor o promovido, etc. Simultáneamente, hay una diferenciación muy fina de la información respecto a dónde se ubican los estudiantes en escuelas, distritos y niveles más altos del sistema educativo. Las estadísticas educacionales 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. Pero no hay ningún contenido que diferenciar.



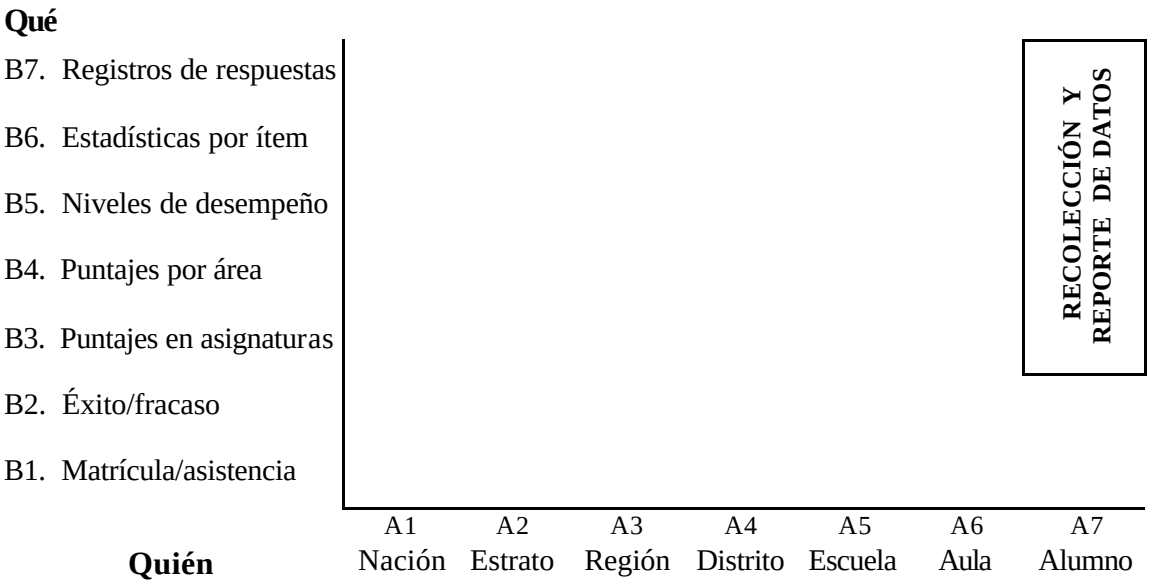
C2. Programas de evaluación y certificación. Desde nuestra perspectiva de las dimensiones de la granularidad, un programa de evaluación y certificación de estudiantes está muy cercano a un sistema de estadística educativa, pero refinado en

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 asimismo realizarse un resumen estadístico a distintos niveles de la jerarquía educacional por requerimiento administrativo.



Los programas de evaluación y certificación son, por su naturaleza, realizados de manera censal en las poblaciones afectadas. Pero a menos que haya un propósito secundario de diagnóstico del éxito o el fracaso estudiantil, o criterios múltiples para la certificación, hay poco interés en la diferenciación del contenido.

C3. Programas de evaluaciones diagnósticas. Muchos sistemas escolares tienen ciertos tipos de programas de pruebas de evaluación que tienen el propósito específico de diagnosticar algunas características educacionales y psicológicas de los individuos (e.g. dislexia). Estos programas 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 apropiada de contenido, relacionada con los tipos de dificultades de aprendizaje que están siendo diagnosticados.



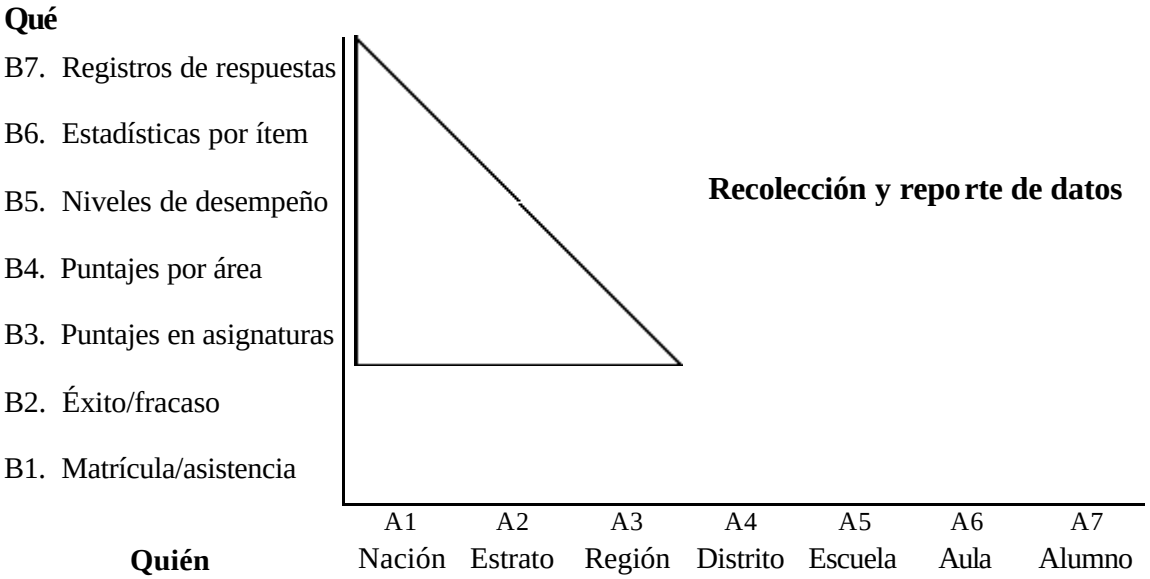
Estos tipos de programas de evaluación o de pruebas generalmente no son considerados evaluaciones nacionales de educación, salvo tal vez en pruebas de despistaje a gran escala, y suelen ser implementados por personal local especializado. Se les menciona en este lugar para destacar que un diagnóstico individual detallado por lo general no es un requerimiento o una posibilidad en un sistema de evaluación a gran escala.

C4. Evaluación nacional por muestreo. Por "evaluación nacional" nos referimos a un tipo de evaluación educacional realizada en diversos países mediante relevamientos muestrales. Un ejemplo temprano es la Evaluación Nacional de Progreso Educativo (NAEP) de los Estados Unidos. Dado que en la misma no se incluye a todos los estudiantes, hay una limitación en el grado de detalle con que pueden desagregarse los datos. Por ejemplo, la NAEP se desagregó originalmente en únicamente cuatro grandes regiones del país y en algunos datos demográficos básicos. Más recientemente ha habido una desagregación hasta el nivel de los estados.

La principal herramienta metodológica de las evaluaciones nacionales es el muestreo de ítemes y de individuos. Lo usual es dividir un conjunto muy grande de ítemes 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 el diagrama para el caso de las evaluaciones nacionales que se presenta más abajo, la recolección y reporte de datos 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 tendremos una precisión adecuada para reportar estadísticas acerca de las respuestas dadas a ese ítem por diferentes poblaciones. En cambio, si se agregan los puntajes correspondientes a varios ítemes, entonces es posible trabajar conjuntamente los resultados de los estudiantes para proporcionar información más detallada sobre unidades más finas. Se

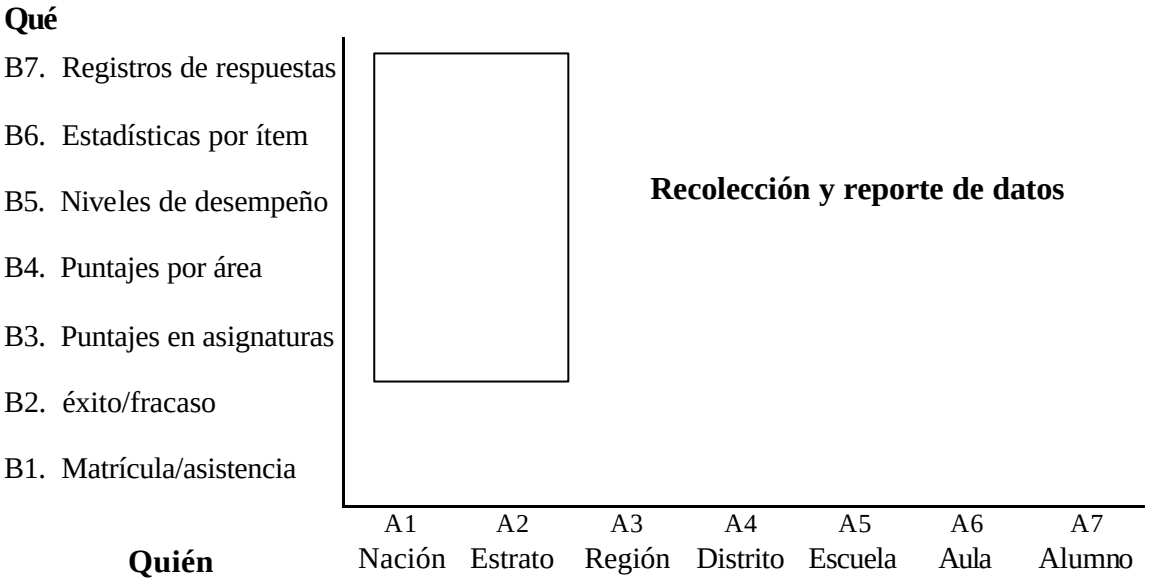
podría incluso calcular puntajes individuales razonables, en base a todo el conjunto de contenidos.



C5. Estudio de investigación curricular mediante muestreo. El propósito de una evaluación investigativa suele ser 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 educacionales y sociales.

En el diagrama relacionado a los estudios de investigación curricular que se encuentra más abajo, 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. Sin embargo, el rectángulo también muestra poco 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. Las variables son importantes pero las 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 educacionales, pero no estamos interesados en informar individualmente sobre las escuelas, clases o estudiantes.



D. El dilema de la granularidad

Nuestro propósito al desplegar en detalle las cuestiones vinculadas a la granularidad -- la determinación acerca de quién y qué es evaluado -- y al 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 educacional.

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 necesitamos obtener medidas detalladas sobre cada municipalidad del país, tendremos probablemente que usar una prueba muy breve y simple, y la cantidad de detalle en los contenidos será mínima. Si, por otro lado, queremos tener una gran profundidad en la medición de un dominio de contenidos como, por ejemplo, información detallada sobre el conocimiento, la comprensión y las habilidades prácticas de los estudiantes en matemáticas, tendremos probablemente que usar rotaciones de ítemes 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 mejora 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 ítemes que provienen de todos el campo de contenidos, además de algunos ítemes 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 ítemes 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 7000 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 estas cuestiones acerca 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 y las 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 el resultado de una evaluación? ¿Se usan los resultados de las evaluaciones de manera apropiada en la toma de decisiones?

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*.

Los y las estudiantes en un sistema educativo participan en un gran número de actividades durante el año escolar, y es común que el éxito con el que enfrentan cada situación de aprendizaje varíe de una ocasión a otra. Es posible que la estrategia óptima para comprender cómo se da (o no da) el logro sea registrar el tipo de éxito que el o la estudiante experimenta al enfrentar cada una de las situaciones que aprendizaje en los que participa al año – lo que, en alguna medida, es lo que docentes comúnmente intentan hacer como parte de su labor de evaluación en el aula.

Por su parte, las autoridades políticas y la sociedad civil tienen interés por tener información acerca del sistema educativo. Este interés obedece a distintas razones, entre las que se pueden citar una preocupación por la calidad de la educación (en muchos países observadores de la educación han sugerido recientemente que los y las estudiantes en América Latina logran poco en la escuela en comparación con los estudiantes de otros países, o con respecto a 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). También hay quienes están preocupados por la equidad en la educación y necesitan descubrir si el sistema educativo favorece en forma desigual a distintos grupos económicos, culturales o lingüísticos y, por cierto, también están aquéllos que 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 y las estudiantes de un país -- o a un número representativo de ellos -- de esa manera.

Por esa razón, los países desarrollan otras estrategias para recolectar información acerca de sus sistemas educacionales (y a menudo, de diversos subsistemas). 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 y las estudiantes, que dura uno o dos períodos lectivos. En esta nueva situación y durante ese lapso de tiempo, el sistema de evaluación pretende que los y las estudiantes demuestren 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 a los estudiantes es una *prueba escrita,* es decir, se hacen preguntas que los y las estudiantes deben de responder en forma escrita.

Las preguntas que se incluyen en las pruebas se encuentran allí porque sus autores consideran que ellas representan bien el tipo de situación que los y las estudiantes deben poder enfrentar con éxito. Es decir, se formulan preguntas que, a criterio de los autores de la prueba, exigen que los estudiantes utilicen lo que aprenden en la escuela para contestarlas correctamente. Por consiguiente, se postula que estas preguntas representan adecuadamente las capacidades adquiridas durante la escolarización. Es así como las pruebas – mediante sus preguntas – pretenden arrojar una representación fiel de las capacidades de los y las estudiantes que se generan en su paso por el sistema educativo.

Ahora bien, ¿cuán fieles son estas representaciones?

Interpretar correctamente y usar apropiadamente la información que nos dan las pruebas significa que debemos preocuparnos por entender el tipo de representación del logro que permiten las mismas. 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. En otras palabras, debemos asegurarnos de entender correctamente qué tipo de información nos dan estos números acerca de las capacidades de estudiantes.

Plantearnos interrogantes sobre una forma apropiada y justificable de interpretar y usar los resultados de nuestras pruebas y encuestas es preocuparnos por lo que se llama en medición la *validez.*

La validez no es una propiedad intrínseca de las pruebas o las 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 validez se refiere a

la calidad de las conclusiones que tomamos a partir de las mediciones y a las consecuencias que las mediciones generan en los procesos que se proponen medir[1]

Algunos ejemplos

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

1. *El Ministerio de Educación se encuentra implementando un nuevo currículum nacional de Matemáticas, cuyo enfoque principal es que los estudiantes aprendan cómo resolver problemas novedosos de la vida real utilizando elementos de razonamiento matemático. Sin embargo, para descubrir qué han logrado los y las estudiantes, 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 y las 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 los estudiantes recurran a este tipo de habilidades para resolverlos.

2. *Se escribe una prueba para descubrir si los estudiantes de educación primaria o básica de 7 años de edad están adquiriendo conocimientos acerca de ciencias*

[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 aquélla 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 hasta la actualidad, 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 Educacional, 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.

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

3. *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 escuelas privadas son más eficaces que las públicas, aun cuando ocurre que éstas no cuentan con textos que aborden uno de los temas más importantes de la prueba.* Aquí, sin duda, 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. ¿Acaso se justifica la interpretación de un bajo promedio como indicador de falta de eficacia del establecimiento? ¿No será más justificado interpretarlo como indicador de una falta de equidad en la distribución de los recursos?

4. *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. Dado este fin, 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 y las 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.

5. *En un país no existe un currículum nacional, sino que cada provincia tiene su propio currículum. La Secretaría de Educación administra una prueba en todas las provincias. Para garantizar que la prueba es justa para todas las provincias, se decide poner sólo preguntas sobre aquellos temas que se enseñan en todas ellas – esto 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?

6. *Se diseña una prueba de lenguaje que entre sus preguntas contiene una sola en la cual los y las 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 y para ello se fija en los resultados de la prueba. Se observa en la prueba 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 mediante los módulos y capacitaciones adquieren 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?

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. Muchos factores pueden afectar el significado que los ministerios u otros usuarios pretenden asignar a los resultados de 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 sustentar con evidencia pertinente el tipo de conclusiones haciendo explícitos de antemano los tipos de uso para los cuales los resultados podrán ser empleados legítimamente, así como los tipos de 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*. La interpretación justa y el uso apropiado de los resultados de las mediciones dependen en gran medida de la solidez del esfuerzo del equipo que diseña las mediciones por asegurar la validación de las mismas.

Opciones para la validación de mediciones en educación

El proceso de validación consiste en acumular evidencia que da sustento o justifica las interpretaciones que se pretende derivar de las pruebas y encuestas. Existe una gran cantidad de opciones en cuando 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 referidas 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 currículum 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 ese tipo de pruebas, se establece (con mayor o menor grado de fundamento) que un promedio determinado *predice* una exitosa carrera universitaria. En el caso de 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.

En América Latina, es poco frecuente que se proporcione documentación acerca de las razones que conducen a las distintas decisiones que se toman en el proceso de construcción de las pruebas. Tampoco es frecuente ofrecer información acerca de los propósitos que se persiguen con respecto a la naturaleza y uso de los resultados, acerca de los grupos entre los cuales fueron validados los instrumentos y sobre las condiciones específicas de la medición. 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 no sólo una definición explícita de las capacidades que interesan sino también de aquéllas que pretendemos 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 de ciencias naturales en la prueba de esa materia.

¿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?, etc.

¿Qué tipos de preguntas permiten comprobar que se dominan las capacidades?

Por ejemplo, si se tiene el objetivo de 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?; ¿se necesitan más bien preguntas que les pidan demostrar los pasos que siguen para resolver problemas o escribir textos?; ¿por qué?. Quizás algunas destrezas o capacidades requieren para ser medidas 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 le va a asignar a cada tipo a la hora de calcular promedios, etc.

¿Cómo se evidencia que lo que predice la prueba ocurre en realidad?

En este sentido, cuando el propósito de una prueba es el de 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 estudiante y 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 convergencia de los resultados. Por otro lado, existen algunas pruebas internacionales comparativas que miden aspectos que también se pretende medir en pruebas de alguna nación o provincia. En estos casos, la participación en estas pruebas internacionales puede servir para propósitos técnicos de validación de las mediciones nacionales. Por otro lado, otra estrategia de validación es contrastar los resultados de una prueba con los resultados de una observación directa a estudiantes o el análisis de sus tareas o 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 al control de ellos?

Es importante describir cómo se asegura que todos los estudiantes estén en igualdad de condiciones para demostrar lo que saben. Es necesario, 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. Si lo que se quiere hacer con la prueba es inferir qué es lo que aprenden o no los y las estudiantes, es muy importante que una contestación errónea represente de verdad la ausencia de un conocimiento y no que se ha interpretado incorrectamente la pregunta, debido a diferencias culturales o regionales en el uso del idioma, por ejemplo. Por otro lado, si se pretende utilizar los resultados de las pruebas para evaluar programas de estudio, opciones pedagógicas o currículum, también 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 contestar preguntas sobre cosas que no les fueron enseñadas en clase. Esto siempre es importante, puesto que existen serios problemas éticos cuando a los estudiantes se les responsabiliza por contenidos que no han tenido la oportunidad de aprender, o cuando a los docentes se les responsabiliza por el logro de sus estudiantes, no habiéndoseles 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, evidencias que en la actualidad muy raramente se reportan. Es perentorio proporcionar estas evidencias y otras que sustenten el contenido y el uso de las pruebas.

Algunas consideraciones finales

Como se estableció anteriormente en la definición formal, la validez es cuestión de *grado*. No existen mediciones perfectamente válidas – mediciones que reproducen fielmente todas aquellas facetas de la realidad educacional que pretenden medir-. Lo que existen son mediciones que son 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. En este sentido, es importante recordar que las responsabilidades con respecto a la validación de las mediciones corresponden tanto a los diseñadores de las mediciones 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 otro lado, 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 la tarea de 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, por ejemplo, permita 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. Es decir, una prueba validada para su uso en un país o en una provincia determinada, no puede ser considerada como validada para el uso con otras poblaciones. Si se desea utilizar el instrumento de medición en una nueva población, compete a quien lo desea utilizar acometer la tarea de su validación para el nuevo contexto. También es necesario tomar 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 transcurrir del tiempo, no atenúen la validez de las mediciones.

La validación es un aspecto central e ineludible del proceso de asegurar que esas mediciones hagan aquello para lo cual fueron diseñadas. Dado que su objetivo es asegurar la congruencia de la medición con la realidad educacional que se supone se está midiendo, se trata de 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 tan solo 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].

[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. La confiabilidad se refiere a tres cosas interrelacionadas. En primer lugar, se refiere a la noción de la estabilidad de la medición. En este sentido, nos preguntamos si las pruebas o encuestas arrojan resultados similoares siempre que se aplican a sujetos similares en condiciones similares. En segundo lugar, se refiere a su nivel de precisión. En este sentido, nos preocupamos por la relación de los resultados de la medición con la "realidad" que mide. En tercer lugar, la confiabilidad se refiere 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. Para un tratamiento detallado de este tema, dirigido a un público no técnico, consultar Moss 1994.

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 auténtica y del desempeño en los sistemas de medición de la región?

Introducción

A los equipos de 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 disyuntiva de desarrollar pruebas de rendimiento para comparar el logro de grupos de estudiantes con otros grupos o para 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, contradictorias. Cuando se quiere comparar el logro de ciertos grupos de estudiantes con los de otros, se puede trabajar dentro del paradigma de medición referida a normas, mientras que cuando se quiere conocer qué conocimientos o competencias específicas logran desarrollar los estudiantes se debe recurrir al paradigma de medición referida a criterios.

Respecto a la disyuntiva planteada, en este capítulo se pretende examinar la realidad de los sistemas de medición de la región. Para ello se ofrece, en primer lugar, un ejemplo de lo que típicamente se encuentra en la región en lo que se refiere al desarrollo y validación de pruebas de conocimiento dentro del paradigma referido a normas. Sobre la base de este ejemplo se hacen observaciones y comentarios con el propósito de señalar limitaciones comúnmente presentes en la tarea del desarrollo y validación de estas pruebas. A continuación, se realiza una revisión de 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 recientemente al empleo de las denominadas pruebas de desempeño y a la evaluación auténtica.

La medición del logro en América Latina: un ejemplo típico

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. Esta aseveración se fundamenta en las descripciones de los procesos de elaboración de las pruebas contenidas en los informes de resultados y otros documentos oficiales de los sistemas. Los elementos del proceso de desarrollo de los instrumentos mencionados en esos documentos incluyen la elaboración de tablas de especificaciones, la producción, aplicación piloto y análisis de ítemes y los reportes de resultados, todos los cuales dan evidencias inequívocas del empleo del modelo de 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 reportados comúnmente, 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, etc...), tienen escaso sentido pedagógico. Este limitado significado pedagógico deriva de la ausencia de información real sobre el logro de conocimientos, destrezas o habilidades específicas de parte de los estudiantes que contienen estos promedios. ¿Qué información de utilidad le comunican los promedios a un maestro de aula que le permita mejorar su trabajo con los niños? ¿Qué utilidad tiene para un curriculista en el Ministerio de Educación conocer el promedio en resolución de problemas de la prueba de matemáticas de quinto grado?

Entre las razones por las cuales se ha recurrido al enfoque referido a normas está la abundancia de experiencia e información internacionalmente disponible sobre los procedimientos que tradicionalmente se han seguido en la elaboración y validación de pruebas dentro del paradigma de pruebas referidas a normas. Además, la limitada formación y capacitación académica en el área de la medición a la cual han podido acceder los funcionarios de los ministerios encargados de desarrollar las pruebas se ha dado sobre los principios de la teoría de pruebas referidas a normas y los procedimientos metodológicos que se sustentan en esos principios.

Otro factor que también ha influido en la selección de este paradigma es la disponibilidad de paquetes estadísticos de computo que permiten realizar análisis de ítemes tradicionales o novedosos y otros análisis técnicos. La existencia, disponibilidad y empleo de estos paquetes produce una sensación de seguridad técnica que se trasluce en muchos de estos informes, aunque muchas veces las interpretaciones que se dan a estos resultados no son del todo correctas o rebasan las posibilidades reales de los análisis.

A continuación se describe un caso que se podría encontrar típicamente en los sistemas de medición del logro en la región. Sobre este caso se harán comentarios y observaciones puntuales referidas a las diferentes etapas de su desarrollo.

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 iniciaron las labores de preparación de pruebas para medir el logro académico para tercero y sexto grados de la educación primaria, en matemáticas y lenguaje. Se definió que las pruebas tendrían como base el curriculum prescrito en esas dos asignaturas y para esos grados. Como primer paso del proceso, se definió una tabla de especificaciones, en la cual los contenidos del currículo se dividieron por áreas y éstas, a su vez, se subdividieron en contenidos de un mayor grado de especificidad. La tabla se balanceó de acuerdo con el nivel taxonómico (ver más adelante) con el cual se quería medir los contenidos especificados y con el número de ítemes con que se quería medir esos contenidos.

Se puede destacar tres hechos relevantes en el ejemplo anterior. Primero, la decisión que generalmente se hace de tomar el currículo prescrito como base de las pruebas. Al hacerlo, se asume que el currículo es conocido y comprendido por todos los maestros. Ellos han recibido la capacitación adecuada para ejecutarlo, tanto en la disciplina como en los elementos didácticos; 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. Sería apropiado tomar esta decisión si el objetivo de la medición es contribuir en la evaluación del currículo implementado, pero si no se incluye una medición de las condiciones de implementación, la evaluación no sería capaz de explicar el fenómeno medido.

Evidentemente, si el objetivo fuese evaluar el currículo real o el enseñado, otra opción sería basar la prueba en una aproximación al currículo enseñado. Una manera de establecer esta aproximación podría ser realizar una consulta a una muestra de maestros sobre los objetivos curriculares (o los contenidos o ambos) que ellos consideran son los fundamentales para un niño que termina el tercero o sexto grado y por lo tanto, los que ellos realmente enseñan y los niños han tenido oportunidad de aprender.

El segundo hecho a destacar es la decisión 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. Esta 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. La intención de dividir el contenido por áreas 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.

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 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 los contenidos de una disciplina es una información fundamental para establecer la validez de la interpretación de los resultados.

El tercer hecho que es conveniente señalar es 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.

> *El paso siguiente es la elaboración de los ítemes para llenar las expectativas establecidas en la tabla de especificaciones. Para cumplir esta tarea, se ha acostumbrado realizar grandes operativos mediante los cuales se ofrecen talleres para la elaboración de preguntas de selección múltiple a maestros en diferentes lugares del país, a los que se les pide redactar ítemes para diferentes contenidos y en distintos niveles taxonómicos. Mediante este procedimiento, se recogen enormes cantidades de ítemes. Se ha justificado esta elaboración masiva de ítemes bajo el argumento que esto 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 en estas tareas o son capacitadas para hacerlo.*

Aquí también, vale la pena hacer un comentario. En primer lugar, la escritura de preguntas de selección o 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. La principal es que tenga un excelente dominio de la materia sobre la que escribirá ítemes. En segundo término debe tener una amplia y probada experiencia en labores de escritura y revisión de ítemes. Conviene aquí insistir en que las preguntas de opción múltiple, si están adecuadamente elaboradas, permiten medir habilidades complejas y que, 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--. La calidad de las preguntas dependerá del conocimiento y experiencia de quien las escribe.

Se debe reconocer que los operativos para que los maestros escriban ítemes pueden tener beneficios políticos para la aceptación del sistema de medición y para capacitar maestros. Sin embargo, debe tenerse claro que, a pesar de las ventajas señaladas, estos operativos no contribuyen en nada a la calidad de la prueba. El argumento que se ofrece, de que de esta manera 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. Además, como se señaló anteriormente, lo que para un

maestro es un ítem de aplicación para otro maestro será de memoria, dada la experiencia de enseñanza que cada uno tiene. Por otra parte, en la práctica se ha comprobado que la inmensa mayoría de los ítemes escritos por maestros en estos operativos, son desechados en la primera revisión.

> *Una vez producidos los ítemes, se someten a una revisión. La complejidad y sistematización de la revisión varía desde una relativamente informal para detectar defectos gruesos en la estructura hasta revisiones hechas con hojas de calificación preparadas con ese propósito expreso o revisiones de relación item-contenido llevadas a cabo por jueces independientes. Estas últimas son casos relativamente raros, pues en la mayoría de las ocasiones las revisiones se dan sobre aspectos formales de estructura y redacción de los ítemes.*

La revisión de los ítemes debería dividirse en dos aspectos. El primero corresponde a una revisión estructural sistemática, hecha por jueces especialistas. Estos jueces tienen que poseer dos características básicas: dominio de la disciplina para la que se escribieron los ítemes y una amplía experiencia en la escritura de ítemes de selección múltiple. Esta revisión requiere que más de un juez revise independientemente los ítemes, empleando un instrumento que sistematice el trabajo, y deje constancia escrita de su opinión y de sus observaciones. Esta revisión puede llevar a reescribir algunos ítemes y a desechar otros.

El segundo aspecto de la revisión de ítemes implica comprobar la relación entre el ítem y el contenido que se supone este mide. Nuevamente, este es un trabajo que requiere la labor de jueces. Estos jueces deben poseer dos rasgos fundamentales en su perfil: tener 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 (alrededor de un 75%) de los jueces tendrá que mostrar acuerdo en la relación de cada ítem con un contenido. La comprobación de la relación del ítem con el contenido que se supone que mide es un elemento central de los procesos de validación

Es fundamental insistir que, en general, en la revisión de los documentos de los sistemas de medición se nota que existe una limitación en el desarrollo y validación de las pruebas empleadas, pues no se le ha puesto atención debida a los procesos de comprobación de la calidad de la estructura de los ítemes y de establecer claramente la relación de cada ítem con el contenido que pretende medir. Esto es particularmente serio si se considera el carácter fundamental de estos dos procesos para establecer la evidencia de validez necesaria para la interpretación y uso del resultado de la medición.

> *Seguidamente, 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 la muestra contenga 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 llevar a cabo análisis estadísticos para conocer índices que caracterizarán a los ítemes. Generalmente estos análisis se ejecutan empleando paquetes estadísticos comerciales(por ejemplo el ITEMAN). Con los parámetros obtenidos (dificultad, correlación ítem-puntaje total de la prueba --discriminación--, frecuencia de respuesta según opción), se realiza una selección de los ítemes que constituirán las pruebas definitivas. Esta selección 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 .40 o correlaciones ítem-puntaje total positivas y significativamente mayores que cero.

Vale la pena señalar dos debilidades básicas observables con frecuencia en la aplicación de la prueba piloto. En primer lugar, casi nunca se establece como objetivo de esta aplicación 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. La recolección requiere de una cuidadosa capacitación de los aplicadores y de la preparación de formas para registrar las opiniones. Generalmente, se podría obtener la información mediante la formulación de una discusión con los estudiantes y una conversación con el maestro.

El segundo punto se refiere a las consecuencias que tienen en los resultados de la aplicación de las pruebas la selección de ítemes basados en los valores de los parámetros establecidos anteriormente (alrededor del 50% de dificultad y una discriminación de 0.4). Es debido a ello que comúnmente los resultados reportados como promedios de los puntajes totales de las pruebas estén 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. Esto es así porque las pruebas están elaboradas y los ítemes son seleccionados para que los resultados estén en el ámbito que resulta. Es un buen ejemplo de la profecía autocumplida.

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.

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).

Resulta conveniente insistir en que, 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. Como se explicó anteriormente, y como se hará más claro en las siguientes secciones de este capítulo, esto significa que se está realizando una interpretación referida a criterios para una prueba referida a normas. 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.

A manera de recapitulación de esta sección, es importante resumir algunos aspectos que se derivan de lo analizado:

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. Un objetivo congruente con el empleo de este paradigma 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.

La medición referida a criterios

Una segunda opción para el desarrollo y validación de pruebas de conocimiento está representada por la medición referida a criterios. Pensamos que esta opción es la más conveniente, fundamentándonos en el hecho de que este paradigma permite obtener información con mucho significado pedagógico. La aplicación de una prueba desarrollada y sometida a un proceso de validación bajo los principios de la medición referida a criterios permite obtener información relevante acerca de los conocimientos, destrezas y habilidades específicas que un grupo de estudiantes logra dominar.

Existen varias diferencias conceptuales entre las pruebas referidas a normas y pruebas referidas a criterios. Estas diferencias provienen del origen psicométrico y edumétrico, respectivamente, de cada uno de los paradigmas. El origen psicométrico está dado por los trabajos de los psicólogos sociales que a finales del siglo XIX y principios del XX emprendieron la tarea de medir las características psicosociales del ser humano. Recién a partir de la década de los sesenta se emprende en educación la

tarea fundamental de la medición de habilidades, destrezas y conocimientos específicos, productos de la escolaridad, es decir, se inicia la edumetría.

Hay dos diferencias fundamentales entre los dos tipos de medición. El paradigma psicométrico tiene la premisa de que los resultados de la medición de cualquier característica humana en una población se comportarán de acuerdo con la curva normal, mientras que el segundo paradigma se fundamenta en el principio de 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. Como consecuencia de esta diferencia, el primero privilegia maximizar la variabilidad y así asegurarse que los resultados de la aplicación de las pruebas se comportan normalmente. En el caso de las pruebas referidas a criterios, la variabilidad no es una característica que importe, por lo tanto no preocupa su valor.

La segunda diferencia estriba en que la medición referida a normas privilegia la comparación entre estudiantes o entre grupos de estudiantes, mientras que la medición referida a criterios privilegia la comparación de los logros de los estudiantes con respecto a las metas de aprendizaje o las competencias que el sistema educativo persigue que los estudiantes alcancen. De estas y otras diferencias se derivan implicaciones para la metodología de desarrollo y validación de las pruebas.

Se pueden señalar varias diferencias metodológicas sustantivas entre los dos paradigmas:

1. La definición de lo que se va medir. En las pruebas referidas a normas ésta 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. Pero en muchos otros casos la definición es aún menos específica, pues proviene o está sustentada en los programas de estudio. Éstos, sobre todo en los años noventa, ofrecen un marco de referencia que es sumamente vago y que da lugar a muy diferentes interpretaciones, pues en muchos casos no ofrecen objetivos siquiera medianamente claros.

Por otra parte, en el paradigma de criterios la definición de lo que se va a medir tiene que ser clara y específica. Consiste en una definición detallada del "dominio del conocimiento" (en la acepción de campo) que abarca el contenido por medir y de las reglas básicas de estructuración de los ítemes con que se va a medir ese dominio. A estas definiciones detalladas se les conoce genéricamente como especificaciones de contenido. Para definir estas especificaciones se emplean diversas técnicas, tales como mapeo de conceptos, algoritmos, objetivos amplificados u objetivos IOX. Algunas de estas técnicas solamente son aplicables a la especificación de los contenidos en una o dos asignaturas, mientras otras tienen una aplicabilidad más generalizada. La selección de la técnica dependerá de varios factores: acceso a la descripción de los pasos necesarios para su aplicación, posibilidad de capacitación del personal en su uso, la generalidad que pueda tener su uso en varias asignaturas y la capacidad de la técnica de cumplir sus dos funciones esenciales: definir el dominio y establecer reglas de estructuración de los ítemes. Es primordial que los ítemes que midan una especificación

estén realmente midiéndola de la misma manera, por lo tanto, deberán ser ítemes muy similares en el contenido que miden y en la forma en que lo hacen.

2. La definición de prueba. En el paradigma normativo, la prueba se define como el conjunto de ítemes que forman una muestra representativa de todos los conocimientos, destrezas y habilidades que se quieren medir con la prueba. El criterio de representatividad puede ser muy variado. Algunos criterios empleados para definir esa representatividad son: el tiempo que se supone se invierte en enseñar cada conocimiento, habilidad o destreza, el nivel taxonómico con que se quiere medir cada conocimiento, habilidad o destreza y la importancia relativa de esos elementos según el criterio de expertos.

En la medición referida a criterios se define una prueba como 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, con una prueba referida a criterios se mide un solo conocimiento, habilidad o destreza. En este caso, un folleto o cuadernillo de prueba estará constituido en realidad por los ítemes pertenecientes a varias pruebas, tantas como especificaciones se estén midiendo.

3. La evidencia de validez. Los procedimientos para obtener evidencia acerca de si la prueba se ajusta a los principios fundamentales de cada uno de los paradigmas y, consecuentemente, a la definición dada anteriormente, difieren sustancialmente.

En el caso de las pruebas *referidas a normas*, 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, es decir la evaluación independiente que hacen de ello los miembros de un grupo de jueces 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.

Para las pruebas *referidas a criterios*, las evidencias de validez serán: (a) la certeza de que los ítemes están construidos de acuerdo con las características propias de las preguntas de selección múltiple –más adelante en el texto se efectuará una referencia a las pruebas de respuesta abierta o de desempeño-, y (b) un valor aceptable en el índice de congruencia entre cada ítem y su especificación de contenido. El proceso de establecer la congruencia entre el ítem y su especificación de contenido requiere el empleo de jueces independientes que, empleando un instrumento diseñado para tal efecto, registren su opinión sobre esa congruencia. Las opiniones de todos los jueces se combinan mediante una fórmula estadística para producir un índice de congruencia para cada ítem. Este índice se convierte en el parámetro principal para decidir cuales de los ítemes construidos para medir una especificación de contenido pasarán a formar parte del banco de ítemes de esa especificación. Entre los ítemes del banco se seleccionarán aleatoriamente aquéllos que se emplearán en la prueba definitiva.

4. Los procedimientos de cálculo de la confiabilidad y de análisis de ítemes. En el caso de las pruebas referidas a normas, tanto los modelos estadísticos tradicionales como los nuevos que se emplean para el cálculo de la confiabilidad y para el análisis de

los ítemes se fundamentan en la maximización de la variabilidad. Maximizar la variabilidad de dos variables que están correlacionadas 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, sin importar con cuál de los tres modelos (estabilidad, equivalencia o consistencia interna) ésta sea calculada.

Los métodos de cálculo para la confiabilidad son algo distintos en las pruebas referidas a criterios. Para medirla, se suele dar mayor importancia a la consistencia que muestran 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 ser la proporción de estudiantes cuyas respuestas en ambas oportunidades demuestren que sí (o no) domina una competencia o especificación de contenidos. También se han propuesto otros modelos para analizar la confiabilidad que son teóricamente similares a los que se aplican en las pruebas referidas a normas, pero han sido criticados porque resultan inconsistentes con el principio fundamental de que asumir una distribución normal de las respuestas no es ni debe ser un requisito de las pruebas referidas a criterios.

En lo que respecta al análisis de ítemes en las pruebas referidas a normas, la maximización de la variabilidad es también importante cuando se calcula la discriminación o el índice de correlación punto-biserial entre el ítem y el puntaje total de la prueba (dos técnicas frecuentemente empleadas en el análisis de ítemes). Los parámetros resultantes del análisis de ítemes, principalmente la dificultad y la discriminación o el índice de correlación ítem- puntaje total de la prueba, se emplean como indicadores fundamentales para la selección de los ítemes que pasarán a formar parte del banco de ítemes del cual luego se seleccionarán aquéllos que constituirán la prueba definitiva. Desde la perspectiva clásica, la selección se hace escogiendo aquéllos que tienen una dificultad de cerca de un 50% y un índice de discriminación superior a .40 o una correlación punto-biseral positiva y significativamente diferente de cero. Además, los ítemes seleccionados deben cumplir con las especificaciones que señala la tabla de especificaciones. .

Para el análisis de ítemes en el paradigma de criterios, se evalúa su dificultad empleando el mismo cálculo que en las pruebas referidas a normas. Para estimar la capacidad de discriminación de los ítemes se han propuesto más de 17 índices distintos, la mayoría de los cuales requieren comparar la respuesta al ítem dada por estudiantes a los cuales se les aplica la prueba en dos ocasiones, antes y después de la enseñanza, o comparar la respuesta dada por grupos de estudiantes instruidos y no instruidos en la competencia o el contenido especificados. Como puede imaginarse, esto puede ser un requisito difícil y costoso de satisfacer. Uno de los índices, sin embargo, no requiere un procedimiento de esa naturaleza, pudiendo calcularse con una sola aplicación de la prueba. Consiste en comparar el rendimiento en un ítem entre estudiantes que dominan el contenido medido por la prueba y los que no lo hacen.

En las pruebas referidas a criterios, la selección de ítemes 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, en un segundo, considerando la

discriminación y la dificultad. Ésta es otra diferencia importante con los procedimientos seguidos en las pruebas referidas a normas, descritos anteriormente.

5. La interpretación de los resultados de la medición. En este capítulo ya se estableció y se explicó esta diferencia. En el paradigma de normas la interpretación del puntaje de la prueba es relativa, en criterios es absoluta. En el primero, 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). En el segundo, 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. Pueden mencionarse Otros aspectos en que difieren los paradigmas se refieren a:

(a) La necesidad de establecer un puntaje de corte en una prueba referida a criterios. Por puntaje de corte se entiende el puntaje que define si un estudiante domina o no la especificación de contenido medida por la prueba. Existen en la literatura diversos procedimientos para definirlo. Este no es un requisito de una prueba referida a normas.

(b) El número de preguntas que constituyen una prueba. En las pruebas referidas a normas el número 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. En las pruebas referidas a criterios 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 ítemes varía entre tres y cinco, mientras que decisiones sumativas e individuales requieren entre ocho y diez ítemes.

Ejemplo de una prueba referida a criterios desarrollada en América Latina.

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. Con el ejemplo que se ofrece a continuación se ilustra el tipo de procedimientos que se aplican en el marco de este enfoque.

> *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 los estudiantes que terminan ese nivel educativo. Ante la necesidad de obtener esta información específica sobre el logro de los objetivos fundamentales del currículo, el Departamento de Pruebas Nacionales(DPN) del Ministerio decide elaborar y someter a 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? ¿Quiénes son las personas más indicadas para realizar la selección de los objetivos fundamentales? Para proceder a 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, que resultó en 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. Una vez terminada esta etapa de selección, se les solicitó realizar la priorización de 20 objetivos entre los 30 seleccionados. Mediante un procedimiento estadístico se definieron los 20 objetivos que tuvieron las más altas prioridades promedio. Los 20 objetivos así seleccionados constituyeron los objetivos fundamentales de cada una de las asignaturas y para cada uno de estos objetivos se desarrollaron especificaciones de contenido.

La siguiente decisión que tomó el DPN fue la selección de 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 realizarla 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 ítemes 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 ítemes.

La fase siguiente fue la escritura de los ítemes. En primer lugar, se decidió que se escribirían 12 ítems para cada objetivo amplificado. En segundo lugar, la escritura de los ítemes estuvo a cargo de los pares de especialistas por asignatura y de dos personas más seleccionadas entre el grupo de escritores de ítemes, 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 ítemes de selección múltiple .

Una vez escritos los 12 ítemes por objetivo amplificado se llevaron a cabo dos procesos de vital importancia. Uno fue la revisión estructural de los ítemes. Con este propósito se contrataron dos personas en cada asignatura. Estos profesionales tenían la doble característica de poseer una gran experiencia en la preparación y revisión de ítemes 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 ítemes que ellos debieron examinar. Como producto de esta revisión se modificaron algunos ítemes y se desecharon unos pocos.

El segundo proceso fue establecer el índice de la congruencia de cada uno de los ítemes con su objetivo amplificado. Para realizar esta tarea se contrataron 10 jueces por asignatura. Estos jueces compartían las características de tener un probado dominio de la materia y de tener 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 item. 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 cuardenillos 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-subvencionadas). La meta que se tuvo fue 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 estudiantes 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, 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.

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. Enfatiza 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. Emplea una variedad de medios que requieren un tiempo sustancial de parte del estudiante para completarlos. 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 de mundo real.

Los medios que emplea la evaluación del desempeño han sido en el pasado usados por los educadores en las aulas: preguntas de respuesta o final abierto, requerimientos de producir un ensayo, resolver problemas, producir materiales o discursos para exhibición pública, producir artefactos y documentos, y producir portafolios o muestras de trabajos realizados a lo largo de un período. Lo realmente novedoso del enfoque radica, por lo tanto, en el énfasis en la medición de conocimientos y habilidades complejas, tal y como se le presentan al estudiante en la vida real.

Es necesario, sin embargo, tomar en cuenta dos características de este enfoque cuando se aplican a muestras masivas de estudiantes: el tiempo que requiere un estudiante para completar una tarea específica (algunas veces todo un curso, como es el

caso de los portafolios) y la limitación técnica de depender del juicio humano para juzgar la calidad de la ejecución de la tarea. Este último 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. 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 de medición, dado el poco poder de generalización del resultado, que es producto del empleo de una única tarea para medir un conocimiento o habilidad compleja. La decisión del logro que se hace sobre la base de una respuesta a una única tarea tiene poca validez.

Si se desea utilizar algunos de los instrumentos de la medición del desempeño en los sistemas de medición del logro en nuestra región, es necesario tomar en cuenta el factor costo. Esto se convierte en una seria limitante para el empleo de este tipo de mediciones, 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 aun más al agregar el pago del personal que califica las tareas, su capacitación y el necesario control que se deberá ejercer para que se mantengan los niveles mínimos de la confiabilidad de la calificación. Estos costos pueden llegar a ser tan altos que se convierten en prohibitivos para la realidad de las economías de nuestros países.

Sin embargo, resulta indudable que es necesario medir el conocimiento y las habilidades complejas y resulta atractiva la idea de aplicar algunos de los medios de la evaluación del desempeño en las pruebas de los sistemas de medición del logro de nuestra región. 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 constituye 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. Resultaría por lo tanto 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?

En la mayor parte de los países latinoamericanos se aplica, junto con las pruebas de logro, cuestionarios dirigidos a recoger información acerca 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. En casos excepcionales, se suele ofrecer información sobre las variables relativas a las familias, es decir, externas a los sistemas educativos, más que sobre las variables escolares sobre las cuales los Ministerios pueden tomar decisiones. Pocos países han desarrollado trabajos sistemáticos de investigación acerca de los factores escolares que explican, generan o están asociados con las diferencias de resultados de aprendizaje entre las escuelas.

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/privado). Ello permite una primera identificación de las disparidades de logro educativo dentro de un país y brindar información a quienes, en distintos niveles, son responsables de la conducción del sistema educativo.

La no utilización de la información "de contexto" plantea al menos dos grandes problemas:

1. la ausencia de información sobre las características socioculturales de las poblaciones a las que enseñan los distintos sectores del sistemas educativo puede llevar a conclusiones erróneas respecto a la eficacia educativa de dichos sectores;

2. la ausencia de información sobre factores estrictamente escolares puede llevar a la conclusión, también errónea, de que finalmente los resultados educativos dependen exclusivamente del entorno social y el sistema educativo no tiene nada que hacer al respecto.

a. El problema de la falta de contextualización sociocultural de la información sobre resultados de las pruebas

En relación al primer aspecto, es preciso señalar que 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 aparecerán como menos eficaces aquellos sectores del sistema educativo que atienden a la población con mayores carencias, al tiempo que aparecerán 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.

> *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 públicas se analizan controlando las variables socioculturales, es decir, cuando se analizan las diferencias entre escuelas públicas y privadas que atienden al mismo tipo de población, las diferencias en la proporción de alumnos que logran un nivel satisfactorio en el conjunto de la prueba se reducen al entorno de 5 puntos porcentuales y, en algunos sectores sociales, son favorables a las escuelas públicas.*

Algo similar a lo ejemplificado en el párrafo anterior ocurre con la presentación de los datos en función de agregaciones político/geográficas. Cuando se presentan los resultados desagregados por provincia, estado, región o departamento, sin ningún tipo de información adicional, la conclusión inmediata para el lector no especializado es que sin duda 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.

Asimismo, 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 escuelas ubicadas en 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 e ignorar la diversidad de situaciones existentes al interior del mundo urbano. 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.

A partir de lo antedicho, parece necesario dedicar tiempo a la reflexión acerca de cómo establecer formas relevantes de caracterización sociocultural de los niveles de desagregación de la información, de modo que la información brindada por el sistema de evaluación permita hacer comparaciones entre establecimientos, departamentos o provincias que atienden poblaciones con algún grado de similitud.

En este sentido, parece recomendable intentar utilizar la información social que suele recogerse 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. La "efectividad" de una escuela no puede medirse únicamente en términos absolutos sino en relación al tipo de alumnado con que trabaja, ya que éste que implica restricciones o ventajas en relación a los resultados posibles.

En ese sentido, parece necesario 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. En Francia por ejemplo uno de los indicadores de efectividad de los liceos es la comparación entre la tasa de aprobación de la prueba de Bachillerato y la tasa anticipada de acuerdo a la relación existente a nivel nacional entre origen social y aprobación.

El principal supuesto de este enfoque es que al trabajar con resultados absolutos se confunden los efectos propios de las escuelas o jurisdicciones con los efectos de la selección de alumnos con que cada una de ellas trabaja. Los mejores niveles de logro en ciertas escuelas o jurisdicciones pueden no obedecer a que la enseñanza sea mejor en las mismas, sino simplemente a que enseñan a los mejores estudiantes.

Finalmente, es preciso señalar que en algunos países es necesario aún trabajar mucho en el diseño de instrumentos que recojan información sociocultural de base que sea relevante, dado que sus 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. Ello implica la necesidad de mejorar los instrumentos y procedimientos que se emplean para medir los aspectos relativos al contexto social.

b. Los problemas relacionados con la información sobre las características de las escuelas y la enseñanza

En relación al segundo de los problemas seleccionados es preciso señalar que 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, años de experiencia del maestro, etc. 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, en el corto plazo, susceptibles de ser modificados 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, en distintas partes del mundo y en algunos países de la región se ha mostrado que, al interior de una misma categoría social, existen diferencias en los niveles de logro de las escuelas, 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. Una primera dificultad parece ser la sobreabundancia de información que al respecto se recoge, que dificulta la selección de una estrategia para su presentación. Se recoge información sobre una gran cantidad de variables, por lo general sin un plan de análisis y difusión previo, que luego hace sumamente difícil decidir cómo organizar la información y cómo vincularla con los datos sobre aprendizaje. Una segunda dificultad, que en cierto modo explica la anterior, radica en el proceso en cierto modo a-sistemático a través del cual se construyen estos instrumentos. Mientras en el proceso de construcción de pruebas en todos los países se sigue algún tipo de proceso de validación, se establece un referente conceptual, se pilotean las pruebas y se realiza una selección de los reactivos a aplicar, en el caso de los instrumentos complementarios casi nada de ello parece ocurrir. 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 rendir y lo que no.

Por ejemplo, normalmente se formula un conjunto de preguntas relacionadas con el clima institucional o con la existencia de objetivos compartidos en la escuela. Recién una vez recogidos los datos se intenta construir un índice o factor a partir de los mismos. La construcción de índices mediante procedimientos estadísticos no debiera sustituir el proceso de construcción de un modelo conceptual que permita definir y otorgar sentido a un conjunto de variables básicas.

En este sentido, parece necesario avanzar en el desarrollo de una metodología para el diseño de los instrumentos de relevamiento de "factores escolares" que incluya:

a. el desarrollo de 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;

b. 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, por un lado, aquellos aspectos sobre los cuales es posible la toma de decisiones tanto desde la política educativa como desde el interior de los establecimientos y, por otro lado, las limitaciones propias de los cuestionarios autoadministrados-;

c. el mejoramiento de los 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 curriculum implementado, los enfoques didácticos, los tipos de actividades realizadas por los niños, la utilización de los materiales y textos además de su mera existencia, etc.-;

d. el pilotaje y validación previa de los instrumentos.

En particular, parece necesario avanzar 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.

Simultáneamente, sería interesante dedicar cierta energía a la publicación de 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 sería de enorme utilidad porque normalmente este tipo de información no existe en los países, contribuiría 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 que permita afinar el tipo de preguntas que es útil formular, aligerar los cuestionarios o ir pasando en evaluaciones sucesivas a indagar nuevos aspectos.

Finalmente, otra tarea a encarar sería el diseño de mejores formas de reportar los resultados de las pruebas, junto con la información sobre contextos sociales y factores escolares. ¿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-?.

c. Los problemas relacionados con la 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 propiamente dicha sobre el modo en que los diversos factores inciden sobre los aprendizajes. Si bien existe abundante investigación en este terreno en los países desarrollados, el modo en que estos factores afectan el aprendizaje está íntimamente relacionado con los contextos nacionales y, aun al interior de un país, seguramente se producen variaciones en función del tipo de escuela y el tipo de población a la que atiende. Es por ello que tiene sentido realizar esfuerzos para acumular conocimiento a nivel nacional sobre los factores escolares que están asociados con los aprendizajes de los niños.

Sin embargo, un primer problema central en este terreno es que normalmente los países han estado evaluando 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 del año escolar y al final del mismo. Sólo por esta vía es posible contar con información sobre lo que efectivamente los alumnos de un grupo aprendieron durante el año, lo que potenciaría la posibilidad de establecer relaciones entre lo que ocurrió en la escuela y el aula ese año y el avance de los alumnos en términos de aprendizaje. Nuevamente, éste es un claro ejemplo de cómo el diseño del sistema de evaluación puede servir para ciertos fines más que para otros.

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. Uno de ellos sería realizar la medición inicial en muestras más pequeñas. Para un trabajo con fines de investigación no es necesario realizar las mediciones en grandes muestras, que por lo general tienen como finalidad permitir la devolución de información a diversos niveles

de desagregación. Otro camino posible es, en países que evalúan grados sucesivos –por ejemplo, 5to. y 6to. 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.

Un segundo problema central radica en que 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.

Un tercer aspecto o problema que resulta relevante señalar es el relativo a las limitaciones propias de este tipo de estudios, dado que muchas veces existe un exceso de expectativas respecto a la validez del conocimiento construido. A través de la investigación de corte estadístico, aun la más sofisticada, se puede construir ciertos tipos de conocimiento y de información, pero muchas veces, para avanzar en nuestra comprensión de los fenómenos, es necesario desarrollar también otro tipo de 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. En estos casos, 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 –por ejemplo, las relativas a las prácticas de enseñanza- de manera adecuada y rigurosa, lo que no puede hacerse cuando se trabaja a escala masiva[3].

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 al mismo, lo que no implica que no sean relevantes en la producción de los resultados.

> *Por ejemplo, la formación de los maestros se mide normalmente a través de la cantidad de años de estudio. Ésta puede tener escasa variación al interior de un país. Como normalmente será difícil medir la calidad de esa formación, el investigador utilizará la cantidad de años como medida de la formación. Dado que el comportamiento de la variable es homogéneo, no ingresará al modelo, y la conclusión será que no es un factor relevante para la efectividad de la enseñanza. Algo similar, pero en sentido inverso, puede ocurrir con el material didáctico. Como las dotaciones del mismo presentan una amplia variación –que además normalmente estará vinculada al nivel socioeconómico del alumnado de*

[3] De todos modos, también en estos casos, la medición de niveles de logro a gran escala ofrece un marco en el cuál apreciar el impacto efectivo de las intervenciones pedagógicas en el aprendizaje de los alumnos.

las escuelas- y son fácilmente medibles, está variable será más proclive a ingresar a los modelos. Por tanto, el investigador recomendará al Ministro invertir en material didáctico y no en formación de maestros.

El ejemplo anterior pretende alertar respecto al uso poco reflexivo de los resultados de los análisis estadísticos y recordar que la sofisticación de los métodos no garantiza por sí misma la validez de las conclusiones.

Asimismo, es necesario preguntarse acerca de qué tipo de decisiones de política educativa es posible tomar a partir de los resultados de un análisis estadístico. Muchas veces parecen existir expectativas excesivas al respecto. 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.

Lo expresado en los párrafos anteriores no debe ser leído en el sentido de descalificar este tipo de investigación, sino en el de tener modestia respecto a lo que pueden aportar. Sin duda las mediciones de aprendizajes y factores asociados contribuyen a iluminar zonas del escenario educativo, a generar conciencia sobre ciertos problemas, a desmitificar soluciones mágicas, a percibir elementos que están presentes en las escuelas con mejores resultados, en fin, a una acumulación de conocimiento que juega un rol fundamental en el momento de delinear las políticas educativas.

En este sentido la realización de "estudios de casos" relevantes, a través de los cuales se observe en detalle y se describa el modo de enseñar de escuelas y maestros cuyos alumnos alcanzan elevados niveles de logro, parece un camino complementario que es necesario recorrer. Probablemente este tipo de estudios permita construir un conocimiento más fácilmente comunicable a los maestros y a otros usuarios en términos que les sean significativos y útiles y que pueda tener impactos importantes en el mejoramiento de las prácticas de enseñanza. Por otra parte, hay áreas específicas de las prácticas de enseñanza y de la vida escolar cuyo análisis requiere de la observación directa de lo que allí acontece.

Con tal finalidad, las bases de datos generadas por los sistemas nacionales de medición constituyen una fuente de información formidable para identificar instituciones relevantes para la realización de este tipo de estudios.

En virtud de todo lo antedicho, parece pertinente plantearse la necesidad de propiciar el establecimiento de asociaciones y convenios de colaboración entre las Unidades de Medición y centros de investigación especializados, de modo de potenciar el aprovechamiento de las bases de datos existentes mediante la realización de trabajos que las Unidades no logran llevar adelante. En principio parece difícil que las Unidades de Evaluación puedan desarrollar efectivamente todas las tareas: diseño de instrumentos, organización de operativos de evaluación, procesamiento de datos y

producción de informes, capacitación a partir de los resultados e investigación sistemática. Al mismo tiempo, no parece razonable subutilizar la información disponible y las posibilidades que brinda a la investigación el hecho de contar con sistemas regulares de medición en funcionamiento.

En este sentido, el establecimiento de asociaciones con institutos de investigación, que aprovechen las bases de datos existentes para la realización de trabajos más sofisticados y que, simultáneamente colaboren en el mejoramiento de los instrumentos de medición y de la calidad de las bases de datos, parece un camino que es necesario empezar a recorrer. 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, sobre los que en la región existe escasa acumulación de experiencia, aún en las instituciones dedicadas a la investigación educativa. En ese sentido, será necesario 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 inclusive al interior de 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 ítemes, porcentaje de respuestas correctas a ítemes individuales, porcentajes de alumnos que alcanzan cierto puntaje en una prueba, etc.? ¿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?

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

Las cuestiones sobre "escalas de reporte" giran en torno a las maneras en que se registran, agregan (o desagregan) y presentan los logros escolares en los informes. Estas cuestiones están en parte atadas a las de la granularidad que fueron tratadas en el primer capítulo -- por ejemplo, si las estadísticas sobre los ítemes 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 --. Pero también incluyen cuestiones importantes referidas a la representación cualitativa y numérica de los resultados, tales como la presentación de ítemes que realmente fueron aplicados en las pruebas, la provisión de modelos ejemplares del trabajo de los alumnos, porcentajes simples para categorías de respuesta (e.g., el índice de dificultad del ítem), o cualquiera de las diversas maneras posibles de resumir, escalar y mostrar las distribuciones del logro escolar. Finalmente y tal vez de la mayor importancia, la cuestión involucra además el tema de la consistencia y comparabilidad de las escalas de reporte a través de distintos contenidos y a lo largo del tiempo.

Los diferentes sistemas nacionales de evaluación usan una variedad de métodos y prácticas de escalas de reporte de resultados. Por "escala de reporte" nos referimos en primera instancia a qué tipos de números se usan para presentar resultados: números o frecuencias simples, porcentajes, percentiles, puntajes en escalas, etc. Pero además, una interpretación más profunda de "escala" involucra las maneras en que se registran, procesan, transforman, agregan y presentan los datos cualitativos y cuantitativos sobre las respuestas dadas por los alumnos. Estos métodos tienen fundamentos teóricos implícitos o explícitos que, dependiendo de cuán bien correspondan con el verdadero aprendizaje de contenidos y su medición, harán que los reportes sean más o menos significativos e interpretables.

Los reportes sobre logros educacionales se han derivado históricamente 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 ítemes mediante el cual se eligió aquéllos que tenían aproximadamente 50% de dificultad, es decir, ítemes que fueron respondidos correctamente por alrededor de la mitad de los alumnos a los que se aplicó la prueba piloto (véase al respecto el capítulo sobre pruebas referidas a normas y pruebas referidas a criterios).

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. Uno está basado en el análisis de los dominios de los contenidos y de las pruebas, y la determinación de la "generalizabilidad" de los puntajes y porcentajes; esta es una extensión de la teoría clásica sobre la confiabilidad. Un segundo tipo de solución utiliza métodos estadísticos para igualar y calibrar pruebas diferentes. Un tercer tipo (que es un caso especial de esos métodos estadísticos) ha comprendido el desarrollo de teorías de respuesta al ítem (IRT) que intentan colocar a los ítemes y a los estudiantes en dimensiones o escalas latentes. Todos estos enfoques tienen ventajas sustanciales pero, al mismo tiempo, presentan peligros sustanciales.

Enfoques básicos de los reportes

Antes de abordar las técnicas, necesitamos considerar las maneras en que se reportan los datos de las evaluaciones y cómo ello depende del nivel de agregación (granularidad).

Respecto a los ítemes individuales, cuando queremos 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 ítems 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.*

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:

1. *Se puede simplemente dar a conocer las estadísticas de ítemes, incluidas las distribuciones de las respuestas correctas y las distribuciones de las respuestas 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 representantivo de un aspecto importante del aprendizaje de la asignatura, o del conjunto de ítemes como representantivo de una muestra amplia de una sub-área, así como de la interpretabilidad de la calificación misma.*

2. *Es posible informar sobre promedios entre ítems 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. *Se puede dar un paso más, llegando a medidas compuestas de todo un dominio de contenido o incluso abarcando varios dominios. La validez en este caso depende de cuán adecuados son los pesos explícitos o implícitos asignados a los subdominios en la construcción del puntaje global.*

Existen diversas alternativas de métricas (escalas numéricas) para reportar los resultados de las pruebas.

1. *Se puede 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. En ambos casos, las interpretaciones se limitan a las muestras específicas (y los universos correspondientes) de ítemes y de estudiantes.*

2. *Con un sistema de análisis IRT, se pueden 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 losestudiantes, o puede cambiarse la muestra de estudiantes sin cambiar los valores de la escala para los ítemes. Se dice que el sistema IRT es "independiente de la muestra", o sea que no es sensible al efecto del muestreso de ítemes o estudiantes. Esta característica idea depende, por supuesto, de la adecuación del modelo IRT para un determinado proceso de medición, universo de ítemes y universo de estudiantes.

3. *Es posible 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. *Es posible 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.).

Además de esto, hay otras comparaciones importantes que pueden y suelen aparecer en los reportes de las pruebas nacionales o que pueden y suelen ser inferidas – aunque no estén explícitamente en los reportes – por los que leen y tratan de interpretar dichos reportes, a menos que los textos de los reportes adviertan explícitamente sobre posibles equívocos al respecto:

1. *A lo largo de las áreas de contenido. El lector se preguntará, por ejemplo, si el desempeño en matemáticas es más alto que el desempeño en lenguaje.*

2. *A través de los grados. El desempeño en matemáticas de los alumnos de sexto grado, ¿es relativamente más alto, en algún sentido significativo, que el desempeño de los alumnos de tercer grado? En otras palabras, ¿los puntajes que los estudiantes de tercer grado obtienen en una prueba para tercer grado son comparables de alguna manera con los puntajes obtenidos por los estudiantes de sexto grado en una prueba para sexto grado?.*

3. *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?*

La verdad es que muchos investigadores y psicometristas dirán que el primer tipo de comparación, entre contenidos, es casi imposible de hacer. Por su naturaleza tendría que hacerse con pruebas distintas, hechas con ítemes muestreados de diferentes universos. ¿Cómo podría determinarse que un aspecto del aprendizaje de las matemáticas y otro de lenguaje plantean la misma dificultad cognitiva para su realización, requieren el mismo esfuerzo pedagógico para ser aprendidos o merecen que su adquisición tenga el mismo valor o reconocimiento social?

También hay problemas similares en el caso de comparaciones entre grados escolares. Por ejemplo, consideremos el problema de establecer criterios para, digamos, matemáticas en tercer grado y otros criterios para matemáticas en sexto grado, y luego determinar qué porcentaje de estudiantes de los respectivos grados logran cumplir esos criterios. ¿Quién puede decir que los rendimientos corresponden a un nivel cognitivo equivaleente, a una dificultad de aprendizaje comparable, o un valor social similar? Sería dificilísimo evaluar la correspondencia porque los criterios serán diferentes (por ejemplo, sumas de enteros para tercer grado, sumas de fracciones para el secto grado).

Consideramos altamente recomendable que en los reportes de las evaluaciones educacionales se enfatice fuertemente la dificultad conceptual de hacer comparaciones directas entre contenidos o entre grados. El análisis debería centrarse dentro de un contenido y/o dentro de un grado, describiéndose los logros y su distribución entre estudiantes y comparándose estos logros con las expectativas o estándares. Es necesario aplicar esa misma cautela también en los aspectos más finos del análisis de resultados. Por ejemplo, sería difícil justificar una afirmación de que a los estudiantes les va mejor en álgebra que en geometría.

Sin embargo, el tercer tipo de 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. Cuando se hace mal y los cambios reportados son meramente consecuencia de errores técnicos de muestreo o de calibración de las pruebas, la información sobre cambios a lo largo del tiempo será casi aleatoria y sólo añadirá más caos y confusión a la planificación educacional.

Generalizabilidad

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 ítemes 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 ítemes particulares y de las condiciones de aplicación, y qué parte de la variación de dichos puntajes obedece a características de los ítemes 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 ítemes 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 ítemes 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 ítemes, 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 ítemes 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 y ello permite calcular la precisión de las mediciones obtenidas y, especialmente, de las comparaciones que es posible hacer a lo largo del tiempo.

Comparabilidad

Por lo general no es factible usar una misma prueba en dos momentos diferentes en el tiempo. Pero hay métodos psicométricos formales y procedimientos estadísticos 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 ítemes a dos formas de pruebas.

Cuando las formas de las pruebas no son estrictamente paralelas sino que tienen una superposición sustancial, es decir, ítemes 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 la pruebas. Uno no puede desarrollar pruebas, *de novo*, cada año. Por ejemplo, para una equiparación estricta se deben desarrollar al menos dos años de pruebas en el primer año. Con otros esquemas de calibración, deben haber ítemes comunes desarrollados por anticipado.

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

La elaboración de escalas

La Teoría de Respuesta al Item (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 antes mencionados. 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.

A veces se puede topar uno con una actitud procrusteana[4] que restringe el dominio de la medición a algún subdominio o subpoblación en el cual el modelo es aplicable. Esto no puede ser recomendado para aplicaciones educacionales en general, a

[4] Procustes era el hostelero griego que "recortaba" a sus huéspedes para que se adaptaran a sus camas!

menos que se haga de una manera muy leve, con la eliminación de una pequeña parte de un dominio de contenido o población estudiantil.

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 ítems, 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 que sean 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. También significa que la precisión de los reportes será tomada en cuenta correctamente. Finalmente significa que las comparaciones al interior de y entre partes de la evaluación serán hechas sólo cuando y 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, mieantras 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.

CONCLUSIONES Y RECOMENDACIONES

Pedro Ravela

El propósito de este capítulo final de cierre del documento es realizar un rápido resumen de las principales reflexiones formuladas a lo largo del mismo 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 en el marco de su Grupo de Trabajo sobre estándares y evaluación, deberían impulsar en los próximos años.

Durante la década de los 90´ se ha desarrollado en América Latina una primera fase de instalación de sistemas de evaluación de aprendizajes a nivel nacional. Ha tenido 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 el 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 –si bien muchas veces ello ocurre más rápidamente en los discursos y documentos que en los presupuestos- resultará 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 de la labor del sistema educativo.

Ahora bien, el panorama resultante de esta primera década de experiencias incipientes en materia de evaluación de aprendizajes a nivel nacional muestra una importante diversidad de enfoques. Por ejemplo, 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. 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 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.

Lo anterior son solamente algunos ejemplos. Pero la variedad de experiencias es enorme y abarca otros aspectos tales como el tipo de conocimientos y competencias que son evaluados, la periodicidad de las evaluaciones, los grados y áreas curriculares que

enfocan, el tipo de variables contextuales sobre las que se releva información, los análisis y formatos de devolución de información, etc. 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 casos, se han adoptado modelos en forma tal vez menos reflexiva. En todos los casos, la fragilidad propia de los primeros pasos dados en un terreno nuevo y desconocido.

Simultáneamente es preciso señalar que, como es natural que ocurra, los sistemas de evaluación compiten por recursos con otras actividades y necesidades igualmente importantes. Por tanto, su sostenibilidad depende de que se aprovechen y maximicen los beneficios que en cierto modo "prometen" a la política educativa. Asimismo, los sistemas de evaluación se encuentran aún en una fase de institucionalización, en el sentido de que en la mayor parte de los casos su continuidad en el tiempo está fuertemente atada a los cambios políticos y/o a los recursos aportados por organismos internacionales de crédito.

En este contexto, las páginas anteriores fueron escritas con la intención de realizar un aporte a la reflexión sobre los próximos pasos a dar para fortalecer los sistemas de evaluación de aprendizajes en América Latina, en 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 en 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 materia de evaluación.

En este documento se ha priorizado cuatro grandes temas centrales, sobre los que se intentó aportar elementos para la revisión del camino recorrido y la reflexión sobre el camino a recorrer.

En primer término, el documento 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, entonces 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. De la misma manera, en el capítulo cinco 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 los sistemas de evaluación en América Latina están entregando a la sociedad, a los maestros y a las autoridades educativas.

En segundo término el documento intenta alertar sobre la necesidad de realizar un análisis más cuidadoso de 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 constante de "formación permanente" de los usuarios –autoridades, opinión pública, medios de comunicación, maestros- respecto a los usos válidos de los distintos tipos de información y respecto al tipo de interpretaciones y conclusiones que no es posible extraer válidamente de la información que aportan estas evaluaciones. Obviamente se trata de un campo altamente especializado, que requiere de un proceso intencional y sistemático de enriquecimiento de la cultura de los usuarios en el tema. Para ello, 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, el documento plantea la necesidad de profundizar la discusión respecto al enfoque de diseño de pruebas más adecuado. El diseño de pruebas en la mayoría de los países de la región 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 si el individuo dominaba o no ciertos campos del conocimiento, sino que el propósito principal era 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 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 su paso por el sistema educativo. Este enfoque parece como el más adecuado para sistemas de evaluación de aprendizajes que no tienen como finalidad la selección de individuos sino la producción de información relevante para la mejora del curriculum y la enseñanza. Para ello es imprescindible enriquecer y mejorar los procedimientos de diseño de pruebas e ítemes. Asimismo, observando las tendencias en curso en los países con mayor trayectoria en materia de evaluación, parece necesario también comenzar a avanzar en el terreno del diseño y corrección estandarizada de pruebas de desempeño.

En cuarto lugar, el documento 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 los aprendizajes ocurren. La mayor parte de los países recoge información sobre variables sociales y escolares, pero en muchos casos la calidad de la misma no es suficiente y menos 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 de tipo institucional y pedagógico con el fin de desarrollar trabajos de investigación 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, desarrollados a lo largo del presente documento -que obviamente no pretenden agotar el diagnóstico de la situación de los sistemas de evaluación de la región- 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[5]:

a. un primer eje estaría relacionado con ***el papel de los sistemas de evaluación en la política educativa,*** es decir, con 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;

b. un segundo eje estaría constituido por ***la necesidad de mejorar la calidad técnica de los diversos aspectos constitutivos de los sistemas de evaluación***, principalmente en lo que hace al diseño de los instrumentos de recolección de información y a los modos de procesar y reportar los resultados;

c. un tercer eje estaría enfocado hacia la discusión sobre ***las estrategias de uso y difusión de los resultados de las evaluaciones***, que si bien está estrechamente relacionado con los dos anteriores, merece una consideración específica.

La distinción entre estos tres ejes es estrictamente analítica, dado que los tres constituyen facetas de un mismo desafío que están estrechamente vinculadas entre sí. En cierto modo, la dilucidación que se dé al primero de estos ejes determina fuertemente las decisiones al interior de los dos restantes.

En cuanto al primer eje, sería necesario 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 hecho de evaluar, por sí mismo puede tener un efecto positivo sobre el sistema educativo -por el mero hecho de dar "visibilidad" a los resultados-, es necesario dar pasos en la dirección de delinear a nivel nacional una estrategia más específica al respecto. En la región es posible identificar una diversidad de estrategias para buscar que el sistema de evaluación tenga impacto en el sistema educativo. En algún caso se ha buscado que el impacto se produzca a través

[5] Habría un cuarto eje relevante a mencionar, cuyo tratamiento escapa a los alcances del presente documento, que es la participación de los países de la región en evaluaciones internacionales y/o la realización de evaluaciones a nivel regional.

del control de los padres sobre la calidad de las escuelas, en otros casos se ha buscado que este impacto se produzca por la vía de utilizar las pruebas para certificar la aprobación de cierto nivel de enseñanza, en otros se ha utilizado la información principalmente para promover la actualización docente y el aprendizaje profesional al interior de las escuelas.

En la Introducción de este documento se planteó al respecto una amplia gama de alternativas de política en materia de evaluación. Asimismo, a lo largo del mismo se mostró como las diversas alternativas técnicas en cuanto al desarrollo del sistema de evaluación dependen de decisiones de política educativa relativas a la finalidad de las evaluaciones. En este sentido, en la mayoría de los países de la región la primera etapa del desarrollo de los sistemas de evaluación de aprendizajes estuvo signada por la premisa básica general de que "evaluar ayuda a mejorar", pero se careció de una reflexión en profundidad sobre las alternativas de política 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[6]. Esta reflexión 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 curricula? ¿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

[6] 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.

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. Estos casos podrían luego difundirse bajo la forma de una serie de publicaciones o en el marco de seminarios de discusión. Un desafío importante a considerar es de qué modo involucrar en estos debates a otros actores de la política educativa y no sólo a los técnicos directamente involucrados en las evaluaciones.

En cuanto al segundo eje de trabajo, 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, a lo largo del documento 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. Un tipo de actividad relevante sería la realización de Seminarios en los que participarían países que estén dispuestos a someter sus instrumentos y procedimientos de evaluación al escrutinio de otros. En estos Seminarios se realizaría un análisis detallado de las fortalezas y debilidades de los instrumentos y procedimientos empleados, con la participación de especialistas de primer nivel de la comunidad internacional. Ello podría servir, asimismo, 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.

Otra iniciativa relevante que sería necesario impulsar es 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, en relación con este segundo eje es preciso 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.

En cuanto al tercer eje es preciso afirmar que en realidad todavía es muy poco lo que sabemos acerca del uso e impacto que la información producida por los sistemas de evaluación de aprendizaje tiene en sus potenciales usuarios. En las páginas anteriores se

ha mencionado la heterogeneidad de experiencias en los países de la región, que incluye países que publican en la prensa un ranking de escuelas que incluye a todos los establecimientos, países en que los resultados por establecimiento son confidenciales, países que entregan públicamente resultados nacionales y provinciales todos los años y países en que después de varios años de aplicación de evaluaciones aún no han dado a conocer ningún resultado. Sin embargo, prácticamente 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;
- 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[7].

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 por tratarse de algo relativamente nuevo en la región, un evento como el descrito aportaría

[7] Recientemente Martin Carnoy y Luis Benveniste de la Universidad de Stanford desarrollaron un estudio comparativo de los sistemas de evaluación de aprendizajes en Argentina, Chile y Uruguay, que incluye aspectos relativos sus impactos en las escuelas y que será publicado próximamente.

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.