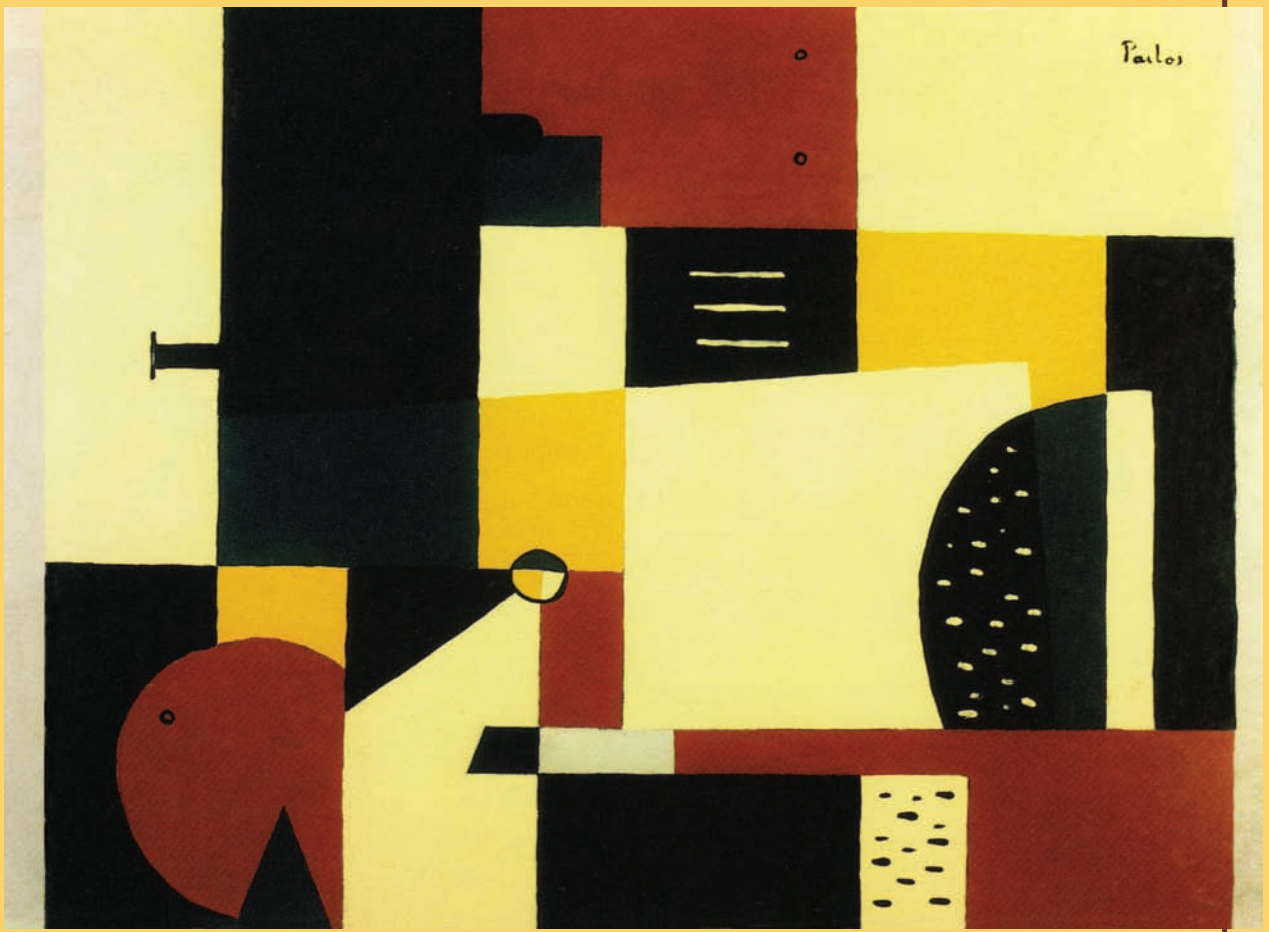
FICHAS DIDACTICAS

Para comprender las evaluaciones educativas

PEDRO RAVELA



PREAL

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

Para comprender las evaluaciones educativas

FICHAS DIDÁCTICAS

Pedro Ravela

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

Para comprender las evaluaciones educativas
FICHAS DIDÁCTICAS

Editor:
Pedro Ravela

Diseño portada:
Paulina Montalva

Imagen portada:
Máquina. Manuel Pailós, Montevideo, 1958. Esmalte sobre tela.

Primera edición:
1.000 ejemplares
Diciembre 2006

I.S.B.N.:956-8589-00-7

Registro de Propiedad Intelectual:
Inscripción Nº 156.754

Diseño e impresión:
Editorial San Marino
E-mail: sanmarino@adsl.tie.cl

Reconocimientos
La publicación de este libro, como las actividades del PREAL, son posibles gracias al apoyo de la United States Agency for International Development (USAID), el Banco Interamericano de Desarrollo (BID), el Banco Mundial, la International Association for the Evaluation of Educational Achievement (IEA), The Tinker Foundation, GE Foundation, entre otras.

CONTENIDO

Página

Introducción

Estas "Fichas Didácticas" sobre evaluación educativa tienen como propósito servir como material de base para facilitar la comprensión de los datos, usos, posibilidades y limitaciones de las evaluaciones estandarizadas por parte de diferentes audiencias.

El material ha sido elaborado en un nivel de dificultad y complejidad intermedia. Puede ser utilizado como base para la organización de talleres o seminarios o materiales de lectura más breves, dirigidos a personas no especializadas pero interesadas en el campo educativo (ciudadanos en general, políticos, periodistas, padres). En este caso, será necesario simplificar el material, convertirlo en un conjunto de presentaciones gráficas, etc.

Pero también puede ser empleado como guía inicial para cursos de evaluación dirigidos a personas profesionalmente vinculadas al mundo educativo (futuros docentes, docentes en servicio, personal técnico de ministerios de educación, etc.). En estos casos, será necesario desarrollar muchos de los temas con mayor profundidad y complejidad, así como remitir a los usuarios a literatura especializada sobre los diversos temas.

El propósito de las Fichas es organizar y ofrecer un conjunto de conocimientos y explicaciones básicas sobre evaluación educativa, que ayuden a comprender mejor los informes resultantes de las evaluaciones nacionales e internacionales estandarizadas, así como los principales debates en relación a este tipo de evaluaciones y al uso de sus resultados.

Cada Ficha busca responder a una pregunta central, explicando y ejemplificando los conceptos básicos necesarios para ello. Si bien las Fichas pueden ser leídas en forma independiente, una línea conceptual las atraviesa a todas. En general, se encontrarán referencias entre unas y otras. Algunos temas son simplemente enunciados en una Ficha, para ser retomados con mayor profundidad en otra.

El conjunto de Fichas constituye una suerte de "manual didáctico sobre evaluación".

El trabajo se inscribe en el marco de las actividades del "Grupo de Trabajo sobre Estándares y Evaluación" del PREAL, entre cuyas finalidades se encuentran las de colaborar para mejorar las evaluaciones nacionales de logro educativo, contribuir a la reflexión sobre los enfoques y estrategias de evaluación y propiciar que sus resultados sean comprendidos y utilizados por la mayor cantidad de actores posible.

La responsabilidad principal en la redacción de las Fichas estuvo a cargo de Pedro Ravela. Tacyana Arce, Patricia Arregui, Santiago Cueto, Guillermo Ferrer, Richard Wolfe y Margarita Zorrilla aportaron valiosas observaciones, comentarios y críticas que permitieron mejorar el trabajo.

Ficha 1
¿QUÉ SON LAS EVALUACIONES EDUCATIVAS Y PARA QUÉ SIRVEN?
un recorrido por las evaluaciones y sus finalidades

Esta primera Ficha recorre la diversidad de evaluaciones que cotidianamente tienen lugar en los sistemas e instituciones educativas, desde las evaluaciones que ocurren a diario dentro de las aulas hasta las evaluaciones internacionales que comparan los resultados de diversos países.

El énfasis en esta Ficha está puesto en mostrar que existen diversos tipos de evaluaciones que responden a distintas finalidades o necesidades del sistema educativo o de la sociedad. Se propone una clasificación de los propósitos de las evaluaciones en seis grandes tipos:

a. acreditación y certificación;
b. selección;
c. toma de decisiones "blandas" o decisiones de mejora;
d. toma de decisiones "duras";
e. establecimiento de incentivos y sanciones;
f. rendición de cuentas.

La Ficha busca mostrar que el primer paso para comprender y juzgar los resultados de cualquier evaluación es conocer sus propósitos y finalidades.

Ficha 2
¿CÓMO SE HACEN LAS EVALUACIONES EDUCATIVAS?
los elementos básicos del proceso de evaluación

La Ficha 2 analiza cinco elementos básicos que subyacen a todo proceso de evaluación:

a. la selección y definición de la realidad a evaluar;
b. la definición de los propósitos de la evaluación;
c. la producción de evidencia empírica;
d. la formulación de juicios de valor;
e. la toma de decisiones o acciones que modifiquen la realidad evaluada.

Se explica en qué consiste cada uno de estos elementos con ejemplos de distintos tipos de evaluaciones educativas y se muestra que lo que define a una evaluación es la formulación de valoraciones o juicios de valor sobre la realidad.

Se enfatiza también que la vocación principal de toda evaluación es incidir sobre la realidad, pero que, al mismo tiempo, la evaluación por sí misma no produce cambios si no hay actores que tomen decisiones a partir de los resultados y valoraciones resultantes de la evaluación.
Se explican y ejemplifican algunas de las limitaciones inevitables de cualquier proceso de evaluación y, por tanto, los cuidados que es necesario tener al analizar y utilizar sus resultados.

Finalmente, se señala de qué modo es posible mejorar las evaluaciones y su potencial de transformación, mejorando los elementos básicos del proceso de evaluación.

Ficha 3
¿CÓMO SE FORMULAN LOS JUICIOS DE VALOR EN LAS EVALUACIONES EDUCATIVAS?
evaluaciones normativas, de progreso y criteriales

La Ficha 3 profundiza en uno de los elementos que definen a una evaluación, la formulación de juicios de valor acerca de los individuos, instituciones o realidades evaluados.

Se explican y ejemplifican tres enfoques principales para formular estas valoraciones:

a. comparar las posiciones relativas entre individuos y grupos (enfoque normativo);
b. comparar el avance o retroceso de un individuo o grupo respecto a su propio desempeño en un momento anterior (enfoque de progreso o de crecimiento);
c. comparar el desempeño demostrado por cada individuo con una definición clara y exhaustiva del dominio evaluado y del nivel de desempeño deseable (enfoque criterial).

En la Ficha se muestra cómo estos enfoques son utilizados normalmente en forma complementaria en las aulas, se explican las virtudes y debilidades de cada uno de ellos, así como sus implicancias para las evaluaciones estandarizadas a gran escala.

Se explica por qué el enfoque criterial es el que mejor permite dar respuesta a las principales preguntas que los diversos actores involucrados en el quehacer educativo esperan que los sistemas nacionales de evaluación respondan; así como las confusiones que se generan cuando los resultados de una evaluación que ha sido desarrollada con un enfoque normativo son interpretados como si la misma fuese de tipo criterial.

Ficha 4
¿CUÁLES SON LOS PRINCIPALES PROBLEMAS COMUNES A TODAS LAS EVALUACIONES EDUCATIVAS?
validez y confiabilidad

La Ficha 4 retoma el esquema de elementos principales del proceso de evaluación de la Ficha 2 para explicar los dos problemas principales que toda evaluación debe enfrentar: la validez y la confiabilidad.

A través de una serie de ejemplos, se muestra cómo estos problemas están presentes, aunque de distinta manera y con distinta importancia, en todas las evaluaciones educativas: en las evaluaciones que los profesores realizan en las aulas, en las evaluaciones estandarizadas, en los concursos para el acceso a plazas docentes, etc. De allí la necesidad de encararlos, para mejorar los modos de evaluar en todas las áreas del sistema educativo.

A partir de los ejemplos se definen conceptualmente, la validez y la confiabilidad.

La Ficha muestra además que, si bien es posible minimizar o controlar estos problemas, los mismos nunca pueden ser eliminados, por lo que siempre están presentes en las evaluaciones.

Por este motivo, toda lectura y uso inteligente de los resultados de una evaluación requiere analizar cómo estos problemas han sido encarados y controlados.

Ficha 5
¿DEBEMOS CREERLE A LAS EVALUACIONES ESTANDARIZADAS EXTERNAS O A LAS EVALUACIONES QUE REALIZA EL DOCENTE EN EL AULA?

los debates ideológicos sobre las evaluaciones

La Ficha 5 se propone explicitar y clarificar algunas de las contraposiciones que suelen plantearse en los debates sobre evaluación educativa, en particular entre la evaluación estandarizada y la evaluación en el aula, entre evaluación de resultados y evaluación de procesos, entre evaluación cuantitativa y evaluación cualitativa.

Se propone además discutir y aclarar algunos de los prejuicios más comunes acerca de las evaluaciones estandarizadas.

La Ficha analiza también lo que puede aportar la evaluación estandarizada, así como sus limitaciones.

La Ficha también pretende mostrar que evaluaciones estandarizadas y evaluaciones en el aula son complementarias y no necesariamente antagónicas. Cada una de ellas permite "ver" o "hacer" algunas cosas, pero no otras.

Finalmente, se explican algunas de las tensiones implícitas en los procesos de evaluación: entre lo local o contextual y lo general, entre la diversidad y la integración social.

Ficha 6
¿QUÉ EVALÚA ESTA PRUEBA? (I)
distintos tipos de actividades en las evaluaciones estandarizadas

La Ficha 6, junto con la Ficha 7, muestran con ejemplos cómo distintas pruebas para una misma disciplina pueden en realidad evaluar cosas muy distintas.
El propósito principal de estas Fichas es alertar y preparar al lector para que, antes de mirar los "números" de las evaluaciones, ponga su atención en mirar qué fue evaluado, en particular, qué tipo de actividades fueron propuestas y qué grado de complejidad tuvieron las tareas que los alumnos debieron enfrentar.

Para ello se muestra la diversidad de tipos de actividades empleadas para evaluar lenguaje y matemática tanto en evaluaciones nacionales de la región como en evaluaciones internacionales.

La Ficha 6 explica y ejemplifica la diversidad de formatos de ítemes que pueden ser empleados en las evaluaciones estandarizadas, así como las ventajas y problemas de cada uno de ellos.

Ficha 7
¿QUÉ EVALÚA ESTA PRUEBA? (II)
contenidos, currículo y competencias

A partir de lo explicado en la Ficha anterior, la Ficha 7 se propone profundizar en el análisis de los distintos marcos conceptuales desde los cuales se elaboran las evaluaciones.

Dos evaluaciones de Lenguaje o de Matemática pueden estar formuladas desde concepciones muy distintas de cuáles son los conocimientos y capacidades fundamentales que los alumnos deberían aprender.

Para ello se presentan distintos ejemplos y se explica el significado de algunos términos de uso cada vez más frecuente, que suelen ser objeto de controversia y uso inadecuado: competencias, contenidos, currículo.

Se analiza la relación entre evaluación y currículo, así como sus variantes según los propósitos de la evaluación.

Se pone especial énfasis en que la evaluación orienta y define en buena medida la enseñanza (lo que se evalúa pasa a ser importante para la enseñanza), por lo cual la evaluación debe estar sustentada en buenas definiciones de los aprendizajes a evaluar y en buenas actividades de prueba.

Ficha 8
¿QUÉ SIGNIFICAN LOS NÚMEROS DE LAS EVALUACIONES? (I)
elementos básicos para comprender los datos estadísticos

Las Fichas 8 y 9 pretenden ayudar al lector a comprender el significado de los números con que se reportan los resultados de las evaluaciones estandarizadas.

Para ello se intenta explicar de manera accesible una serie de conceptos imprescindibles para comprender los datos que se incluyen en los reportes.

En primer lugar, la Ficha 8 explica en forma sintética los dos principales marcos metodológicos existentes para la construcción de las pruebas y sus puntajes: la Teoría Clásica de los Tests y la Teoría de la Respuesta al ítem.

En segundo lugar, la Ficha explica las dos formas básicas que pueden tomar los datos estadísticos: promedios y distribución de frecuencias. La Ficha 8 se concentra en el uso de promedios para el reporte de resultados, en tanto la Ficha 9 lo hace con el reporte mediante distribuciones de frecuencias.

A lo largo de la Ficha se muestran y explican ejemplos de reportes que utilizan diferentes tipos de promedios, analizando además el tipo de interpretaciones válidas y los errores comunes que pueden cometerse. Se explica en particular la importancia de analizar la dispersión de puntajes cuando se trabaja con promedios.

La Ficha explica además los conceptos de error estándar de medición e intervalo de confianza. Se explica brevemente qué es importante tener en cuenta para realizar comparaciones entre los resultados de evaluaciones realizadas en distintos momentos.

Ficha 9
¿QUÉ SIGNIFICAN LOS NÚMEROS DE LAS EVALUACIONES? (II)
elementos básicos para comprender los datos estadísticos

La Ficha 9 es una continuación de la anterior. Está focalizada en el reporte de resultados a través de la distribución de frecuencias de los alumnos en categorías de desempeño. Para ello, se intenta ejemplificar y explicar de qué manera los puntajes de una prueba son transformados en niveles de desempeño que permiten describir mejor qué conocen y qué son capaces de hacer los alumnos.

Se explica la diferencia fundamental existente entre el hecho de establecer niveles de desempeño de carácter descriptivo y el hecho de establecer un estándar o expectativa respecto a lo que todos los alumnos deberían conocer y ser capaces de hacer al finalizar un grado o ciclo educativo. Para ello, se muestran y explican ejemplos de reportes de evaluaciones nacionales e internacionales.

Finalmente, se explican algunos de los procedimientos técnicos existentes para establecer "puntos de corte" en una prueba, es decir, para definir cuál es el puntaje que los alumnos deberían alcanzar para que su resultado sea considerado satisfactorio.

Ficha 10
¿POR QUÉ LOS RANKINGS SON MODOS INAPROPIADOS DE VALORAR LA CALIDAD DE LAS ESCUELAS?
hacia nuevas formas de utilizar los resultados por escuela

La Ficha 10 se propone explicar la necesidad de evitar la elaboración de rankings que sobresimplifican la realidad de las instituciones y sistemas educativos, conduciendo a juicios inapropiados acerca de la calidad de los mismos.

En primer término, se explican los problemas técnicos que implica la construcción de este tipo de ordenamientos.

La Ficha muestra y explica cómo los *rankings* dan una falsa apariencia de ordenamiento riguroso, ignoran aspectos relevantes de la calidad educativa, ignoran la importancia que

en los resultados de las evaluaciones estandarizadas tiene la composición social del alumnado de escuelas y países.

En segundo término, se discute la disyuntiva política existente entre utilizar ordenamientos de escuelas con carácter público o limitarse a utilizarlos como fuente de información para la toma de decisiones al interior del sistema educativo. La experiencia internacional muestra que es más efectivo utilizar los resultados por escuela para establecer programas de apoyo a las escuelas con dificultades, que para establecer mecanismos de competencia entre los centros educativos.

La Ficha también muestra que es posible construir comparaciones válidas de los resultados de instituciones y sistemas educativos si se toman algunas precauciones tales como tener en cuenta la composición sociocultural del alumnado, trabajar con grupos de escuelas más que con un *ranking* único e individualizado que asigne posiciones definidas y diversificar los indicadores de calidad.

Ficha 11
¿QUÉ ES LA "RENDICIÓN DE CUENTAS"?
hacia la responsabilidad compartida

La Ficha Nº 11 se focaliza en un tema de creciente importancia en los debates de política educativa: la "rendición de cuentas" o "responsabilización por los resultados".

La Ficha se propone explicar los diversos significados del término y las visiones muchas veces contrapuestas que existen acerca de la misma.

Se exponen y analizan los principales debates en torno a la rendición de cuentas y algunas experiencias relevantes que se llevan adelante en diversos lugares del mundo. Se explica asimismo los efectos "perversos" o no deseados que algunos mecanismos de rendición de cuentas pueden desencadenar.

Al mismo tiempo, se destaca la importancia de la rendición de cuentas como mecanismo de involucramiento de la ciudadanía con la educación, de transparencia democrática y de mejora de la gestión y de la enseñanza que se brinda a los estudiantes.

Para ello se analizan las peculiaridades de la acción educativa que por un lado hacen necesaria la rendición de cuentas pero que, por otro, exigen encararla con una lógica articuladora de los esfuerzos de distintos actores y no como búsqueda y señalamiento de culpables de los problemas educativos.

Así la Ficha enfatiza la multiplicidad de actores que deberían hacerse responsables por los resultados en el sistema educativo: los directivos y docentes, pero también las autoridades centrales y locales, los técnicos, las familias y los propios alumnos.

Ficha 12
¿CÓMO USAR LAS EVALUACIONES PARA MEJORAR LA EDUCACIÓN?
los resultados como herramienta de aprendizaje y mejora

Complementando a la anterior, la Ficha 12 muestra algunos enfoques alternativos con los que los países de la región y del mundo "desarrollado" utilizan los resultados de las evaluaciones estandarizadas para intentar que estas contribuyan a mejorar el trabajo de escuelas y docentes y los logros educativos de los estudiantes.

Se ilustran los diversos enfoques con ejemplos de experiencias nacionales y estaduales, y se vincula el tratamiento de este tema con lo analizado en la Ficha 1 en torno a los propósitos y consecuencias de las evaluaciones.

Finalmente, se plantea la importancia de que a la hora de elaborar los reportes de las evaluaciones se tenga presente la diversidad de audiencias a las deben estar dirigidos y algunas características específicas que deberían tener los informes para cada audiencia.

Ficha 13
¿QUÉ SON LOS "FACTORES ASOCIADOS"?
intentando comprender los sistemas educativos

La Ficha 13 explica el significado de los denominados "factores asociados", es decir, aquellos factores externos e internos a la organización escolar que influyen sobre los aprendizajes de los alumnos.

La Ficha pretende, por un lado, mostrar la importancia de investigar sobre los aspectos relevantes que tienen influencia sobre los aprendizajes y que pueden ser objeto de decisiones y acciones al interior de las instituciones y políticas educativas.

Por otro lado, la Ficha se propone explicar el modo en que se investigan estos factores, sus alcances y sus limitaciones. También orienta al lector sobre modos apropiados e inapropiados de hacer investigación sobre estos temas, para que pueda leer en forma analítica y crítica reportes en este campo.

Finalmente, la Ficha incluye un resumen de los principales hallazgos de las investigaciones sobre los factores que inciden en los aprendizajes.

Ficha 14
¿CÓMO ANALIZAR UN REPORTE DE EVALUACIÓN?
las preguntas que el lector debe hacerse ante un informe de resultados

La Ficha 14 intenta ser un resumen de cierre del conjunto del trabajo que oriente al lector acerca de lo que contienen o deberían incluir los reportes de las evaluaciones –marcos conceptuales, datos, definiciones, ejemplos de ítemes, etc.– relacionándolo con los temas analizados a lo largo de las Fichas.

Para ello, la Ficha propone al lector un conjunto de preguntas fundamentales que debe formularse ante todo reporte de resultados de una evaluación:

1. ¿Cuál fue el propósito o finalidad de la evaluación?
2. ¿Qué fue evaluado?
3. ¿Cuál fue el universo estudiado?
4. ¿Cuál fue el enfoque de la evaluación?
5. ¿Qué tipo de datos se proporcionan y qué significan esos números?
6. ¿Qué grado de precisión tiene la información?
7. ¿En qué grado y de qué modo se contextualiza la presentación de los resultados?
8. ¿Quiénes son los destinatarios de la información?
9. ¿Qué consecuencias e implicancias tienen los resultados?

¿QUÉ SON LAS EVALUACIONES EDUCATIVAS Y PARA QUÉ SIRVEN?

Un recorrido por las evaluaciones y sus finalidades

1. La evaluación en los sistemas educativos

La evaluación está presente de manera permanente en la vida cotidiana de los sistemas educativos, docentes y estudiantes.

En las aulas, los maestros y profesores evalúan a sus alumnos, algunas veces con el propósito de conocer qué han aprendido y cuáles son sus dificultades, de modo de ayudarlos en su proceso de aprendizaje o con el propósito de otorgarles una calificación.

Estas evaluaciones suelen ser realizadas a través de diversas modalidades: pruebas escritas, observación y registro del comportamiento de los alumnos y de sus intervenciones en la clase, "toma de la lección" a alumnos individuales, etc.

Al finalizar el año lectivo, generalmente los profesores deben decidir cuáles alumnos aprueban el curso y cuáles no. Esta decisión puede derivarse tanto de la acumulación de información sobre lo que el alumno aprendió a lo largo del año, como de una prueba o examen final. El examen, en muchos países, es elaborado y calificado a nivel local, en cada centro educativo. En otros casos, se trata de exámenes nacionales que son elaborados a nivel central y aplicados a todos los alumnos por igual.

Los propios profesores y maestros son visitados y evaluados por sus directores, supervisores o inspectores. Algunas de estas evaluaciones tienen como finalidad orientar al docente para realizar mejor su trabajo; otras sí asignan al docente una calificación, que normalmente tiene importancia para su avance en la carrera. A veces una misma evaluación busca cumplir con ambos propósitos.

Estas evaluaciones también se realizan a través de diversos procedimientos e instrumentos: observación del docente trabajando en clase con sus alumnos, una entrevista al docente, revisión de la planificación del curso, etc.

Los docentes también son evaluados en muchos países para acceder a su plaza. En algunos países esto se hace a través de un conjunto de pruebas teóricas y prácticas que se denominan "concurso". Como resultado de estas pruebas suelen ocurrir dos cosas distintas:

- en primer lugar, se distingue entre los docentes que han logrado los requisitos mínimos para acceder a la plaza y aquellos que no lo han logrado;
- en segundo lugar, se ordena a los candidatos en función del resultado obtenido, de modo que los mejor evaluados tendrán prioridad a la hora de elegir plaza.

En otros países, esta labor de selección es realizada directamente por los directores de los centros educativos y, a veces, por agencias especializadas en evaluación y certificación de docentes.

Los instrumentos de evaluación pueden ser muy diferentes: una prueba de conocimientos sobre la disciplina que el docente va a enseñar, la observación del docente dictando una clase o una entrevista.

Recientemente ha comenzado a utilizarse como instrumento de evaluación el denominado "portafolio", que es una colección de muestras de su trabajo que el docente prepara siguiendo criterios dados. El portafolio puede incluir planes de clases, consignas de actividades propuestas a sus alumnos, actividades de evaluación preparadas por él o ella, trabajos realizados por los alumnos, artículos publicados, etc.

En muchos países los estudiantes también deben presentarse a evaluaciones con el fin de acceder a la educación terciaria. En estos casos, normalmente se les aplica una o varias pruebas, a partir de las cuales los jóvenes son ordenados en función de sus aptitudes para continuar estudios terciarios. El resultado de las pruebas es utilizado por las instituciones de educación terciaria para seleccionar entre los candidatos que aspiran a ingresar a ellas.

En un número creciente de países se aplican pruebas de logro educativo a los alumnos con la finalidad de conocer en qué medida están siendo logrados los conocimientos y competencias que se espera que los estudiantes hayan adquirido cuando finalizan un ciclo o nivel del sistema educativo. Estas pruebas no tienen consecuencias directas para los alumnos. Su finalidad principal es tener un diagnóstico de algunos resultados de la labor educativa y, a veces, de su evolución a lo largo del tiempo.

Las pruebas de este último tipo están dirigidas a evaluar al sistema educativo y no a los alumnos. Normalmente se trata de pruebas de carácter nacional, pero también las hay de carácter provincial o regional.

Además, en los últimos años han cobrado creciente importancia las evaluaciones internacionales, que buscan comparar lo que han aprendido los estudiantes en diferentes países. TIMSS y PISA son las evaluaciones internacionales más conocidas. A nivel regional, la UNESCO realizó en 1997 una evaluación latinoamericana a nivel de la escuela primaria, y actualmente está preparando un segundo estudio para el 2006.

Estas evaluaciones de carácter diagnóstico muchas veces sirven como evaluación de las políticas educativas en curso y se supone que aporten información útil para tomar decisiones que permitan mejorar el sistema educativo.

Los centros o instituciones educativas también son, muchas veces, objeto de evaluación. Por ejemplo, en muchos países existen procesos de evaluación de las instituciones de educación superior para acreditarlas como universidades.

Los centros de educación primaria y media son evaluados de muy diversas maneras. En algunos países esto se hace mediante procedimientos de tipo predominantemente administrativo-burocráticos. En otros, como en el caso de Escocia, la Inspección tiene un esquema de evaluación integral que tiene en cuenta una amplia gama de aspectos de la vida de la institución. Finalmente, en algunos países se utilizan los resultados de los estudiantes en pruebas o exámenes nacionales como medida de la calidad de los centros educativos.

También el currículo es objeto de evaluación. En algunos casos, se evalúa el grado en que el mismo es considerado adecuado y relevante por diversos actores sociales. La pregunta que se busca responder es en qué medida lo que se intenta enseñar en las escuelas es lo que los alumnos necesitan aprender para continuar estudiando en el nivel superior, para ejercer la ciudadanía o para estar en condiciones de conseguir un empleo.

En otros casos, la evaluación del currículo está dirigida a conocer en qué medida el mismo está siendo realmente enseñado en las escuelas y los alumnos están teniendo oportunidades para aprender lo que se espera que aprendan.

Para terminar este recorrido, que no pretende ser exhaustivo, cabe mencionar que los

proyectos o innovaciones que se introducen en el sistema educativo también son a veces objeto de evaluación. Por ejemplo, se evalúa en qué medida un programa destinado a otorgar becas a alumnos de sectores pobres para que no abandonen la escuela está beneficiando a los destinatarios esperados y está logrando efectivamente retener a los alumnos.

Estas evaluaciones pueden servir tanto para tomar decisiones que permitan mejorar el proyecto, o programa como para decidir acerca de su continuidad o finalización.

2. Los propósitos de las evaluaciones

Este "recorrido" por las evaluaciones educativas tuvo un primer objetivo: mostrar al lector la diversidad de modalidades de evaluación que continuamente están ocurriendo al interior de los sistemas educativos. Con ello se pretende ubicar a las evaluaciones "estandarizadas" como una modalidad, entre otras, en el conjunto de las evaluaciones educativas.

Es importante tener conciencia de que en educación continuamente se está evaluando y que la gran mayoría de estas evaluaciones tiene actualmente importantes debilidades y problemas que analizaremos a lo largo de las Fichas. Estas debilidades no son exclusivas de las evaluaciones estandarizadas, sino que también están presentes en las evaluaciones realizadas por los docentes en las aulas, así como en las evaluaciones de docentes y de centros educativos, etc.

Todos los tipos de evaluación son cruciales para la mejora del sistema educativo. Para mejorar la calidad de los sistemas educativos es importante no solamente realizar buenas evaluaciones estandarizadas, sino que, además, es necesario mejorar el modo en que los docentes evalúan a sus alumnos en las aulas, las evaluaciones de que son objeto los propios docentes, las evaluaciones mediante las cuales se selecciona a los directivos y supervisores, las evaluaciones de los proyectos e innovaciones, etc.

El "recorrido" por las evaluaciones educativas tuvo, además, un segundo objetivo más específico: mostrar al lector que las evaluaciones educativas se distinguen, primero que nada, por sus propósitos o finalidades. Con este objetivo, la Figura 1 ofrece un "mapa" de las evaluaciones educativas. Las casillas sombreadas muestran cuáles son los sujetos o realidades que pueden ser evaluados y cuáles son las finalidades o propósitos para los que pueden ser evaluados

Un "mapa" de las evaluaciones educativas según su objeto y finalidad ***Figura 1***

	Finalidades					
Sujetos o realidades a evaluar	Acreditación y/o certificación	Ordenamiento y selección	Toma de decisiones "blandas"	Toma de decisiones "duras"	Incentivos y sanciones	Rendición de cuentas
Alumnos	▲	▲	▲		▲	
Docentes	▲	▲	▲		▲	
Directivos/ supervisores	▲	▲	▲			
Centros educativos	▲	▲	▲		▲	▲
Curriculo	▲		▲	▲		
Proyectos/ innovaciones			▲	▲		
Sistemas educativos			▲			▲

Toda evaluación educativa tiene alguna de las siguientes finalidades, que responden a necesidades de la labor educativa o a funciones específicas que el sistema educativo debe desempeñar en la sociedad:

a. acreditación y/o certificación;
b. ordenamiento para la selección;
c. toma de decisiones de mejora (decisiones "blandas");
d. toma de decisiones "duras" (como, por ejemplo, discontinuar un proyecto);
e. establecimiento de incentivos para individuos o instituciones; y
f. rendición de cuentas y responsabilización por los resultados.

2.1. Acreditación y/o certificación

Toda evaluación cuyo propósito sea otorgar algún tipo de constancia formal, con reconocimiento social, de que un individuo o institución posee ciertas características o cualidades, es una evaluación de "acreditación". Generalmente se emplea el término "acreditación" para referirse a las instituciones y el término "certificación" para referirse a los individuos.

La "certificación" es una función social ineludible que las instituciones educativas deben cumplir. Al final de un grado o de un ciclo educativo, es necesario que las instituciones indiquen qué estudiantes han logrado los conocimientos y competencias estipulados para ese nivel –lo cual además implica que se encuentran preparados para realizar los estudios correspondientes al nivel siguiente–.

En otros casos, las instituciones educativas deben certificar que un individuo se encuentra capacitado para desempeñar un determinado oficio o profesión –carpintero, analista de sistemas, médico, profesor, etc.–, otorgándoles un "título" o "certificado". La sociedad en general otorga confianza y validez a estas certificaciones.

De la misma manera, cuando se recurre a un procedimiento de evaluación para proveer cargos de dirección o supervisión, o cargos técnicos en la estructura de gobierno del sistema educativo, se supone que la evaluación certifica que los individuos poseen los conocimientos y competencias necesarias para el desempeño del cargo en cuestión.

También existe la necesidad de "acreditar" (dar "crédito", en el sentido de confianza pública) a las instituciones y programas educativos. Esto ocurre con mayor frecuencia en el nivel terciario. Las instituciones que desean ser reconocidas como universidades deben someterse a algún proceso de evaluación que las acredite como tales. En otros casos son carreras específicas las que son sometidas a evaluación para ser reconocidas como de carácter universitario.

Todo lo expresado anteriormente no implica que las evaluaciones se realicen en forma adecuada. Muchas veces se trata de trámites meramente administrativos más que de evaluaciones propiamente dichas. En algunos casos la certificación profesional se otorga por la mera acumulación de cursos aprobados, sin que la institución responsable tenga cabal conciencia del papel social que está desempeñando al otorgar un título. Pero en esta Ficha solo se pretende mostrar cuáles son las necesidades del sistema educativo y de

la sociedad que dan lugar a los diferentes tipos de evaluaciones. En las siguientes Fichas se analizarán los problemas y dificultades de los distintos tipos de evaluación.

2.2. Ordenamiento para la selección

Un segundo tipo de necesidad para el cual se requiere del uso de evaluaciones es el ordenamiento de individuos o instituciones.

Un primer caso se verifica cuando las plazas en las instituciones de nivel terciario son limitadas y el número de aspirantes es superior a la cantidad de las mismas. En estos casos, muchos países utilizan pruebas de aptitud cuya finalidad es producir un ordenamiento de los candidatos, que es utilizado por las instituciones –a veces junto con otra información sobre los mismos– para seleccionar a quienes habrán de ser admitidos.

Un segundo caso muy común en los sistemas educativos son las evaluaciones dirigidas a ordenar a los candidatos a ocupar cargos o plazas vacantes –docentes, de dirección o de supervisión–. También en este caso se realizan evaluaciones cuya finalidad es ordenar a los candidatos, teóricamente en función de su capacidad para la tarea, de modo que los más capaces sean los que accedan a dichos cargos. Estas evaluaciones suelen tener poca visibilidad, pero son de enorme importancia porque determinan la calidad de los profesionales que tendrán responsabilidades importantes en la conducción de la educación[1].

Un tercer caso se verifica cuando las instituciones educativas son ordenadas a partir de una evaluación. Un caso típico son los *rankings* de establecimientos en función de los resultados de sus alumnos en algún tipo de prueba o examen, que se publican con la pretensión de ofrecer a las familias y al público información acerca de la "calidad"[2] de las instituciones.

En otros casos es preciso ordenar a las instituciones para diversos propósitos como, por ejemplo, seleccionarlas para participar en un programa de apoyo de carácter compensatorio para las escuelas que trabajan con las poblaciones más desfavorecidas u otorgarles algún tipo de premio o incentivo por ser las instituciones que realizan mejor su labor educativa o que logran mejoras notorias en su situación.

En todos los casos mencionados más arriba existe el supuesto –que no siempre se cumple– de que las evaluaciones logran ordenar a los individuos o instituciones en función de

1) Muchas de estas evaluaciones cumplen simultáneamente dos finalidades: certificar la idoneidad para el cargo y ordenar en función del grado de idoneidad.

2) El entrecomillado es deliberado dado que, como veremos más adelante, la mayoría de estos rankings no da información apropiada sobre la calidad de las instituciones.

propiedades relevantes para el fin para el cual se realizará la selección y que el ordenamiento resultante es "justo" y "preciso", es decir, que realmente los mejores ocupan los primeros lugares y que los errores en los puntajes atribuidos son mínimos, de modo de evitar que individuos o instituciones con menor capacidad o mérito queden ubicados en el ordenamiento por encima de otros que son mejores.

2.3. Toma de decisiones "blandas" o decisiones de mejora

Todas las evaluaciones pueden –y deberían– ser utilizadas con el fin de comprender mejor la realidad con la que se está trabajando y contribuir a mejorarla.

Normalmente se utiliza el término "evaluación formativa" para designar a las evaluaciones cuyo propósito principal es servir de base para tomar decisiones y emprender acciones de mejora de aquello que ha sido evaluado. Es el caso de:

i) las evaluaciones de los alumnos realizadas por su profesor con el fin de comprender el proceso de aprendizaje y apoyar a cada alumno individualmente en función de sus dificultades;
ii) las evaluaciones de los docentes cuya finalidad es brindarles orientación profesional para desarrollar mejor su labor de enseñanza y de relacionamiento con los alumnos;
iii) las evaluaciones de centros educativos cuya finalidad es desencadenar procesos de diálogo entre los actores de las instituciones para buscar caminos de mejora en relación a los problemas existentes;
iv) las evaluaciones del currículo o de proyectos específicos cuya finalidad es detectar necesidades de cambio y mejora;
v) las evaluaciones nacionales e internacionales de logros educativos cuyo propósito es contribuir a identificar las principales debilidades de los sistemas educativos y orientar la reflexión y formulación de las políticas educativas.

En todos estos casos las evaluaciones no tienen una consecuencia formal específica para actor alguno. De allí la utilización del término decisiones "blandas", aunque, obviamente, toda divulgación de resultados genera consecuencias de algún tipo. Lo que se pretende con este tipo de evaluaciones es contribuir a mejorar la comprensión de la situación educativa y propiciar acciones y decisiones que permitan cambiar y mejorar.

2.4. Toma de decisiones "duras"

En otros casos, las evaluaciones están dirigidas a tomar decisiones "duras", en el sentido de que tienen consecuencias importantes y directas para un proyecto o institución.

Obviamente podrían entrar en esta categoría las evaluaciones de "acreditación y/o certificación", que tienen consecuencias directas e importantes para los actores evaluados, pero se ha optado por mantenerlas como una categoría diferente y reservar el término decisiones "duras" para aquellas evaluaciones de proyectos, innovaciones o programas educativos cuya finalidad específica es resolver acerca de la continuidad o terminación del proyecto o programa en cuestión.

2.5. Establecimiento de incentivos

Toda evaluación cumple, en forma explícita o implícita, una función de establecimiento de incentivos para la mejora del desempeño de individuos e instituciones.

Las evaluaciones con consecuencias directas para individuos o instituciones establecen un incentivo para lograr aquello que será evaluado: los estudiantes quieren aprobar el curso, los docentes necesitan obtener una buena calificación de su supervisor o inspector para acceder a un cargo o avanzar en su carrera, etc.

Los docentes suelen utilizar la calificación como forma de establecer incentivos o sanciones para los alumnos. Muchas veces lo hacen amenazando con una baja calificación cuando no logran motivar a sus alumnos para que estudien. Otras veces lo hacen asignando a los alumnos calificaciones bajas al inicio del año, de modo que puedan ir "subiéndolas" a lo largo de este. En este último caso la calificación no refleja lo que el alumno ha aprendido en un período de tiempo, sino que es utilizada como incentivo para mantener el esfuerzo de los alumnos por mejorar a lo largo de todo el año.

También las evaluaciones de carácter formativo cuando están bien realizadas, constituyen un incentivo para mejorar.

Normalmente los individuos encuentran un estímulo personal y profesional en aquellas evaluaciones que les brindan oportunidades para aprender a desarrollar mejor su trabajo.

Finalmente, en los últimos años y a partir de propuestas surgidas principalmente del ámbito

de los economistas, se ha abierto camino la idea de emplear las evaluaciones para establecer incentivos de carácter económico para individuos o instituciones educativas. La idea básica es ofrecer pagos complementarios en función de los resultados de los alumnos en pruebas estandarizadas.

El razonamiento que se hace es el siguiente: en la mayoría de los sistemas educativos los incrementos saíÈriales han estado vinculados principalmente a la antigüedad o al acceso a cargos superiores, pero no hay premios económicos para los buenos docentes, con lo cual no habría estímulos para desempeñarse mejor y para permanecer en el aula.

Por esta razón, se propone establecer incentivos económicos vinculados a diversos aspectos: al logro educativo de los alumnos en pruebas estandarizadas, a los esfuerzos de capacitación académica, al riguroso cumplimiento de las obligaciones en materia de puntualidad y asiduidad en la asistencia, etc.

Estos esquemas admiten muchas variantes. En algunos casos se utilizan estímulos individuales y en otros colectivos –al conjunto de los docentes de acuerdo al resultado general del centro educativo–. En general se construyen índices que incluyen otros indicadores además de los resultados de los alumnos en pruebas estandarizadas. En las Fichas 10 y 11 se analizará con mayor profundidad estos enfoques.

2.6. Rendición de cuentas y responsabilización

Un sexto y último propósito de las evaluaciones educativas es "rendir cuentas" de lo que se ha logrado.

El término "rendición de cuentas" comenzó a ser utilizado en los años 80, pero la idea básica que subyace tiene una larga historia.

Antiguamente los exámenes que rendían los estudiantes eran públicos. Éste era un procedimiento a través del cual, además de evaluar al estudiante, tanto éste como su profesor daban cuenta ante los interesados –las familias, los amigos, los colegas, la sociedad– de lo que el profesor había enseñado y de lo que el alumno había aprendido.

Como resulta obvio, esta práctica se tornó inviable a medida en que los sistemas educativos adquirieron carácter masivo. Con el paso del tiempo y su crecimiento cuantitativo, los

sistemas educativos se tornaron "opacos", en el sentido de que sus resultados dejaron de ser fácilmente "visibles" para la sociedad, –aunque sí "intuibles".

En muchas áreas de la vida de las sociedades como la salud, el empleo, las condiciones de vida de la población, etc., la generación sistemática de información e indicadores permite hacer visibles a los ciudadanos situaciones que generalmente son ya intuidas: la magnitud del desempleo, las condiciones de pobreza, la prevalencia del SIDA, etc. La información permite también apreciar la evolución de estas situaciones a lo largo del tiempo.

En la educación existe una necesidad similar, que tradicionalmente fue satisfecha a través de la información sobre evolución de la matrícula y cobertura alcanzada en distintos niveles etarios, las tasas de reprobación, etc.

Sin embargo, desde las últimas décadas del siglo XX, una de las preocupaciones principales ha pasado a ser qué tanto aprenden los alumnos en su paso por el sistema educativo. Esta pregunta ha pasado a tener una importancia central en un mundo en que el conocimiento y las capacidades de los individuos son cada vez más importantes, tanto para las oportunidades en la vida y el desarrollo de las potencialidades de las personas, como para las posibilidades de los países de desarrollarse en diferentes aspectos: cultural, económico, social.

En este contexto es que surgen con fuerza las evaluaciones estandarizadas, tanto a nivel de cada país como a nivel internacional, uno de cuyos propósitos es justamente hacer "socialmente visibles" los resultados de la labor educativa y permitir al menos dos formas de "rendición de cuentas" complementarias:

- de los profesores y centros educativos ante las familias de sus alumnos, la administración y ante la sociedad en general;

- de la administración educativa ante la sociedad, las familias y el sistema político.

La rendición de cuentas –también denominada "responsabilidad por los resultados"– admite al menos dos enfoques diferentes.

Uno es la **"búsqueda de culpables"**. Las escuelas culpan a las familias, los profesores los alumnos, la administración a los profesores, y los profesores a la administración, por los

resultados insatisfactorios. En esta actitud cada actor busca "deslindar" o eludir su propia responsabilidad y atribuirla a otros. La responsabilidad es de "otro"[3].

Un enfoque completamente distinto parte del supuesto de que los logros educativos de los estudiantes son el resultado de un complejo conjunto de factores –incluido el esfuerzo individual de cada alumno por aprender– y que cada actor tiene la responsabilidad de hacerse cargo de buscar los caminos para mejorarlos dentro del ámbito de decisiones que le competen. **La responsabilidad es compartida** por políticos y administradores, técnicos, directivos y docentes, instituciones formadoras, así como por los propios alumnos y sus familias. Este tema es tratado con mayor detalle en la Ficha 11.

3) Esta "elusión" de la responsabilidad también puede hacerse por la vía de la descalificación genérica de toda evaluación de carácter externo y la caracterización de la educación como algo intangible y no susceptible de medición ni evaluación, con lo cual no hay forma de establecer responsabilidades de ningún tipo.

Síntesis final

Esta primera Ficha tuvo como finalidad introducir al lector en el vasto mundo de las evaluaciones educativas desde la perspectiva de sus propósitos. Se ha intentado mostrar que existen muchas y diversas formas de evaluación en la vida diaria de los sistemas educativos y que cada una de ellas responde a uno o más propósitos o finalidades.

De un modo general, es posible clasificar a las evaluaciones en dos grandes grupos: aquellas que tienen consecuencias directas importantes para individuos o instituciones –los exámenes para aprobar un curso, las pruebas de selección, las evaluaciones que definen una calificación para el maestros o determinan premios en dinero– y aquellas que tienen como propósito principal aprender para mejorar, pero que no tienen consecuencias "fuertes".

En la literatura anglosajona se suele denominar a estos dos tipos de evaluaciones con los términos "*high stakes*" y "*low stakes*", respectivamente (lo que puede traducirse como "altas" y "bajas" implicancias, respectivamente). En la literatura educativa es usual denominar a las primeras como evaluaciones "sumativas" y a las segundas como "formativas" o "diagnósticas".

Es importante comprender que ambos tipos de evaluaciones son relevantes, ya que responden a necesidades de la vida de la sociedad y de los sistemas educativos. Muchas veces se descalifica a las

evaluaciones "sumativas" porque solo tienen en cuenta los resultados, pero no los procesos. Bajo el mismo argumento, se descalifica a las evaluaciones estandarizadas a nivel nacional o internacional. Sin embargo, ambos tipos de evaluación cumplen con un propósito necesario.

Desde una perspectiva opuesta, muchas veces se piensa la evaluación únicamente en términos de evaluaciones con consecuencias directas fuertes y se desconoce el papel de la evaluación como instancia formativa, sin consecuencias directas, cuyo propósito principal es comprender mejor la realidad para ayudar a los individuos y a las instituciones a aprender para realizar mejor su trabajo.

Ambos tipos de evaluaciones son necesarias en distintos momentos y contextos. Ambos tipos de evaluaciones tienen sus ventajas y sus problemas, que serán analizados en las siguientes Fichas.

La evaluación –bien realizada– puede ser una herramienta de cambio de enorme potencial. Si los sistemas educativos mejoraran los distintos tipos de evaluación que ocurren a diario, ello tendría un enorme impacto en el sistema educativo: los alumnos recibirían mejor apoyo en sus procesos de aprendizaje, las evaluaciones de certificación serían más justas y garantizarían que los individuos están preparados para lo que se supone fueron formados, el sistema seleccionaría a los individuos más competentes para desempeñar funciones de conducción y responsabilidad institucional, los profesores y las escuelas aprenderían más de su experiencia, las familias conocerían mejor qué es lo que sus hijos están intentando aprender y qué dificultades tienen, la sociedad en general tendría mayor conocimiento y compromiso con la educación, las políticas educativas podrían estar sustentadas en una base de información más sistemática.

¿CÓMO SE HACEN LAS EVALUACIONES EDUCATIVAS?

Los elementos básicos del proceso de evaluación

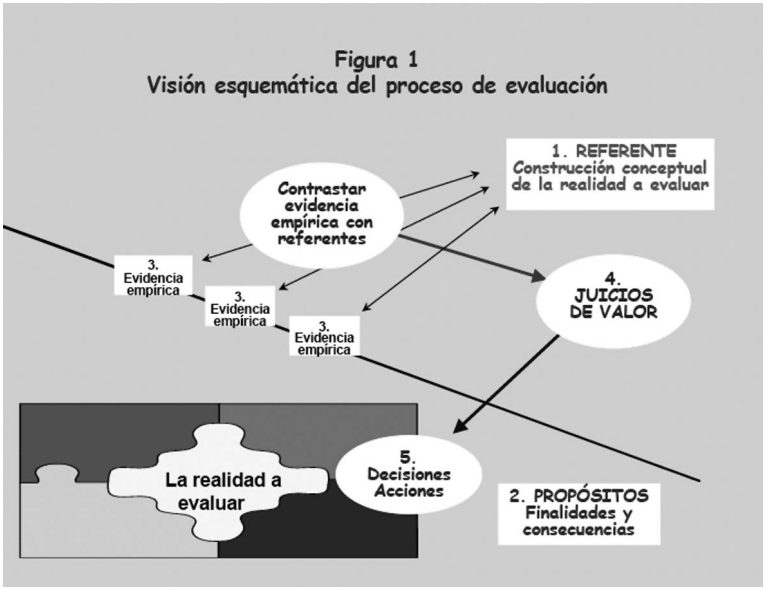
El propósito de esta Ficha es explicar al lector cinco elementos centrales que subyacen a todo proceso de evaluación:

1. la selección de la realidad a evaluar y la "construcción conceptual" de la misma;
2. la definición de los propósitos de la evaluación;
3. la producción de evidencia empírica (información, datos acerca de la realidad evaluada);
4. la formulación de juicios de valor sobre la realidad evaluada;
5. la toma de decisiones o acciones que transformen dicha realidad.

Comprender el significado de estos cinco elementos es fundamental para entender cómo funcionan las evaluaciones, conocer sus posibilidades y limitaciones y leer de manera más inteligente sus resultados.

Los cinco elementos centrales aparecen representados en forma de esquema en la Figura 1.

Figura 1
Visión esquemática del proceso de evaluación



FICHA 2

1. La selección y "construcción conceptual" de la realidad a evaluar

El primer paso obvio de cualquier evaluación es decidir qué se quiere evaluar. En principio esto parece sencillo: queremos evaluar el desempeño de los docentes, la calidad de una universidad o si los estudiantes han aprendido a leer.

Sin embargo, cualquiera de las expresiones anteriores involucra aspectos de la realidad cuyo significado puede ser definido de diversas maneras. Existen diversas perspectivas y posibilidades para definir "desempeño docente", "calidad universitaria" o "capacidad de lectura".

Un error en muchas evaluaciones es dar por supuesto que la realidad está al alcance de la mano, que hay algo predefinido inequívocamente que es "el buen desempeño docente", "la capacidad de lectura" o "la calidad de una institución universitaria" y que, por lo tanto, el problema de la evaluación se limita a diseñar instrumentos y recoger información.

Por el contrario, cualquier "realidad" debe ser definida, es decir, requiere de una "construcción conceptual" que involucra conocimientos, visiones y valores acerca de dicha realidad.

La realidad no es algo de lo que podamos tener conocimiento directo. Es construida por los seres humanos y puede ser percibida y conceptualizada de diversas maneras. Por esta razón, en la Figura 1 está representada con un "rompecabezas" y "separada" de nuestra percepción por una línea negra, que representa la "opacidad" de la realidad.

A la construcción conceptual de la realidad que queremos evaluar se la denomina "referente". El "referente" siempre tiene una connotación valorativa, porque expresa lo deseable o lo que se desea alcanzar. Elaborar y explicitar el "referente" es la primera tarea central de toda evaluación.

Los Recuadros 1 y 2 ejemplifican dos referentes distintos para la evaluación del desempeño docente.

En el Recuadro 1 se incluyen las pautas empleadas en Uruguay para la evaluación del desempeño de los docentes por parte de los Inspectores, de acuerdo a lo estipulado en el Estatuto del Funcionario Docente. Esta pauta, junto con un formulario que incluye un

espacio para que el Inspector registre algunas observaciones o comentarios sobre cada uno de los aspectos anteriores, es todo con lo que cuenta el Inspector para evaluar a cada docente, a través de una única visita al aula para observar una clase.

Como se puede apreciar, se trata de una definición conceptual absolutamente ge-neral y poco precisa. ¿Qué significa, por ejemplo, "posibilidades de desarrollo del trabajo creativo"?

El Inspector debe asignar al docente un puntaje entre 0 y 100. La escala tiene juicios de valor asociados a distintos puntajes (por ejemplo, un puntaje entre 90 y 100 significa "excelente"), pero no existen pautas explícitas que definan qué significa ser "excelente", por ejemplo, en la "capacidad técnico pedagógica". En cierto modo, se da por supuesto que dicha expresión tiene una única interpretación posible que es conocida y compartida por docentes e Inspectores. Como ello no ocurre, los resultados de las evaluaciones que se realizan dependen fuertemente de la subjetividad de cada Inspector.

En el Recuadro 2 se incluye una parte del referente elaborado por el "Educational Testing Service" (ETS) para la evaluación de docentes. Es un referente mucho más elaborado, preciso y explícito, que recoge las investigaciones y enfoques más recientes en relación a la profesión docente y la práctica de la enseñanza.

Este referente se basa en cuatro grandes dimensiones (véase la primera parte del Recuadro 2), cada una de las cuales es luego explicitada a través de entre cuatro y seis criterios, que a su vez explicitan el significado de la dimensión. En la segunda parte del Recuadro 2 se incluyen los criterios correspondientes a las dimensiones 'A' y 'B'.

Pero además de establecer criterios para cada dimensión, cada uno de los 19 criterios resultantes está explicado en un documento mediante tres o cuatro párrafos que los describen y sustentan en la investigación y literatura más recientes en relación al desempeño docente.

Como resulta obvio, un dispositivo de evaluación desarrollado sobre este último referente tiene mucho mayor interés y utilidad que el incluido en el Recuadro 1: el docente tiene más claro sobre qué bases está siendo evaluado, es más fácil construir instrumentos para relevar información cuanto más explícito esté el referente, es más fácil también entrenar a los evaluadores para que las evaluaciones de distintos docentes sean menos subjetivas y más equitativas.

FICHA 2

Recuadro 1 **Pautas para el Informe de Inspección empleado en Uruguay**

Juicios sobre la aptitud docente

1. "Capacidad técnico pedagógica"
2. "Conducción del proceso de enseñanza-aprendizaje"
3. "Orientación dada al curso; planeamiento y desarrollo del mismo"
4. "Aprendizaje realizado por los alumnos y capacidad para seguir aprendiendo"
5. "Clima de trabajo, cooperación e iniciativa"
6. "Respeto al alumno y promoción de su capacidad de autodeterminación"
7. "Posibilidades de desarrollo del trabajo creativo"

Recuadro 2 **El "referente" de desempeño docente empleado por el Modelo de Evaluación PRAXIS[1]**

DIMENSIONES

Organizar los contenidos para el aprendizaje de los alumnos
Crear un entorno para el aprendizaje de los alumnos
Enseñar para que los alumnos aprendan
Profesionalismo Docente

CRITERIOS

A.1. Familiarizarse con los aspectos relevantes de las experiencias y conocimientos previos de los alumnos
A.2. Formular metas claras de aprendizaje que sean apropiadas para los alumnos
A.3. Demostrar comprensión de los vínculos entre el contenido que se aprendió anteriormente, el contenido actual y el contenido que se trabajará más adelante
A.4. Crear o seleccionar métodos de enseñanza, actividades de aprendizaje y materiales u otras fuentes de instrucción apropiados para los alumnos y para las metas de aprendizaje
A.5. Crear o seleccionar estrategias de evaluación apropiadas para los alumnos y para las metas de aprendizaje
B.1. Crear un clima que promueva la equidad
B.2. Establecer y mantener una relación interpersonal adecuada con los alumnos
B.3. Comunicar expectativas de aprendizaje desafiantes a cada alumno
B.4. Establecer y mantener normas coherentes de comportamiento en clase
B.5. Hacer que el entorno físico sea lo más seguro y conducente al aprendizaje que sea posible

1) Fuente: DWYER, C.; "Evaluación de los maestros". En: ALVAREZ, B. y RUIZ-CASARES, M. (editores), 1997; Evaluación y reforma educativa. Opciones de política; Capítulo 7, pp.187-222. PREAL, Santiago de Chile.

Lo mismo ocurre en cualquier otro aspecto de la realidad que se quiera evaluar. Por ejemplo, hay múltiples modos de definir capacidad de lectura, los que pueden ser más o menos generales o específicos y pueden o no recoger los conocimientos más actualizados en el campo del aprendizaje del lenguaje.

En las Fichas 6 y 7 este tema será abordado nuevamente, a través de ejemplos sobre cómo distintas evaluaciones estandarizadas de lectura definen de manera distinta el aprendizaje de los alumnos y utilizan actividades de prueba muy distintas. Por este motivo, los datos que entregan esas evaluaciones tienen sentidos diferentes.

2. La definición de los propósitos de la evaluación

El segundo elemento clave para una buena evaluación es definir claramente los propósitos de la misma: las preguntas que busca responder, el tipo de consecuencias que tendrá la evaluación, los usos que se pretende dar a sus resultados, quiénes harán uso de los mismos. Estos aspectos fueron tratados en la Ficha 1.

Lo que interesa destacar en esta Ficha es que la definición de los propósitos y consecuencias de una evaluación tiene importancia decisiva para el resto del diseño de la misma.

Por ejemplo, si el propósito de una evaluación es calificar a los docentes, y esta calificación tendrá consecuencias para sus carreras funcionales, se requiere una gran confiabilidad y precisión en los puntajes que se les asignen, porque diferencias de puntajes muy pequeñas entre los individuos pueden determinar que unos accedan a un cargo y otros no. En este caso, la precisión de los puntajes es fundamental para que el proceso sea equitativo y para que el sistema educativo se asegure de promover a los profesionales más competentes.

En cambio, si la evaluación tiene como finalidad principal orientar a los docentes para mejorar su trabajo como forma de desarrollo profesional, la asignación de puntajes precisos pasa a tener importancia secundaria, dado que no está en juego la carrera de los individuos.

De la misma manera, una evaluación de alumnos cuyos resultados tendrán consecuencias "fuertes", tales como decidir si aprueban o no un curso, debería restringirse a aquello que fue realmente enseñado durante el curso.

En cambio, si se trata de una evaluación estandarizada de carácter diagnóstico que no tendrá consecuencias para los estudiantes, es legítimo evaluar aspectos relevantes, aun cuando no hayan sido enseñados, justamente para entregar al cuerpo docente una señal en el sentido de que dichos aspectos deberían estar siendo enseñados.

Otro aspecto importante es el relacionado con los destinatarios de la evaluación. Definir con claridad desde el comienzo quiénes utilizarán los resultados de la evaluación, y para qué propósitos, también tiene implicancias importantes para el diseño de la evaluación. Será necesario asegurarse que el tipo de información y los juicios de valor que la evaluación produzca sean apropiados y comprensibles para los destinatarios.

3. La producción de evidencia empírica

No es raro que cuando se piensa en evaluar algo, la primera tarea que se emprenda sea el diseño de instrumentos. Sin embargo, de acuerdo a lo analizado en los apartados anteriores, recién después de haber definido con claridad el referente y los propósitos de la evaluación tiene sentido trabajar en el diseño de instrumentos para recoger información.

El problema central en este terreno es reconocer que no tenemos acceso directo a la realidad. Ante un profesor, no podemos, saber directamente si como profesional es bueno, regular o malo. Del mismo modo, no tenemos acceso directo a la mente y el corazón de los alumnos. Por lo tanto, necesitamos construir "mediaciones" que nos permitan aproximarnos a la realidad que queremos evaluar.

A estas mediaciones se las denomina en forma genérica como "evidencia empírica", entendiendo por tal toda pieza de información que muestra un aspecto de la realidad que nos interesa conocer y evaluar.

Por ejemplo, si queremos evaluar el desempeño de un docente, necesitaremos recoger información sobre aquellos aspectos que hayamos definido como relevantes en nuestro referente: cómo planifica sus clases, cómo explica los temas a sus alumnos, qué actividades les propone que ellos realicen, en qué grado se esfuerza para que los alumnos se interesen en los temas, etc.

Los "instrumentos" de una evaluación son todos aquellos dispositivos construidos para recoger evidencia empírica en forma sistemática sobre los aspectos relevantes de la realidad a evaluar.

Los instrumentos para la recolección de evidencia empírica en las evaluaciones educativas pueden ser muy variados:

- pruebas escritas de diverso tipo (de ensayo, de respuesta construida, de opción múltiple, etc.);
- pruebas prácticas como, por ejemplo, dar una clase o conducir una reunión de profesores;
- registros de observación de diverso tipo, como por ejemplo, de las actividades de los alumnos;
- carpetas con trabajos producidos por los alumnos o los profesores (denominados "portafolios" desde hace algún tiempo en la literatura anglosajona);
- encuestas de opinión de padres, alumnos o profesores.

La amplitud y diversidad de las fuentes de recolección de evidencia empírica es enorme. La enumeración anterior de instrumentos es solamente ilustrativa.

Cuanto más variada sea la diversidad de fuentes de información que se utilice en una evaluación, más sólida será la misma. Sin embargo, al mismo tiempo hay restricciones de tiempo y costos que hacen necesario limitar la cantidad de instrumentos.

Un aspecto muy importante que será tratado con detalle en la Ficha 4 es cuidar la adecuada relación y coherencia entre el referente de la evaluación y los contenidos de los instrumentos de recolección de evidencia empírica. Un problema común en muchas evaluaciones es que se define de una manera la "realidad" a evaluar, pero los instrumentos recogen información sobre otros aspectos no contemplados en el referente. Esto, así como otros errores en el diseño de las evaluaciones y en la interpretación y uso de sus resultados, genera un problema de "validez" en la evaluación.

Otro problema, siempre presente, deriva del hecho de que la realidad es inabarcable, por lo que inevitablemente nuestra información tendrá limitaciones que no hay más remedio que aceptar.

Por ello es importante evitar los simplismos, tanto el de creer que un número puede definir la calidad de una persona o de una institución, como el simplismo opuesto: creer que, como hay limitaciones en la información, solo es válida la observación directa y el

juicio subjetivo, con la consecuente descalificación genérica de las evaluaciones sistemáticas.

De lo antes expuesto se desprende que la selección de los instrumentos a emplear en una evaluación debe ser coherente con el referente y con los propósitos de la evaluación, así como con cierta dosis de sentido común y realismo en cuanto a los recursos, el tiempo y la viabilidad práctica de la evaluación.

Siempre es necesario realizar un balance entre la cobertura más exhaustiva posible de los diversos aspectos de la realidad que se quiere evaluar, la inversión (principalmente de tiempo y de dinero) y el tipo de consecuencias que la evaluación tendrá.

Por ejemplo, puede ocurrir que la evaluación sea tan exhaustiva que el tiempo necesario para el procesamiento de la información y la producción de resultados tenga como consecuencia que éstos lleguen tarde y que la evaluación no tenga impacto porque sus resultados no llegaron en el momento oportuno.

Esto ocurre en las evaluaciones nacionales de carácter estandarizado: muchas veces cuando se presentan los resultados ha pasado tanto tiempo desde que se realizó la evaluación, que los resultados ya no son pertinentes para la toma de decisiones.

Asimismo, los profesores constantemente necesitan hacer en el aula el siguiente balance: si evalúan con demasiada frecuencia, probablemente no tengan tiempo para corregir las evaluaciones que proponen y para hacer devoluciones significativas a sus alumnos.

4. La formulación de juicios de valor

La esencia de la evaluación es establecer un juicio de valor. Este surge principalmente de contrastar la evidencia empírica con el referente para formular valoraciones sobre la realidad.

Ejemplos de juicios de valor son: "este alumno tiene un desempeño satisfactorio en Matemática", "este alumno tiene dificultades en el uso de la puntuación en la producción escrita", "este maestro es excelente enseñando Ciencias". En estos casos se trata de juicios de valor absolutos, formulados en relación al referente de la evaluación. Se los suele denominar como "referidos a un criterio".

Otras veces, los juicios de valor están basados en comparaciones entre individuos o entre instituciones: "estos alumnos son mejores que aquellos", "este alumno está dentro del 10% de estudiantes con mejor desempeño", "éstas son las escuelas con mejores resultados en contextos desfavorecidos". En estos casos se trata de juicios de valor "relativos", porque están basados en la comparación entre unidades de análisis. A este tipo de juicios de valor se suele denominar como "normativo" (por referencia a la curva normal, según se explicará en la Ficha 3).

En ocasiones, los juicios de valor son conjuntos de apreciaciones que no se resumen en un puntaje o una categoría, sino que se utilizan en forma global para orientar el mejor desempeño de un individuo. Es el caso de un profesor orientando a un alumno o de un supervisor orientando a un docente. En estos casos puede haber valoraciones que se expresan en el marco de la interacción directa entre evaluador y evaluado, sin necesidad de resumirlas en un juicio taxativo o en un puntaje.

Al análisis de diferentes modos de establecer los juicios de valor está dedicada la Ficha 3.

Lo central a destacar en este punto es que lo más importante en una evaluación no son los datos en sí mismos, sino la valoración de la realidad que es objeto de evaluación, valoración que se construye a partir de la constrastación entre los datos y el referente.

5. La toma de decisiones o acciones que modifiquen la realidad evaluada

La vocación de toda evaluación es tener alguna consecuencia sobre la realidad. Las consecuencias pueden ser de diverso tipo –Ficha 1– pero una evaluación no es tal si no pretende tener algún tipo de consecuencia, al menos una nueva comprensión de la realidad evaluada que ayude a los actores a pensar en nuevos modos de actuar.

Ninguna evaluación se hace simplemente por curiosidad, sin intención de que sus resultados sean empleados de uno u otro modo para fines específicos.

En la Ficha 1 se hacía una distinción entre las evaluaciones con consecuencias formales y explícitas y evaluaciones formativas sin consecuencias formales, pero que modifican la realidad a través del aprendizaje de los individuos, que logran nuevas visiones sobre su propio desempeño, el de sus alumnos o el de las instituciones de las que forman parte.

En este punto es necesario señalar algunas dificultades.

Cuando una evaluación tiene consecuencias "fuertes", a veces ocurre que se producen efectos no deseados o "efectos perversos".

Por ejemplo, un examen de finalización de la primaria –con consecuencias para los alumnos– cuyo propósito explícito es garantizar que todos logran niveles de aprendizaje similares, puede tener como efecto "perverso" un incremento de los niveles de repetición y deserción si muchos alumnos no consiguen aprobar.

Otro ejemplo podría ser la introducción de incentivos monetarios a las escuelas en función de sus resultados, cuyo propósito explícito es incentivar a todos a mejorar, pero porque puede tener como efecto "perverso" la desmoralización de los "perdedores" y un incremento de la segmentación interna del sistema educativo. Estos aspectos serán tratados con mayor detalle en la Ficha 11.

Cuando las evaluaciones son de carácter formativo, el tipo de consecuencias que tengan dependerá del grado en que los actores se apropien de los resultados y los utilicen para generar nuevas comprensiones sobre la realidad que les permitan aprender y mejorar su modo de actuar. Que ello efectivamente ocurra requiere que se implementen estrategias y acciones específicas de divulgación de resultados (a este aspecto está destinada la Ficha 12).

Una dificultad central en relación al impacto de los procesos de evaluación en la transformación de la realidad es el divorcio que a veces se produce en la práctica entre evaluadores, tomadores de decisiones y quienes se desempeñan como los directivos y docentes.

Muchas veces los evaluadores consideran que su labor termina cuando producen el informe de la evaluación y que la responsabilidad de tomar decisiones y usar los resultados corresponde a otros. Pero, al mismo tiempo, suele ocurrir que los tomadores de decisiones y quienes están en las escuelas –los educadores– no comprenden adecuadamente los resultados, no se interesan por ellos o no los consideran relevantes. Por lo tanto, no los utilizan y la evaluación corre el riesgo de tornarse un ejercicio estéril y de perder legitimidad.

Por ejemplo, es bastante común que las autoridades educativas quieran evaluaciones pero luego no utilicen sus resultados en la toma de decisiones ni emprendan las acciones necesarias para revertir los problemas que la evaluación devela.

Este problema tiene una doble arista. Por un lado, es resultado de que muchas veces prima una lógica político-partidaria y de corto plazo en la toma de decisiones. Al mismo tiempo, obedece a que muchas veces lo que producen los evaluadores no es lo que los tomadores de decisiones necesitan.

La vocación principal de toda evaluación es modificar la realidad, pero la evaluación por sí misma no produce cambios si no hay actores que usen los resultados y tomen decisiones a partir de las valoraciones resultantes de la misma.

En este sentido, es preciso enfatizar que, si bien no son los evaluadores quienes deben tomar decisiones o emprender acciones, sí es su responsabilidad comunicar los resultados de manera apropiada. Esto incluye escribir reportes comprensibles pero, sobre todo, propiciar y participar de instancias de diálogo "cara a cara" con otros actores –políticos, autoridades, docentes, unidades de currículum, formadores de docentes, etc.– ayudando a comprender el significado de los datos y lo que nos dicen sobre la realidad, escuchando las dudas y demandas, ayudando a pensar en términos de alternativas para encarar los problemas.

Síntesis final

Esta segunda Ficha se propuso explicar los cinco elementos centrales de todo proceso de evaluación. Intentó además alertar al lector o usuario de evaluaciones sobre los aspectos clave que es necesario tener presente al diseñar, interpretar o analizar una evaluación. Atender adecuadamente a estos cinco aspectos permitirá mejorar sustancialmente el impacto de las evaluaciones sobre el sistema educativo.

- Toda evaluación debe estar basada en un "referente" explícito, claro y apropiado. Esto significa partir de una definición conceptual de la realidad a evaluar que recoja las perspectivas más actualizadas y que, al mismo tiempo, esté expresado de manera clara y comprensible. De este modo es posible diseñar mejor los instrumentos para la evaluación y, al mismo tiempo, permitir a los usuarios interpretar con más facilidad los resultados de la evaluación y, por tanto, hacer uso de ellos.

- Los "propósitos y consecuencias" de las evaluaciones deben estar definidos desde el comienzo. Esto permite, al diseñar la evaluación, tomar decisiones apropiadas en relación a las consecuencias que se espera tenga la evaluación. El hecho de que los propósitos estén claros desde el comienzo facilita a los diversos actores involucrados saber a qué atenerse y qué pueden esperar de la evaluación. La reflexión inicial sobre los propósitos incluye el análisis de la complejidad de los efectos que la evaluación puede generar, de modo de minimizar efectos no deseados.
- La "evidencia empírica", si bien siempre será limitada, debe intentar cubrir de manera apropiada la diversidad de aspectos de la realidad a evaluar definidos en el referente. La consistencia entre los instrumentos y el referente hace que la evaluación tenga mayor validez y, por tanto, que sus resultados sean interpretables y utilizables. Al mismo tiempo, el uso e impacto de la evaluación mejoran cuando los instrumentos empleados son suficientemente precisos (teniendo en cuenta las consecuencias que la evaluación tendrá) y no dependen excesivamente de la subjetividad del evaluador, es decir, son confiables. Los conceptos de validez y confiabilidad serán analizados en la Ficha 4.
- La esencia de una evaluación es "establecer un juicio de valor" acerca de la realidad evaluada. Esta valoración surge principalmente de contrastar la evidencia empírica con el referente (en la Ficha 3 se analiza con más detenimiento cómo se establecen los juicios de valor). El uso apropiado de los resultados de una evaluación requiere tener conciencia de que los datos no "son" la "realidad", sino una aproximación a ella, por lo cual es muy importante hacer siempre un uso reflexivo y ponderado de los mismos, evitando caer en visiones simplistas. El principal aporte de una evaluación es ayudar a reflexionar y comprender mejor la realidad, con el fin de enriquecer la toma de decisiones. Por el contrario, las visiones simplistas pueden hacer que la evaluación pierda su potencial y conduzca a debates estériles o a decisiones inapropiadas.
- El sentido último de toda evaluación es "provocar cambios en la realidad". Para que las modificaciones se produzcan, es fundamental que los evaluadores consideren como parte de su trabajo –y por tanto del diseño de la evaluación– la divulgación apropiada de los resultados a las audiencias que los utilizarán. Esto requiere una preocupación especial por hacer reportes comprensibles pero, además, invertir tiempo en contactos directos con las diferentes audiencias para explicar los resultados y ayudar a la reflexión acerca de las implicancias de los mismos.

FICHA 3

¿CÓMO SE FORMULAN LOS JUICIOS DE VALOR EN LAS EVALUACIONES EDUCATIVAS?

Evaluaciones normativas, de progreso y criteriales

1. Tres modos de formular juicios de valor

Evaluar consiste básicamente en valorar una realidad (una institución, las competencias de un conjunto de individuos, un sistema educativo) comparando evidencia empírica sistemática con un referente o definición conceptual acerca de lo deseable para dicha realidad. En las evaluaciones educativas hay tres maneras principales de formular juicios de valor.

1.1. Enfoque "normativo"

Un primer enfoque, llamado "normativo", pone el foco de atención en ordenar a los individuos, instituciones o subsistemas –regiones, provincias, estados – evaluados con el fin de compararlos entre sí.

Este enfoque está fuertemente relacionado con pruebas cuyo propósito es la selección. En estas pruebas no importa tanto qué es lo que un individuo específicamente "sabe" o domina, sino si "sabe" más o menos que los otros. El centro de la preocupación no está puesto en describir los conocimientos y competencias de los individuos, sino en conocer en qué lugar del conjunto se encuentra cada individuo –entre los primeros, en el medio, entre los últimos–, justamente porque el propósito es seleccionar a los mejores candidatos.

Los juicios de valor que se formula en este enfoque son del tipo "Juan es mejor que José", "Juan está ubicado dentro del 10% mejor".

1.2. Enfoque "criterial"

Un segundo enfoque consiste en privilegiar la comparación del desempeño de un individuo con una definición clara y precisa de lo que se espera que conozca y sea capaz de hacer en un determinado dominio (por ejemplo, comprensión de textos escritos).

Muchas veces se define distintos niveles de logro en ese dominio (unos más básicos, otros más avanzados) y se busca establecer en qué nivel se encuentra cada individuo.

Los juicios de valor que se formula en este enfoque son del tipo "Juan se encuentra en un nivel avanzado de desempeño en lectura" o "José no alcanza un nivel mínimamente aceptable de desempeño en lectura".

1.3. Enfoque de "progreso"

Un tercer enfoque suele denominarse de "progreso", también de "crecimiento" o "aprendizaje". Este pone el foco en analizar cuánto ha cambiado un individuo, institución o subsistema en relación a un punto de partida o línea de base anterior.

En este caso, lo que interesa comparar es la situación de un individuo o institución con respecto a un momento anterior en el tiempo.

Los juicios de valor que se formula, en este enfoque son del tipo "Juan avanzó –o retrocedió– tanto desde la evaluación anterior"; "José no cambió desde la evaluación anterior".

Es importante notar que el enfoque de "progreso" generalmente opera dentro de un enfoque criterial, es decir, que normalmente lo que se hace es valorar el crecimiento o progreso del individuo dentro del dominio evaluado: "Juan estaba por debajo del nivel aceptable y ahora está en un nivel destacado en Matemática". Sin embargo, puede analizarse el progreso dentro de un enfoque normativo. En estos casos se valora el cambio de posiciones relativas: "Juan antes estaba entre los peores estudiantes del grupo y ahora está en el promedio".

2. Los tres enfoques en el aula

Para comprender las diferencias entre los tres enfoques, veamos cómo funcionan en las evaluaciones que los profesores realizan dentro de las aulas.

Normalmente, cuando un docente evalúa a sus alumnos para calificarlos, es decir, para formular un juicio de valor sobre el desempeño de cada uno de ellos, utiliza en forma combinada los tres enfoques.

Por un lado, tiene en cuenta qué alumnos están mostrando un nivel de desempeño satisfactorio en relación a los propósitos del curso y cuáles no –enfoque criterial–.

Simultáneamente, es probable que el profesor tenga en cuenta el punto de partida de cada uno (enfoque de progreso): "Juan aún no alcanza un nivel satisfactorio, pero empezó muy mal y se ha esforzado enormemente". Por tanto, se le asigna una calificación un poco más alta que la que correspondería. En cambio, Lucía tiene un buen desempeño, pero no se ha esforzado mucho. En realidad podría lograr mucho más. Por tanto, se le asigna una calificación algo inferior a la que le correspondería.

Finalmente, los profesores suelen establecer comparaciones entre sus alumnos para decidir la calificación (enfoque normativo). Por ejemplo, toman las cinco o seis pruebas de carácter destacado y las comparan entre sí. A Clara se le asigna un punto más en la escala de calificación que a José, porque si bien ambas pruebas son igualmente destacadas, la de Clara es un poco mejor.

Este modo de proceder es sabio, porque compensa las debilidades que cada enfoque tiene por separado. Particularmente riesgoso es utilizar el enfoque normativo como único enfoque. En estos casos, se califica a los alumnos teniendo en cuenta únicamente la comparación de los desempeños dentro del grupo de estudiantes, pero no se tiene en cuenta cuáles son los mínimos aceptables. Dicho en otras palabras, si un grupo es malo, un alumno puede ser "el mejor del grupo", pero aun así no alcanzar un nivel de desempeño satisfactorio.

3. La parábola de la montaña

Para comprender mejor las diferencias entre los tres enfoques, resulta útil recurrir a la metáfora de una montaña. Imaginemos a un grupo de niños escalando una montaña. La evaluación consiste en formular un juicio de valor acerca de qué tan bien lo hacen[1].

En la Figura 1 los puntos representan a los individuos. Con el enfoque normativo podemos saber quiénes han llegado más alto y quiénes han quedado más abajo. Pero no sabemos nada acerca de qué tan cerca están de llegar a la cima de la montaña.

Puede ser que todos estén muy abajo o que todos estén muy cerca de la cima. Lo único que sabemos es que unos estaán más alto que otros, pero no qué tan lejos o qué tan

1) La idea original de esta metáfora fue tomada de: Committee for Economic Development, 2000. Research and Policy Committee. "Measuring what matters: using assessment and accountability to improve student learning / a statement by the Research and Policy Committee of the Committee for Economic Development".

FICHA 3

Figura 1 **El enfoque normativo**

Figura 2 **El enfoque criterial**

cerca de la cima o del pie de la montaña están.

Volviendo al ejemplo del aula, un profesor que utilizara un enfoque exclusivamente normativo solamente se preocuparía por calificar a sus alumnos en función de quiénes son mejores y quiénes peores, pero no se estaría preocupando sobre qué tanto han logrado de los aprendizajes esperados para el curso o qué tanto han progresado desde su situación al inicio del curso.

La Figura 2 representa el enfoque "criterial". En este enfoque podemos ver la montaña completa, ya que lo que define a este enfoque es una descripción completa y detallada de los desempeños en el área de conocimientos que será evaluada. El foco de atención está puesto en la altura a la que logró llegar cada individuo.

Además, es posible trazar metas a diferentes alturas, representadas por líneas, que equivalen a los denominados "niveles de desempeño". Como sabemos que no todos llegarán hasta las zonas más altas, se puede establecer una zona a la que se espera que todos los alumnos lleguen (la línea intermedia) y otra zona que indica que quien no ha llegado allí está en serias dificultades (la línea más baja), por lo que necesita apoyo adicional o especial.

La Figura 3 representa el enfoque de "progreso" o "crecimiento". El foco de atención en este caso está puesto en cuánto ha avanzado o retrocedido cada individuo desde donde estaba en un momento anterior. En las Figuras 3 y 4, el rombo representa la situación inicial, y el círculo, la situación final de cada individuo. La línea representa el cambio en el tiempo.

El enfoque de progreso **Figura 3**

Con el enfoque de progreso por sí mismo sabemos cuánto avanzó o retrocedió cada individuo, pero seguimos sin saber a qué zona de la montaña ha llegado cada uno, como en la Figura 3. Puede ser utilizado dentro de un enfoque normativo: la preocupación se centra en observar los cambios de posiciones relativas. Juan, que estaba último, ahora está entre los primeros. Pero también puede emplearse dentro de un enfoque criterial. En este caso, podemos saber, por ejemplo, si Juan pasó de estar en un nivel no aceptable a estar en un nivel aceptable (Figura 4).

El riesgo de utilizar el enfoque de progreso de manera aislada es que dos alumnos pueden haber avanzado lo mismo, pero uno puede estar mucho más cerca de la cima que el otro. Si bien es muy importante tener en cuenta cuánto se ha esforzado cada alumno y cuánto ha progresado desde el punto de partida, se corre el riesgo de que un alumno haya progresado mucho, pero que, sin embargo, se encuentre aún muy lejos de lo esperable.

El enfoque de progreso y criterial **Figura 4**

En principio puede decirse que el juicio de valor en función del avance del individuo es

más razonable en los niveles educativos inferiores (educación inicial, primaria) pero que, sin embargo, se hace más riesgoso a medida que se avanza hacia niveles superiores.

Por ejemplo, es discutible otorgar un título de bachiller o de finalización de la secundaria, que habilita a estudios universitarios, a un estudiante que se ha esforzado mucho, que ha avanzado, pero que aún se encuentra muy lejos de poseer las capacidades mínimas indispensables para desempeñarse en el nivel terciario.

Recuadro 1 **Pruebas normativas y pruebas criteriales**

Se denomina prueba normativa a "aquella que ha sido diseñada para brindar una medida del desempeño que es interpretable en términos de posiciones relativas entre individuos o entidades". En la construcción de estas pruebas se eliminan tanto las preguntas muy fáciles como las muy difíciles. Esto se hace porque, como la finalidad principal es ordenar o comparar a los individuos, las preguntas que mejor sirven a esta finalidad son las de dificultad intermedia. Las preguntas muy difíciles solo las responden unos pocos alumnos, por lo que una prueba con muchas preguntas difíciles no servirá para ordenar a la totalidad de los alumnos. Lo mismo ocurre con una prueba con muchas preguntas fáciles, pero por la razón contraria: la mayoría de los alumnos podrá responder a casi todas las preguntas.

Se denomina prueba criterial a "aquella que ha sido diseñada para brindar una medida del desempeño que es interpretable en términos de un dominio de tareas de aprendizaje claramente definido y delimitado". En la construcción de estas pruebas, las preguntas se seleccionan de modo tal que permitan describir toda la gama de niveles de desempeño posibles, desde los más simples hasta los más complejos. Luego es posible informar qué proporción de los alumnos se encuentra en cada nivel de desempeño. Esto significa que con las pruebas se "dibuja" la montaña entera de los aprendizajes esperados en ciertas áreas, lo cual puede operar como cuadro de referencia para el el cuerpo docente y las familias de los alumnos.

Una prueba criterial puede (o no) incluir la definición de un parámetro o línea de corte que establezca cuál es el puntaje o nivel al que se espera que todos los alumnos lleguen al finalizar un grado o nivel de enseñanza (véase la Ficha 9).

(*) Traducido de LINN, R. & GRONLUND, N., 2000; Measurement and Assessment in Teaching (8ª edición), pp. 42-43. Prentice Hall.

Más problemático aun es el empleo de este enfoque en el nivel terciario, en que la institución formadora debe certificar ante la sociedad que el egresado está apto para desempeñar una determinada profesión. En estos casos debe primar el enfoque criterial, a los efectos de garantizar que el individuo posee las competencias necesarias para desempeñarse en la profesión.

4. Implicaciones para las evaluaciones nacionales e internacionales

¿Cuáles son las implicaciones que los conceptos anteriores tienen en relación a las evaluaciones estandarizadas?

En América Latina, en las últimas décadas, muchos sistemas nacionales de evaluación han trabajado con pruebas construidas desde un enfoque normativo, a pesar de que no es el más apropiado para la finalidad que tienen dichos sistemas. Esto ha ocurrido así, principalmente, por desconocimiento y falta de acumulación conceptual y práctica en el área de la evaluación.

Con dichas pruebas es posible comparar los resultados entre escuelas, así como entre subsistemas provinciales o estaduales, pero no se puede informar qué tan cerca o lejos están los alumnos de lo que se espera de ellos. Uno de los problemas principales en la interpretación de los resultados de las evaluaciones nacionales se produce cuando se mezclan los enfoques normativo y criterial o, más precisamente, cuando se interpreta de manera criterial los resultados de una prueba normativa.

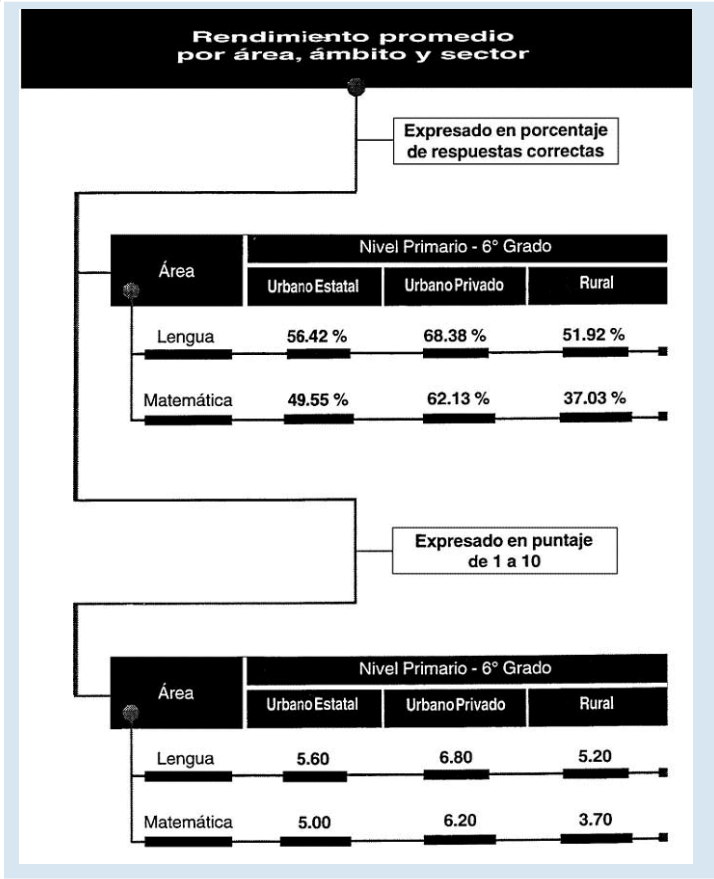
El caso de Argentina algunos años atrás sirve como ilustración de este problema. Se aplicó pruebas que no definían niveles de desempeño ni establecían un nivel de suficiencia, sino que evaluaban los aprendizajes "mínimos". Las pruebas no evaluaban los niveles de desempeño más altos, es decir, no se tenía en cuenta la totalidad de la montaña. Pero los resultados eran presentados –y, por tanto, interpretados por la prensa y el público– como calificaciones (véanse la Figura 5 y el Recuadro 2).

Se interpreta el puntaje 7 como indicador de un resultado satisfactorio, cuando en realidad solo hubiese sido satisfactorio un puntaje de 10, ya que se dice que 10 significa tan solo "dominio de los contenidos mínimos y elementales de una materia". En una prueba normativa los resultados nunca indican el grado de dominio de los aprendizajes, sólo permiten interpretaciones en términos de comparaciones entre individuos o entidades.

FICHA 3

La evaluación internacional PISA es un ejemplo de evaluación estandarizada dentro de un enfoque criterial, en la medida en que se enfatiza la descripción de una escala de desempeños de los estudiantes en Lectura, desde los más complejos –Nivel 5– hasta los más simples –Nivel 1– y se reporta qué porcentaje de los alumnos se encuentra en cada nivel (véase la Figura 6)[2].

Figura 5 **Reporte de resultados en Argentina bajo la forma de porcentaje promedio de respuestas correctas y "calificaciones"**



Fuente: Ministerio de Cultura y Educación de la Nación, Argentina, 1997. Operativo Nacional de Aprendizajes 1996; pág. 30.

2) *En la ficha 7 se incluye una descripción más detallada de los niveles de desempeño de Lectura en PISA.*

Recuadro 2 " La nota no llega a 7"

"Un 6,69 en Lengua y un 6,70 en Matemática son las notas promedio que los alumnos que el año último terminaron el secundario obtuvieron en las pruebas tomadas por el Ministerio de Educación para medir la calidad de la enseñanza.
En la última evaluación, la nota subió a 6,7, tanto en Lengua como en Matemática.
En séptimo grado, la nota en Lengua fue 6,6, un punto por encima del resultado de 1997.
En Matemática, los alumnos de séptimo grado obtuvieron un 6, cuando la nota había sido 5,4 en 1997.
A pesar de la mejora, los resultados siguen siendo poco alentadores: en ningún caso el promedio nacional llegó al 7, que por lo general es el mínimo necesario para aprobar una materia en el colegio.
Además, el 10 no significa, en el criterio ministerial, "excelencia", sino únicamente que el alumno domina los contenidos mínimos y elementales de cada materia".
(La Nación Line, Argentina, 28/05/99)

Figura 6 Niveles de desempeño en lectura en la evaluación o "prueba" PISA

Nivel 5

Los estudiantes situados en el Nivel 5 de la escala combinada de lectura son capaces de llevar a cabo tareas lectoras sofisticadas, tales como:

- manejar información difícil de encontrar en textos con los que no están familiarizados;
- mostrar una comprensión detallada de tales textos e inferir qué información del texto es relevante para la tarea;
- valorar críticamente y elaborar hipótesis, basándose en conocimientos especializados, así como incluir conceptos que pueden ser contrarios a las expectativas.

Nivel 4

Los estudiantes que se sitúan en el Nivel 4 son capaces de solucionar tareas lectoras complejas, tales como localizar información entremetida en el texto, reconstruir el significado a partir de los matices del lenguaje y valorar críticamente un texto.

Nivel 3

Los estudiantes que se sitúan en el Nivel 3 son capaces de resolver tareas lectoras de complejidad moderada, tales como localizar múltiples elementos de información, establecer conexiones entre las diferentes partes de un texto y relacionar el texto con los conocimientos cotidianos habituales.

Nivel 2

Los estudiantes que se sitúan en el Nivel 2 son capaces de solucionar tareas lectoras básicas, tales como localizar información presentada directamente, llevar a cabo diversos tipos de inferencias de bajo nivel, deducir el significado de una parte bien definida del texto y utilizar algunos conocimientos externos al texto para comprenderlo.

Nivel 1

Los estudiantes situados en este Nivel sólo son capaces de realizar correctamente las tareas lectoras menos complejas desarrolladas para el proyecto PISA, tales como localizar un único elemento de información, identificar el tema principal de un texto o hacer una conexión sencilla con los conocimientos cotidianos.

Fuente: Ministerio de Educación, Cultura y Deporte de España-Instituto Nacional de Calidad y Evaluación (INCE)/OCDE, 2001; Conocimientos y destrezas para la vida: Primeros Resultados del Proyecto PISA 2000. Resumen de Resultados. INCE, Madrid.

FICHA 3

La información acerca de los porcentajes de alumnos en cada nivel de desempeño es más útil que los promedios, dado que éstos últimos no dicen nada acerca de lo que aquéllos conocen y son capaces de hacer. Por esta razón, la información sobre niveles de desempeño resulta más útil para los docentes y responde mejor a las demandas de la opinión pública.

Es preciso notar que si bien PISA es una evaluación criterial, dado que "dibuja" toda la montaña y establece distintas metas o niveles en ella, no define qué nivel deberían alcanzar todos los estudiantes. Esto es razonable, dado que se trata de una evaluación que involucra a varias decenas de países. La tarea de establecer qué nivel deberían alcanzar los alumnos queda librada a los análisis que cada país realice y a las metas que quiera plantearse.

Un tercer ejemplo es el caso del SAEB –el sistema nacional de evaluación de Brasil– que es particularmente interesante porque construye una descripción de niveles similar a la de PISA –aunque no tan rica– a la que se le agrega un juicio de valor respecto a qué niveles son aceptables y cuáles no.

Es decir, además de describir los diversos niveles de desempeño, el SAEB establece un juicio de valor acerca de cuál es el nivel adecuado que deberían alcanzar los alumnos al finalizar, en este caso, la educación media (véase la Figura 7).

La decisión acerca de cuál es el nivel de desempeño que deberían alcanzar todos los alumnos al final de un grado o ciclo es materia siempre debatible y sólo puede realizarse a través de lo que se denomina la opinión de "jueces". Este tema será tratado en detalle en la Ficha 9.

Una implicación final relevante de los enfoques planteados, en relación a las evaluaciones nacionales, tiene que ver con el enfoque de "progreso". Lo característico de este enfoque es que requiere realizar varias mediciones en el tiempo, (al menos dos, por ejemplo, al inicio y al final del año lectivo). Sólo así es posible evaluar específicamente qué han aprendido los estudiantes al cabo de un cierto período de tiempo, durante el cual han participado en un conjunto de experiencias educativas.

En estos casos se puede hablar con propiedad de evaluación de "aprendizaje" –en el sentido de cambio–. Cuando se evalúa únicamente el resultado al final de un grado o ciclo, el término adecuado es evaluación de "logro".

Niveles de desempeño en Matemática en SAEB – 2001 *Figura 7*

Estágio	População	%
Muito Crítico	99.969	4,84
Crítico	1.294.072	62,60
Intermediário	549.306	26,57
Adequado	123.800	5,99
Total	**2.067.147**	**100,00**

Fonte: MEC/Inep/Daeb

Muy crítico
No logran responder a consignas operacionales elementales compatibles con el 3er. grado de Educación Media (construcción, lectura e interpretación gráfica; uso de propiedades de figuras geométricas planas y comprensión de otras funciones). Los alumnos en este nivel alcanzaron, como máximo, el nivel 3 de la escala SAEB.

Crítico
Desarrollan algunas habilidades elementales de interpretación de problemas, pero no logran traducir lo que está siendo pedido en el enunciado a un lenguaje matemático específico, estando, por lo tanto, por debajo de lo exigido para el 3er. grado de la EM (construcción, lectura e interpretación gráfica; uso de algunas propiedades y características de figuras geométricas planas y resolución de funciones logarítmicas y exponenciales). Los alumnos, en este nivel, alcanzaron los niveles 4 ó 5 de la escala SAEB.

Intermedio
Presentan algunas habilidades de interpretación de problemas. Hacen uso de lenguaje matemático específico, pero la resolución es insuficiente en relación a los que es exigido para el 3er. grado de EM (reconocen y utilizan algunos elementos de geometría analítica, ecuaciones polinómicas y reconocen algunas operaciones con números complejos). Los alumnos en este nivel alcanzaron los niveles 6 ó 7 de la escala SAEB.

Adecuado
Interpretan y saben resolver problemas de forma competente; hacen uso correcto del lenguaje matemático específico. Presentan habilidades compatibles con el grado en cuestión (reconocen y utilizan elementos de geometría analítica, ecuaciones polinómicas y desarrollan operaciones con números complejos). Los alumnos en este nivel alcanzaron los niveles 8, 9 ó 10 de la escala SAEB.

Fuente: Instituto Nacional de Estudios y Pesquisas Ed ucacionales (INEP), 2002; Relatório SAEB 2001 - Matemática. Brasilia, Ministerio de Educación.

FICHA 3

En las evaluaciones estandarizadas, el enfoque de progreso tiene una virtud fundamental: es el único que permite analizar cuál es el efecto de la labor de los docentes y centros educativos en el aprendizaje de los alumnos, dado que tiene en cuenta el punto de partida en que estaban estos antes de que el profesor o el equipo docente realizara su labor de enseñanza.

Este aspecto será tratado con mayor detalle en la Ficha 13, pero cabe adelantar que un juicio de valor acerca de la "eficacia" o "calidad" de un docente, un centro educativo o un subsistema que tome como indicador únicamente el resultado final, sin considerar el punto de partida, es poco apropiado, porque las poblaciones estudiantiles varían enormemente en cuánto a su capital cultural y su acumulación de conocimientos previos.

Un ejemplo de evaluación nacional que se aproxima de alguna manera a un enfoque de "progreso" es el caso de Chile (véase la Figura 8), en que se reporta cuánto cambiaron los puntajes de cada escuela entre evaluaciones sucesivas y comparables.

La debilidad en este caso es que en realidad los alumnos no son los mismos, por lo que los cambios de puntajes en las escuelas pueden ser resultado de cambios en la selección de los alumnos por parte de la escuela en el lapso transcurrido entre las dos evaluaciones o, también, de cambios socioculturales en el perfil del estudiantado de un centro educativo derivados de cambios sociales en su entorno.

Figura 8 **Reporte de progreso de los establecimientos educativos en Chile**

RESULTADOS POR ESTABLECIMIENTO

	ALUMNOS EVALUADOS	LENGUA CASTELLANA Y COMUNICACIÓN	MATEMÁTICA
2° MEDIO A	35	250	250
2° MEDIO B	30	250	250
ESTABLECIMIENTO	65	250	250
VARIACIÓN EN RELACIÓN A 1998		▲ +1	▼ -2
DIFERENCIA CON EL PROMEDIO DE SU GRUPO SOCIOECONÓMICO		● +1	▲ +2
DIFERENCIA CON EL PROMEDIO COMUNAL		▼ -5	● +1
DIFERENCIA CON EL PROMEDIO REGIONAL		● +2	▼ -5
DIFERENCIA CON EL PROMEDIO NACIONAL		● +3	▲ +3
PUNTAJE MÁXIMO NACIONAL		350	376
PUNTAJE MÍNIMO NACIONAL		182	182
GRUPO SOCIOECONÓMICO DEL ESTABLECIMIENTO	E		

Fuente: Ministerio de Educación de Chile/SIMCE, 2002; Informe de Resultados 2001. 2° medio.

Para aplicar efectivamente el enfoque de "progreso" se requiere evaluar dos veces en el tiempo a los mismos alumnos, por ejemplo, al inicio y al final del año lectivo, o al final y al inicio de un ciclo completo, lo cual requiere identificarlos y efectuar un "seguimiento longitudinal" de la misma cohorte de estudiantes.

Síntesis final

El sentido último de toda evaluación de aprendizajes o logros educativos es formular un juicio de valor acerca de lo que los niños y jóvenes están aprendiendo, con el fin de propiciar transformaciones y mejoras en el sistema educativo. Los juicios de valor pueden ser formulados con énfasis diversos: comparando a los alumnos entre sí (enfoque normativo), comparando el desempeño de cada alumno con una definición clara del desempeño esperado (enfoque criterial), comparando el desempeño de cada alumno con su propio desempeño en un momento anterior en el tiempo (enfoque de progreso).

En sus inicios, la mayor parte de los sistemas nacionales de evaluación de aprendizajes en América Latina adoptaron un enfoque normativo en la construcción de sus pruebas. Esto ocurrió debido a la insuficiente acumulación de conocimiento sobre evaluación en la región y a la consiguiente adopción a crítica de metodologías propuestas por consultores.

Sin embargo, este enfoque no permite dar respuesta a las preguntas principales que la opinión pública, los docentes, las familias y los responsables de la conducción educativa esperan ver respondidas: ¿en qué grado los alumnos están logrando lo que se espera que hayan aprendido al finalizar un determinado ciclo escolar? y ¿qué políticas y prácticas educativas son las que en mayor grado aseguran que la mayor parte de los alumnos alcancen dichos aprendizajes, aun cuando provengan de sectores sociales desfavorecidos?

Para poder dar respuesta a estas preguntas es necesario adoptar una combinación de los enfoques criterial y de progreso en el diseño de los sistemas de evaluación. Sólo a partir de descripciones claras, completas y detalladas de lo que se espera que los alumnos aprendan, junto con pruebas apropiadas, es

posible responder a la primera pregunta. Y solo a partir de sistemas que evalúen los progresos de los alumnos en el tiempo es posible identificar de manera apropiada cuáles son las políticas y prácticas educativas que mejor contribuyen a dichos logros, es decir, responder a la segunda pregunta.

En los últimos años, buena parte de los sistemas nacionales de evaluación de la región han comenzado a moverse en esta dirección. Chile, México, Perú, República Dominicana, y Uruguay son algunos casos. Por lo tanto, es esperable que estos sistemas progresivamente aporten información más relevante, y que su contribución general a las políticas y prácticas educativas mejore sustantivamente en los próximos años, siempre y cuando exista continuidad en estos procesos de cambio y en la consolidación de las nuevas políticas de evaluación.

¿CUÁLES SON LOS PRINCIPALES PROBLEMAS COMUNES A TODAS LAS EVALUACIONES EDUCATIVAS?

Validez y confiabilidad

Esta Ficha tiene como objetivo ayudar al lector a desarrollar su capacidad para leer críticamente las evaluaciones.

La formulación de juicios de valor es el centro de la evaluación. Pero, como se mostró en las Fichas anteriores, esta no es una actividad objetiva y aséptica, en la medida en que:

a. intervienen valores y visiones del mundo y de la realidad evaluada que son construidas por los evaluadores e, idealmente, "concertadas" con otros actores;

b. los datos y percepciones que poseemos acerca de la realidad evaluada son siempre aproximaciones parciales a la misma.

Por lo tanto, todo lector inteligente de evaluaciones debería, antes de aceptar las conclusiones y valoraciones resultantes, mirar con ojo crítico y particular cuidado el modo en que la evaluación fue realizada.

1. Validez y confiabilidad

Se puede agrupar los principales tipos de problemas que debe enfrentar cualquier evaluación en torno a dos conceptos: validez y confiabilidad.

El concepto de validez refiere al grado en que los juicios de valor que se formulan en la evaluación están adecuadamente sustentados en evidencia empírica y están efectivamente relacionados con el "referente" definido para la evaluación.

El concepto de confiabilidad refiere a la precisión de las medidas y de la evidencia empírica empleada en la evaluación.

Dada la abstracción de estos dos conceptos para quien no pertenece a los campos de la evaluación o la investigación, comenzaremos por ejemplificarlos, para luego presentar una definición más elaborada.

2. Ejemplos de problemas de validez

2.1. La prueba no evalúa lo que se supone debe evaluar

Este es un tipo de problema bastante común en diversos tipos de pruebas.

> *2.1.a. Un curso de Historia puede tener como propósito lograr que los alumnos desarrollen su capacidad para analizar críticamente los factores sociales, políticos y económicos que incidieron en la generación de ciertos acontecimientos históricos. Sin embargo, luego se los examina con una prueba de ensayo que fundamentalmente requiere de la memorización de acontecimientos, datos y fechas, así como de la capacidad para organizar un* ***relato*** *escrito de ellos con una prosa adecuada.*

En este caso, la prueba no recoge evidencia empírica suficiente y apropiada para determinar si los estudiantes adquirieron las capacidades que fueron definidas como propósito del curso.

> *2.1.b. Una actividad de Matemática tiene un alto contenido de consignas verbales. Los alumnos que tienen menor competencia para la lectura no comprenden lo que se les está pidiendo, por lo que sus resultados son malos no porque no sepan razonar matemáticamente, sino porque no entienden "de qué se trata". En cambio, los alumnos con mayor habilidad para la lectura tendrán más posibilidades de resolver la actividad.*

En este caso, la actividad no está evaluando lo que se supone debe evaluar. Evalúa lectura antes que capacidades o conocimientos matemáticos.

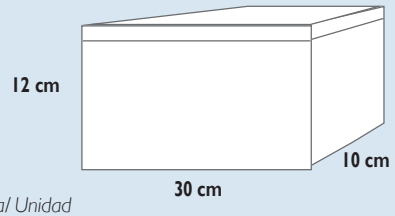
> *2.1.c. En la Figura 1 se puede observar otro ejemplo de una actividad de Matemática que no evalúa lo que dice evaluar.*

Se trata de un ítem que, teóricamente, está dirigido a evaluar la capacidad del alumno para resolver problemas. Sin embargo, si se lo analiza detenidamente se puede constatar que

Una actividad de "resolución de problemas" ***Figura 1***

El volumen de esta caja de zapatos es:

a) 3,6 cm^3
b) 300 cm^3
c) 360 cm^3
d) 3.600 cm^3



Fuente: Administración Nacional de Educación Pública/ Unidad de Medición de Resultados Educativos, 2000. Evaluaciones Nacionales de Aprendizajes en Educación Primaria en el Uruguay. 1995-1999. UMRE, Montevideo.

en realidad no hay ningún "problema" a resolver. Para llegar a la respuesta correcta es necesario, en primer lugar, que el alumno haya memorizado la fórmula de cálculo del volumen de un poliedro. En segundo lugar, el alumno necesita ser capaz de multiplicar números enteros.

Tal como está planteada, la actividad solo puede ser considerada un "problema" desde una concepción muy simplista y escolar de lo que es "resolución de problemas".

Para estos casos, en que no existe consistencia entre el instrumento empleado para recoger evidencia empírica (la prueba) y el referente de la evaluación (aquello que ha sido definido como capacidades o aprendizajes esperables en los alumnos), se utiliza técnicamente la expresión "validez de constructo".

2.2 La prueba no abarca adecuadamente lo que los estudiantes deberían haber aprendido

Una prueba puede ser coherente con la definición de qué se espera que los alumnos aprendan pero no cubrir adecuadamente los diferentes aspectos definidos en el referente. Los siguientes son algunos ejemplos.

2.2.a. Una prueba tiene como objetivo evaluar el dominio del currículo de Ciencias vigente para un determinado grado, pero sus actividades sólo cubren la cuarta parte de

los conocimientos científicos relevantes que los alumnos deberían dominar e ignora los restantes.

2.2.b. Otro ejemplo de este tipo de problema es la antigua práctica de "sortear" temas para en examen a través de un "bolillero" o "balotario". En este caso, se sorteaba un tema de todos los que el alumno debería dominar y el alumno debía exponer exclusivamente sobre ese tema.

El resultado de un alumno en situaciones de evaluación como las descritas en los ejemplos está fuertemente determinado por el azar.

Cuando la prueba no contempla en forma al menos aproximada los diferentes contenidos del curso, un alumno que estudie y aprende la mitad de los temas tiene un 50% de probabilidades de obtener un resultado excelente, y la misma probabilidad de ser reprobado con una pésima calificación.
Para estos casos, en que la prueba no cubre adecuadamente la diversidad de conocimientos y competencias definidos en el referente, se suele utilizar la expresión "validez de contenido".

Este tipo de problemas son en cierto modo inevitables, dado que toda prueba se realiza en un lapso limitado, por lo que difícilmente puede cubrir la totalidad de los conocimientos y competencias trabajadas en un curso.

Sin embargo, existen procedimientos para enfrentar el problema.

En el caso de las pruebas estandarizadas, se busca, en primer término, que las actividades de la prueba sean una buena muestra de la diversidad de contenidos y competencias, que se pretende evaluar y se prioriza aquellos que son considerados más relevantes.

En segundo término, cada vez más se utiliza simultáneamente varias pruebas diferentes que se distribuyen aleatoriamente entre los alumnos. Estas pruebas contienen algunos bloques de actividades en común y otros bloques de actividades que son diferentes. Este procedimiento técnico permite ampliar la muestra de conocimientos y competencias que son evaluados, sin perder la posibilidad de comparar puntuaciones entre los estudiantes.

El hecho de que no todos los alumnos realicen la misma prueba no es relevante en este caso, porque el propósito de una evaluación de este tipo no es establecer comparaciones

entre alumnos individuales pero, además, porque se han desarrollado procedimientos estadísticos para otorgar a los alumnos puntajes comparables, independientemente de qué conjunto de actividades hayan respondido (véase en la Ficha 8 la "teoría de respuesta al ítem").

En el caso de las evaluaciones de certificación que realizan los docentes en los centros educativos, el problema se resuelve por la vía de no hacer depender el resultado final de un alumno de una sola prueba o examen, estableciendo en cambio un sistema de evaluación que incluye varias pruebas y, además, otro tipo de actividades y productos que el alumno realiza durante el curso.

Los ejemplos analizados hasta el momento (1.1 a 2.2) implican una primera advertencia para el usuario de evaluaciones: es necesario analizar el grado de consistencia entre lo que la evaluación se propuso evaluar –el referente– y los instrumentos empleados para ello.

Realizar directamente este análisis normalmente no está al alcance del lector. Por lo tanto, los reportes de las evaluaciones deberían incluir información técnica explícita acerca de los procedimientos seguidos para minimizar estos problemas.

Del mismo modo, los centros educativos deberían hacer explícita a alumnos y familias la manera en que su sistema de evaluación busca asegurar la coherencia con los objetivos de los cursos y los propósitos educativos de la institución.

2.3. El dispositivo de evaluación no es apropiado para predecir el desempeño futuro de los evaluados

Algunas evaluaciones tienen como propósito principal seleccionar personas para desempeñar determinados cargos en la estructura educativa o estudiantes para acceder a determinados programas educativos.

Un caso particularmente importante lo constituyen las evaluaciones que se realizan en el sistema educativo para seleccionar directivos o supervisores, normalmente mediante concursos que, además de pruebas, utilizan otras fuentes de evidencia empírica tales como los antecedentes académicos y funcionales de los candidatos.

En este caso los resultados de la evaluación deben tener la capacidad de anticipar dos cosas principales:

a. qué candidatos tienen las aptitudes imprescindibles para el cargo y cuáles no;

b. qué candidatos tienen más aptitudes y cuáles menos; es decir, se espera que la evaluación prediga quiénes serán mejores en el desempeño de los cargos, a efectos de que tengan prioridad para acceder a los mismos.

Este tipo de procesos de evaluación suele estructurarse en torno a un análisis de los méritos –la formación y títulos alcanzados por cada candidato, su trayectoria en el sistema educativo, sus publicaciones, etc.– y a la realización de varias pruebas, algunas de carácter teórico y otras de tipo práctico como, por ejemplo, conducir una reunión de docentes o analizar una clase dictada por un profesor.

Del conjunto de los elementos anteriores suele derivarse un puntaje final, que es el que determina el ordenamiento de los candidatos y sus posibilidades de acceder a los cargos disponibles.

Este tipo de evaluaciones suele tener tres debilidades principales:

2.3.a. En primer lugar, no suele existir un referente explícito para la evaluación, es decir, una descripción elaborada y apropiada de qué tipo de conocimientos y competencias se requiere para desempeñar el cargo de director de una escuela, cuáles son más importantes que otras, cuáles son los niveles básicos e imprescindibles de competencia y cuáles serían niveles destacados de competencia. Como consecuencia de ello, las pruebas suelen ser elaboradas y evaluadas a partir de la visión personal e implícita que los miembros del tribunal o jurado encargado de la evaluación tienen acerca de los temas anteriores.

2.3.b. En segundo lugar, muchas veces estas evaluaciones no establecen con claridad qué aspectos son prioritarios para el desempeño del cargo. Por ejemplo, suele tener mucho más peso en el puntaje final la antigüedad del individuo en el sistema, su desempeño en una prueba teórica sobre educación o la acumulación de certificados de participación en seminarios, cursos y talleres, que las habilidades del candidato relacionadas con la gestión de una organización compleja o sus capacidades relacionadas con las relaciones humanas y el liderazgo.

2.3.c. En tercer lugar, dada la cantidad de candidatos a evaluar, normalmente en estas evaluaciones intervienen varios tribunales o jurados diferentes, pero no existen procedimientos explícitos para garantizar la máxima consistencia posible entre estos diversos evaluadores. Por lo tanto, el resultado de un candidato suele depender del tribunal o jurado que le toque en suerte –éste es también un problema de confiabilidad que analizaremos más adelante en esta Ficha–.

Como resultado de la acumulación de los problemas anteriores, normalmente estas evaluaciones no consiguen su propósito de identificar a los candidatos más aptos para desempeñar cargos de responsabilidad y ordenarlos de manera más o menos adecuada a su capacidad para desempeñar el cargo. A este tipo de problemas se le denomina técnicamente como de "validez predictiva".

El ejemplo que acabamos de utilizar pone de manifiesto un problema particularmente grave de los sistemas educativos: por lo general, éstos carecen de mecanismos apropiados para seleccionar válidamente a quienes desempeñarán cargos de conducción –así como también a los formadores de docentes–, con lo cual se generan diversas dinámicas perversas: muchos directores no son reconocidos técnicamente por sus docentes y muchos supervisores no son los docentes más competentes, con lo cual las cadenas de autoridad, de transmisión de conocimiento práctico y de aprendizaje institucional se debilitan. Simultáneamente, muchos individuos competentes no son seleccionados o desisten de presentarse a estos procesos de selección, con lo cual el sistema desperdicia talento, conocimiento y capacidades.

2.4. Los usos o consecuencias de la evaluación van más allá de lo que los resultados permiten

En los últimos años se han incrementado las propuestas para establecer incentivos económicos para las escuelas o los docentes individuales, en función de los resultados de sus alumnos medidos a través de una prueba nacional estandarizada.

La versión simple de estas propuestas adolece de serios problemas. En particular, implica calificar la calidad de los centros educativos a partir de evidencia empírica muy limitada: los resultados de una generación de alumnos en pruebas de Lenguaje y Matemática, por ejemplo. Como resulta obvio, lo que las familias y la sociedad esperan de los centros educativos es mucho más que esto. Al focalizar los incentivos en un tipo limitado de

resultados lo que se consigue es propiciar la reducción del abanico de prioridades de las escuelas a mejorar sus resultados en las pruebas estandarizadas. Este es un "efecto perverso" o no deseado del uso de la evaluación.

Este tipo de casos, en que se pretende hacer un uso de los resultados de una evaluación que va más allá de lo que la evaluación permite y de aquello para lo cual fue diseñada, constituye lo que técnicamente se denomina problemas de "validez de uso" o, también, "validez de consecuencias".

La validez de consecuencias alerta al lector sobre la necesidad de analizar la consistencia entre los propósitos para los cuales fue diseñada una evaluación y los usos que se hace de sus resultados. Y también, invirtiendo los términos, llama la atención de quienes encargan o diseñan un sistema de evaluación hacia la necesidad de definir clara y explícitamente cuáles son sus propósitos, qué tipo de decisiones se pretende tomar, para luego establecer un diseño de la evaluación acorde con ellos.

2.5. La situación en que se desarrolla la prueba afecta el desempeño de los individuos

Un último tipo de amenazas a la validez de las evaluaciones educativas tiene relación con el grado en que el desempeño de los individuos en una prueba se ve afectado por las condiciones de aplicación de la misma.

2.5.a. Ejemplos de este problema son, en primer lugar, las situaciones de examen tradicional en que el desempeño de los individuos se ve fuertemente afectado por nervios o angustia ante la situación de evaluación.

2.5.b. En segundo término, las evaluaciones estandarizadas que no tienen consecuencias para los individuos que las realizan –es decir, por ejemplo, que no formarán parte de alguna calificación necesaria para ser promovido a otro grado, ciclo o nivel educativo o profesional– conllevan el riesgo de que los alumnos no realicen todo el esfuerzo de que son capaces, por lo que los resultados casi seguramente serán algo inferiores a lo que realmente son capaces de lograr los estudiantes. Este problema es particularmente importante en la educación media, cuando las pruebas son respondidas por adolescentes.

2.5.c. En tercer lugar, es preciso mencionar los casos en que los instrumentos de

evaluación tienen "sesgos", en el sentido de que favorecen el desempeño de ciertos grupos. Por ejemplo, las actividades de una prueba pueden resultar más motivadoras para las niñas que para los varones o pueden contener situaciones más familiares para los niños y niñas de medios urbanos que para los de medios rurales.

En estos casos, se habla técnicamente de "validez de las condiciones de aplicación".

3. El concepto de validez

El elemento común a todos los ejemplos analizados hasta el momento es que se trata de situaciones en que la evaluación no evalúa realmente aquello que se propuso evaluar o en que el uso de sus resultados va más allá de lo que la evaluación permite.

Ninguna evaluación está exenta de este tipo de problemas, pero todas deben dar cuenta de las acciones tomadas para minimizarlos. Y los involucrados en un proceso de evaluación –quienes la encargan, quienes la llevan adelante, quienes son evaluados, quienes usan los resultados o se informan de ellos– deben estar alertas a estas "amenazas a la validez".

Si bien hasta el momento hemos tratado a la validez como una propiedad de las evaluaciones, las elaboraciones más recientes del concepto tienden a plantearlo en términos de una propiedad de las interpretaciones y usos que se hacen de los resultados de una evaluación.

"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 uso legítimo 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].

En términos de los ejemplos que acabamos de presentar, este giro en el enfoque implica poner la atención en el grado en que las interpretaciones y consecuencias de una evaluación son apropiadas, dadas la evidencia empírica y la teoría disponibles.

1) La deficinión corresponde a Gilbert Valverde (2001); "La interpretación justificada y el uso apropiado de los resultados de las meiciones". En Ravela, P. (editor); Los Próximos Pasos:¿Hacia dónde y como avanzar en la evaluación de aprendizajes en América Latina?. PREAL/GTEE

En el ejemplo del proceso de selección de directores, el problema no es que las pruebas y la evaluación de méritos sean malas en sí mismas. El problema es si la decisión de seleccionar a los directores tiene sustento suficiente en el conjunto de evidencia empírica utilizado para ello.

En el ejemplo de la prueba de ensayo en Historia, no es que la prueba en sí misma sea mala, sino que no es posible interpretarla como evidencia de logro de los objetivos explícitos del curso.

La importancia de este cambio de perspectiva radica en que enfatiza la responsabilidad que los evaluadores y usuarios de las evaluaciones tienen en cuanto al uso apropiado de las mismas, en lugar de limitar el tema de la validez a un problema técnico de los instrumentos.

Los docentes, los técnicos, los formuladores de políticas, los periodistas y los ciudadanos, tienen la responsabilidad de analizar y preguntar por el grado en que el uso y consecuencias de una evaluación tienen un sustento adecuado.

4. Confiabilidad

La confiabilidad de una evaluación refiere a la consistencia y precisión de sus resultados.

A continuación se proponen algunos ejemplos de problemas de confiabilidad en evaluaciones.

4.1. Los resultados de una prueba dependen de la subjetividad de los evaluadores

Este es un tipo de problema de confiabilidad muy extendido en las evaluaciones educativas, dado que en muchas de ellas inevitablemente debe intervenir el juicio subjetivo de individuos que actúan como evaluadores.

4.1.a. Es sabido que el resultado de un alumno en una prueba escrita aplicada y corregida por su maestro puede estar influido por el momento en que su prueba es corregida, al inicio, en el medio o al final del proceso de corrección.

Dependiendo del maestro, el cansancio puede operar en la dirección de tender a asignar calificaciones más bajas al final, tanto como a tornarse más benevolente y tender a asignar calificaciones más altas.

Independientemente del problema del cansancio, como el docente generalmente corrige sin criterios o estándares claros y detallados, sino más bien de tipo holístico y subjetivo, a medida en que corrige producciones de distintas calidades sus criterios se van modificando en el proceso, y normalmente no hay tiempo para volver atrás y recalificar todo con criterios homogéneos.

> *4.1.b.. Otro caso típico está constituido por todas aquellas evaluaciones en las que intervienen diversos evaluadores. Por ejemplo, cuando varios tribunales se conforman para corregir una misma prueba en un concurso para directores, o cuando en pruebas estandarizadas con preguntas abiertas es necesario recurrir a correctores para codificarlas.*

En estos casos, es necesario establecer procedimientos de control de la confiabilidad de las puntuaciones otorgadas por los diferentes correctores.

Por ejemplo, en la prueba internacional PISA se apartan cien ejemplares de cada cuadernillo de prueba y cada uno de éstos es corregido en forma independiente por cuatro correctores, sin que ninguno de ellos conozca los códigos asignados por los demás. Luego se comparan los códigos y se establece un índice de confiabilidad que mide el grado de consistencia de las correcciones. Si la consistencia es baja, ello puede dar lugar a la invalidación del proceso de corrección.

Como procedimiento previo a este tipo de controles de confiabilidad, es imprescindible establecer pautas y criterios detallados y precisos para la corrección, así como un entrenamiento y supervisión de los correctores. Esto no siempre ocurre en los casos de pruebas de concurso.

4.2. Los resultados de una prueba son poco precisos en el ordenamiento de los sujetos o entidades evaluados

Toda calificación numérica de los conocimientos y capacidades de un individuo, así como de la "calidad" de la educación de un centro educativo o de un país, está sujeta a error de

medición. Ninguna medida es absolutamente precisa. Esto implica que todo ordenamiento de individuos, instituciones o países, en base a una calificación numérica, debe ser realizada y analizada con sumo cuidado.

4.2.a. En la mayoría de los concursos de selección –como el ejemplo de los directores anteriormente mencionado–, existen niveles de error importantes que no están controlados. Esto significa que Ana obtuvo 85 puntos, pero bien podría haber obtenido 80, así como también 90, dependiendo de diversos imponderables. Y Lucía obtuvo 80 puntos, pero si le hubiese tocado en suerte otro tribunal en la prueba teórica podría haber alcanzado los 95 puntos. En este caso, el componente de azar en la puntuación final tiene una consecuencia muy importante, porque determina quién va a ocupar un cargo de director y quién no (Ana tendrá prioridad sobre Lucía, pero podría haber sido al revés).

El ordenamiento final de los candidatos en general no refleja con precisión un ordenamiento en cuanto a sus capacidades para el cargo. Si no hay procedimientos de control de la calidad del proceso de determinación de puntajes ni estimación de la magnitud del error posible en los mismos, es imposible saber con propiedad qué tan grave es el problema.

Dada la trascendencia que los procesos de selección de mandos medios tienen para la calidad del sistema educativo, este problema debería ser encarado de algún modo como, por ejemplo, a través de mecanismos de control de la comparabilidad de los puntajes otorgados por distintos tribunales o mediante márgenes de error que permitan establecer cuántos puntos hacen que una diferencia de puntajes entre dos candidatos sea significativa.

4.2.b. La mayor parte de los rankings de escuelas o de países que suelen tener amplia difusión en la prensa se basan en el ordenamiento de dichas entidades en función de una cifra, el promedio de los puntajes alcanzados por sus estudiantes. Sin embargo, no todas las diferencias de puntajes tienen un significado relevante.

En este tipo de evaluaciones el error posible de la medición puede ser calculado mediante procesos estadísticos. El error posible suele representarse gráficamente utilizando una barra para indicar el puntaje promedio, y una "caja" (técnicamente denominada "intervalo de confianza") que marca los límites de precisión de dicho promedio (véase la Figura 2).

El significado de esta "caja" es el siguiente: el valor real del promedio de cada país se ubica, con un 95% de confianza, en algún lugar dentro de la caja, no estrictamente en la línea que indica la media.

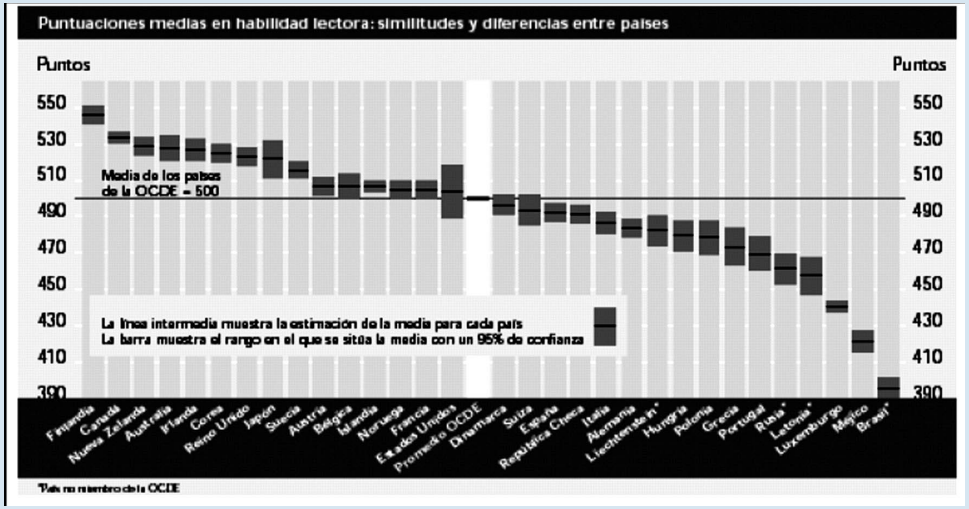
En otras palabras, cada país obtuvo una media sujeta a error y el valor de esa media puede variar dentro de los límites del intervalo de confianza, es decir, el valor correspondiente a un país puede ser algo mayor o algo inferior al que indica la media.

Este hecho tiene consecuencias muy importantes.

Si las "cajas" correspondientes a dos países diferentes se superponen, es decir, si tienen en común una parte de la escala de puntajes, esto significa que no puede afirmarse con propiedad que el resultado de uno sea mejor o peor que el resultado del otro.

En términos técnicos se dice en estos casos que la diferencia de puntajes no es "estadísticamente significativa", lo cual implica que, debido al error de medición, no podemos saber si el país que aparece con un puntaje más bajo en realidad no es mejor que el otro.

Resultados de PISA 2000 en Lectura ***Figura 2***



Fuente: Ministerio de Educación, Cultura y Deporte de España-Instituto Nacional de Calidad y Evaluación (INCE)/ OCDE, 2001; Conocimientos y destrezas para la vida: Primeros Resultados del Proyecto PISA 2000. Resumen de Resultados. INCE, Madrid.

Para que la diferencia sea "estadísticamente significativa", las "cajas" no deben superponerse. Eso garantiza que, cualquiera sea el puntaje "verdadero" de los dos países, está garantizado que uno de ellos está por encima del otro.

Como ejemplo de lo anterior, en la Figura 2, correspondiente a la evaluación PISA 2000, Finlandia tiene un resultado superior a Canadá, pero no hay diferencias significativas entre este último país y Nueva Zelanda, Australia e Irlanda.

En el otro extremo del gráfico, el resultado de Brasil es inferior al de México y el de México al de Luxemburgo, pero las diferencias no son significativas entre Letonia, Rusia, Portugal y Grecia.

De todos modos debe subrayarse que "estadísticamente significativo" no significa que las diferencias entre dos países sean de gran magnitud. En realidad normalmente las diferencias entre países "adyacentes" en un *ranking* son pequeñas, aún cuando sean "estadísticamente significativas". Esto último solo quiere decir que existe una diferencia real, pero la misma puede ser grande, mediana o pequeña.

Este tema es abordado con más detalle en las Fichas 8 y 10.

Síntesis final

La confiabilidad y la validez son conceptos relacionados pero diferentes, ambos estrechamente relacionados con el uso apropiado de los resultados de las evaluaciones.

La validez está referida al grado en que una evaluación realmente evalúa aquello que se supone evalúa –el aspecto sustantivo–. Es el concepto principal para analizar una evaluación, porque está relacionado con el significado de qué es lo que está siendo evaluado. La confiabilidad, en cambio, puede ser definida como la consistencia y precisión de los resultados de una prueba o de una evaluación.

La confiabilidad es condición necesaria pero no suficiente para la validez. Una prueba puede ser confiable pero no válida, es decir, se puede lograr una medida muy precisa, pero de algo que no es lo que en realidad interesaba evaluar.

Por lo tanto, lo primero que debe ser garantizado en cualquier evaluación es la validez. Lo primero que el usuario de las evaluaciones estandarizadas debe preguntarse es en qué medida aquello que se evalúa en las pruebas es relevante y deseable como logro educativo. Lo mismo se aplica a las evaluaciones que realizan los docentes y los centros educativos.

La confiabilidad es una cuestión de grado: los resultados nunca son perfectamente precisos, siempre están sujetos a error. Este error puede ser grande o chico y lo importante es poder estimarlo. Lo grave en una evaluación es que no exista ninguna estimación de error, porque entonces no hay forma de interpretar adecuadamente los resultados.

Es por esta razón que puede considerarse que la confiabilidad forma parte de la validez. Si una prueba es muy apropiada desde el punto de vista de sus contenidos –validez de constructo y de contenidos– pero arroja resultados muy imprecisos- –baja confiabilidad–, estará seriamente afectada la validez de uso de esos resultados. Es el caso del concurso de selección de directores anteriormente empleado como ejemplo.

El problema de la confiabilidad es una cuestión de grados en función de los propósitos de la evaluación[2]. Es decir, según cuál sea el propósito y consecuencias de una evaluación, se podrá tolerar un mayor o menor nivel de error en la precisión. En una prueba cuyo fin es realizar un diagnóstico de los aprendizajes en el país, el nivel de error en la estimación de los puntajes de cada individuo no es cuestión de vida o muerte. Se puede aceptar un monto de error mayor que en el caso de una evaluación de cuyo resultado dependa la posibilidad del estudiante de ingresar a una universidad o la posibilidad de un docente de acceder a un cargo de director. En estos casos, en que las pruebas tienen consecuencias "fuertes" para las personas, la precisión de los puntajes debería ser mayor.

Los sistemas nacionales de evaluación en América Latina se encuentran realizando importantes esfuerzos por mejorar la calidad de sus pruebas y la pertinencia de los conocimientos y competencias que son objeto de evaluación. Al mismo tiempo, es necesario mejorar los modos de reportar los resultados, incorporando información sobre los márgenes de error de los mismos y sobre la significancia de las diferencias de resultados entre escuelas o jurisdicciones, a efectos de que la interpretación de dichas diferencias sea apropiada.

2) LINN, R. Et GRONLUND, N., 2000; Measurement and Assessment in Teaching (8° edición), pp. 131-133. Prentice Hall

¿DEBEMOS CREERLE A LAS EVALUACIONES ESTANDARIZADAS EXTERNAS O A LAS EVALUACIONES QUE REALIZA EL DOCENTE EN EL AULA?

Los debates ideológicos sobre las evaluaciones

El propósito de esta Ficha es explicitar y clarificar algunas de las principales contraposiciones que suelen plantearse en los debates sobre evaluación educativa: entre la evaluación estandarizada externa y la evaluación en el aula; entre evaluación de resultados y evaluación de procesos; entre evaluación cuantitativa y evaluación cualitativa.

El objetivo es mostrar que por detrás de estos debates se encuentran implícitas las principales tensiones presentes en todo proceso educativo: entre lo individual y lo colectivo; entre lo local y lo general; entre el respeto a la diversidad, por un lado, y la integración social, por otro.

La Ficha se propone discutir y aclarar algunos de los prejuicios más comunes acerca de las evaluaciones estandarizadas, así como explicar la parte de verdad que contienen muchas críticas a dichas evaluaciones. Explícitamente, la Ficha se propone argumentar contra el mito, sostenido por muchos educadores, que sostiene que solo las evaluaciones realizadas por ellos mismos en las aulas tienen validez y legitimidad.

En tal sentido, la Ficha se propone mostrar que las evaluaciones estandarizadas y las evaluaciones en el aula son complementarias y no antagónicas. Cada una permite "ver" o "hacer" algunas cosas, pero no otras. La evaluación externa sirve para poner el foco de atención en aquello que todos los alumnos deberían aprender pero, por supuesto, no puede ni pretende dar cuenta de todos los aprendizajes. La evaluación en el aula, cuando se hace bien, puede ser mucho más rica en su apreciación de los procesos de aprendizaje de alumnos específicos, pero no puede nunca ofrecer un panorama de lo que ocurre a nivel del conjunto del sistema educativo.

Por tanto, se trata de una falsa oposición que el título de la Ficha refleja deliberadamente. No se trata de creerle a unas ó a otras, sino de usar ambas de acuerdo a diferentes propósitos y a lo que cada una de ellas puede dar.

1. Tres falsas oposiciones

Las tres principales falsas oposiciones que normalmente forman parte de los discursos contrarios a las evaluaciones estandarizadas son las siguientes:

1.1. Enfoque cuantitativo vs. cualitativo

Según esta postura, lo que invalida a las evaluaciones estandarizadas es su pretensión de "cuantificar" el aprendizaje, que es un fenómeno esencialmente cualitativo. Esta crítica suele ir acompañada por la afirmación de que las evaluaciones estandarizadas solo se preocupan por los "resultados", pero no por los "procesos". Generalmente incluye también una visión que considera despectivamente como "positivista" cualquier intento de analizar con herramientas estadísticas la realidad social.

1.2. Evaluación interna vs. externa

El segundo gran defecto de las evaluaciones estandarizadas sería su carácter externo al aula y a los procesos que ocurren dentro de ella. Según esta postura, sólo el docente que está cotidianamente dentro del aula y en contacto con los alumnos puede "conocer", y por tanto evaluar, los aprendizajes.

Las evaluaciones estandarizadas, por su carácter externo y "alejado" del aula, no tendrían la capacidad de captar lo que los alumnos aprenden y, por lo que tampoco tendrían nada que aportar a los docentes.

1.3. Diversidad y contextualización vs. estándares y control central

El tercer argumento contrario a las evaluaciones estandarizadas apunta justamente a su carácter estandarizado: se evalúa a la diversidad de alumnos de un país o región con un único instrumento y, por tanto, con una definición común a todos, acerca de lo que deberían aprender. Un intento de esta naturaleza, afirman los críticos, desconoce la heterogeneidad cultural, la diversidad de aprendizajes que se verifican en los múltiples contextos sociales y escolares, así como la diversidad de la enseñanza que brindan los profesores. La pretensión de evaluar con instrumentos estandarizados obedecería, en definitiva, a una pretensión de control estatal sobre la enseñanza y el trabajo docente, cuyos efectos serían negativos.

En las páginas que siguen ilustraremos y discutiremos cada una de estas tres líneas argumentales opuestas a la evaluación estandarizada.

2. Cuantitativo vs. cualitativo

El párrafo incluido en el Recuadro 1, tomado de un artículo periodístico y firmado por una profesora, es ilustrativo de la primera de las falsas oposiciones que nos interesa analizar.

Antes que nada, cabe señalar que el tono peyorativo del artículo, denotado por el uso del término "ordalía", así como la referencia a la intención "del poder" de deslumbrarse con algo "mega" (aparentemente la autora del artículo desconoce que un censo de 40.000 alumnos no tiene nada de "mega" en comparación con las evaluaciones de varios millones de alumnos que se realizan anualmente en países como Brasil o México), implica un abordaje del tema poco propicio para el debate en el plano de las ideas. Más bien se propone recurrir a la descalificación.

Desde el punto de vista conceptual, es equivocada la aseveración de que a este tipo de evaluaciones solo le preocupa *cuánto* saben los estudiantes y no *qué* saben.

Recuadro 1

"Ordalías evaluativas, como el 'censo de aprendizajes' aplicado a más de 40.000 estudiantes de tercer año liceal en 1999 son antes que nada el gesto del poder que puede deslumbrarse a sí mismo con algo verdaderamente 'mega'. Sin embargo, difícilmente pueden ser consideradas acciones educativas en la medida en que (siempre fieles a la pasión cuantitativa) saber cuánto saben los estudiantes en general no tiene demasiado sentido para alumnos o para profesores en particular. Tal vez si hubieran querido saber qué saben, la cosa hubiera sido distinta, pero en la concepción de base de estos procedimientos llamados evaluatorios, todos los estudiantes saben lo mismo, solo que algunos más que otros, y eso es lo único que interesa.

Ana Zavala, Semanario Brecha, Montevideo, 19 de mayo de 2000.

En realidad, toda evaluación estandarizada comienza justamente por definir qué deberían saber los estudiantes. Lo que luego por lo general se cuantifica, no es cuánto saben los alumnos, sino cuántos alumnos saben lo que deberían saber, que no es lo mismo (la cuestión de "*lo que deberían saber*" remite a otro aspecto que trataremos un poco más adelante). Cuando se trabaja con niveles de desempeño que describen lo que son capaces de hacer los alumnos con lo que han aprendido (algo intrínsecamente cualitativo), se cuantifica la cantidad o proporción de alumnos en cada nivel.

FICHA 5

¿Debemos creerle a las evaluaciones estandarizadas externas o a las evaluaciones que realiza el docente en el aula?

Sí es cierto que los resultados de muchas evaluaciones estandarizadas, que reportan en términos de porcentajes de respuestas correctas, suelen ser erróneamente interpretados como si mostrasen qué porcentaje "saben" los alumnos de lo que deberían saber.

Pero la oposición cuantitativo-cualitativa es una falsa oposición. Se trata de enfoques diferentes para el abordaje de la realidad, cada uno de los cuales da cuenta de un aspecto de la misma. Ninguno da cuenta por completo de ella, más bien son enfoques que se complementan y que, normalmente, se utilizan en forma conjunta.

Toda cuantificación parte siempre de un análisis cualitativo, es decir, de una descripción y categorización de situaciones de la realidad. Un indicador de pobreza, por ejemplo, es un dato cuantitativo que parte de una definición cualitativa de cuáles son las situaciones que deben considerarse como "pobreza".

Una prueba, tanto la que diseña el profesor en el aula como la que se diseña a nivel central en una unidad especializada en evaluación, es siempre un instrumento diseñado para captar un fenómeno cualitativo: lo que el alumno conoce y lo que puede hacer con el conocimiento.

Obviamente, puede haber pruebas buenas, regulares y malas, tanto entre las estandarizadas como entre las elaboradas por los profesores, pero ese es un problema distinto (véase la Ficha 6).

El hecho de que los resultados de una prueba cualquiera se cuantifiquen o no, depende de los propósitos de la evaluación. Para un profesor en el aula, seguramente no es necesario calcular qué porcentaje de sus alumnos obtuvieron un resultado satisfactorio en su curso. No siempre necesita "cuantificar" los resultados, aunque muchos probablemente lo hagan.

Un centro educativo puede tener interés en "cuantificar" los resultados de sus estudiantes en las pruebas o exámenes diseñados y aplicados dentro de la institución.

Cuando uno se coloca en la perspectiva de comprender la realidad de un sistema educativo que involucra a decenas o cientos de miles de alumnos, con toda su complejidad y magnitud, no tiene otro camino que cuantificar distintos aspectos de la realidad, uno de los cuales son los alumnos en función de sus desempeños. Pero estos siguen siendo "cualitativos".

Lo importante en este punto es captar la diferencia entre el carácter del instrumento –que siempre intenta captar una realidad "cualitativa"– con el hecho posterior de cuantificar o no los resultados de la aplicación del instrumento.

Como siempre, lo que hace la diferencia es el propósito y contexto de la evaluación. Una evaluación cuyo cometido es dar cuenta de la situación de un sistema educativo necesariamente debe cuantificar, salvo que alguien crea que se puede describir, analizar y comprender un sistema educativo exclusivamente sobre la base de estudios de casos.

Note el lector que aun en el caso de que sólo se acepten como válidas las pruebas y las evaluaciones elaboradas y aplicadas por los profesores, a la hora de dar cuenta de la realidad del sistema educativo uno no tendría más remedio que cuantificar esas evaluaciones. Eso es lo que se hace, por ejemplo, cuando se utiliza como indicador una tasa de aprobación de exámenes diseñados y aplicados localmente en cada centro educativo o una tasa de repetición.

Finalmente, llevada al extremo, la postura "anticuantitativa" implicaría no usar información en la conducción del sistema educativo y tomar decisiones exclusivamente sobre la base de opiniones de los actores, sean éstos políticos, sindicalistas, profesores, académicos, etc.

¿Cuál es la parte de verdad en la crítica "anticuantitativa"?

En primer término, es cierto que muchas veces se "mistifican" los números y se pierde de vista que estos son solamente una aproximación a ciertos aspectos de una realidad compleja e inabarcable. Ello ocurre cuando se reduce la calidad del trabajo educativo de un profesor o de un centro de enseñanza al resultado promedio de sus alumnos en una prueba o cuando se reduce la calidad de la educación en un país a la posición en un *ranking*.

Otra actitud que denota una mistificación de los números es la de trabajar, analizar, discutir y hasta decidir políticas a partir de ellos, sin saber qué significan. Es muy común, por ejemplo, que se formulen juicios de valor sobre la calidad de un sistema educativo a partir de resultados de pruebas que en realidad son muy malas y no evalúan lo realmente importante. Muchos analistas, decisores, economistas e investigadores emplean los resultados sin tener una idea más o menos acabada de qué fue lo evaluado y en qué medida es relevante para los análisis que se proponen realizar.

En segundo término, es cierto que la complejidad de los fenómenos educativos no se

capta exclusivamente a partir de miradas "macro", sino que es importante recurrir también a estudios en profundidad, a la observación de instituciones y aulas, a estudios de casos, a entrevistas a los actores, que aportan otras perspectivas y comprensiones de la realidad.

En tercer término, es cierto que uno de los problemas de las evaluaciones en gran escala es que muchas veces obligan a emplear únicamente instrumentos que puedan ser procesados mediante lectura óptica, lo cual puede empobrecer la calidad de las pruebas porque no logran evaluar las competencias más complejas y relevantes.

3. Evaluación interna vs. evaluación externa

El argumento contra las evaluaciones estandarizadas basado en su carácter externo al aula plantea también una falsa oposición (véase el Recuadro 2).

En primer lugar, debe decirse que no es cierto que estar dentro del aula y en relación continua con los alumnos garantice conocer lo que aprenden. A nadie escapa que muchos profesores, a pesar de la convivencia cotidiana con los alumnos, no tienen una idea cabal de los procesos de aprendizajes de estos.

Esto ocurre, a veces, porque la cantidad de alumnos que un profesor debe atender hace imposible tal seguimiento personal. Pero debe decirse también, aunque resulte duro, que muchos profesores simplemente no poseen las categorías conceptuales necesarias para comprender los procesos cognitivos de sus alumnos.

Recuadro 2

"Si se comparan una evaluación dentro del proceso de aprendizaje y una externa, 'hace una diferencia como del día y la noche', insiste Zavala. El gran problema de las evaluaciones externas es que 'se pierden de la misma más de la mitad'... '**La única posibilidad** de saber lo que el otro sabe es una continua relación pedagógica, ya que según la confiabilidad de las pruebas el desempeño puede ser igual, inferior o superior en el alumno'. Por ello los cambios en el rendimiento son imperceptibles en las evaluaciones externas. 'Solamente cuando estás en el juego real de la [relación] interpersonal pedagógica te das cuenta cuándo

> a Fulanito le tiene que haber pasado algo, o que Menganito debe haber copiado porque nunca hizo algo así y además su trabajo se parece sugestivamente al de Zutanito que se sienta al lado. Esto no puede detectarse con lector óptico'...
>
> *Diego Sempol, Semanario BRECHA, Montevideo, 8 de agosto de 2003*

Debe decirse también que a muchos profesores simplemente no les interesa captar lo que pasa con sus alumnos: simplemente les preocupa "dar" sus clases para los alumnos que puedan seguirlos.

Con esto no se pretende descalificar genéricamente a los profesores, sino simplemente señalar que, como en todas las profesiones, hay buenos y malos profesionales, y que la "cercanía" con los estudiantes no garantiza en absoluto evaluaciones ecuánimes y apropiadas.

En segundo lugar, hay una confusión respecto a los propósitos y tipos de evaluaciones. La mayoría de las evaluaciones estandarizadas no pretende evaluar a los alumnos individualmente, sino obtener un panorama general del sistema educativo. En este contexto, las situaciones particulares tienden a compensarse. Esta es una ley estadística. ¿Qué significa? Que si bien es cierto que Fulanito el día de la prueba "andaba mal" y no hizo lo mejor que podía, al mismo tiempo a tal otro alumno le fue mejor de lo normal. En el conjunto nacional, estas excepciones individuales tienden a compensarse unas con otras. De modo que la fotografía que se obtiene es razonablemente cercana a la realidad.

Lo importante es captar que en una evaluación estandarizada no importan las situaciones individuales y excepcionales, en la medida en que el propósito no es evaluar a "Fulanito", sino tener una visión global del sistema.

Este tipo de dificultades sí se plantea en los casos en que las evaluaciones tienen consecuencias para los alumnos (exámenes de certificación o pruebas de selección), pero el tipo de problemas que se plantean no es diferente de los que se exponen ante una situación de examen tradicional.

Por otra parte, las evaluaciones internas están sujetas a fenómenos de subjetividad. Muchas veces los profesores son condescendientes con sus alumnos, a veces justamente, porque conocen sus problemas y otras veces porque son poco exigentes, todo lo cual da lugar a

problemas conocidos: muchos estudiantes llegan a la universidad o al mundo del trabajo sin los conocimientos y competencias imprescindibles.

La responsabilidad por este problema no puede ser atribuido a las evaluaciones estandarizadas, sino, más bien, a las evaluaciones internas que diseñan e implementan los profesores.

Las evaluaciones externas justamente pretenden ofrecer al profesor la posibilidad de ubicar el desempeño de sus alumnos y sus propias evaluaciones en el marco de referencia más amplio de los aprendizajes en el conjunto del sistema educativo.

En este sentido, evaluaciones internas y externas pueden y deben complementarse, en la medida en que ambas tienen limitaciones y potenciales.

4. Diversidad y contextualización vs. estándares y control central

Los argumentos incluidos en los Recuadros 3 y 4 ponen de manifiesto una tensión ineludible y siempre presente en la acción educativa.

La educación tiene siempre un doble propósito: potenciar el desarrollo único y original de los individuos y, al mismo tiempo, incorporarlos a la sociedad de la que forman parte y permitirles manejar los códigos de ésta; fomentar el desarrollo de las culturas locales pero, al mismo tiempo, garantizar la integración nacional a través de un conjunto de valores y conocimientos compartidos. Esta tensión inevitablemente aparece siempre en las decisiones relativas al currículo y a la evaluación. Si se sigue la línea argumental planteada en los Recuadros 3 y 4, no solo no debería haber evaluación estandarizada, sino que tampoco debería existir currículo, ya que también es una imposición social y política desde el Estado. Llevando el argumento al extremo, no habría educación, pues nadie tendría derecho a decidir qué deben aprender o cómo deben formarse otros, ni siquiera los propios padres.

Recuadro 3

"Ambas pruebas tienen así un punto de partida extremadamente polémico: la aceptación de criterios –a nivel internacional en un caso y local en el otro– de lo que se supone deberían lograr hacer los alumnos para afrontar los "desafíos del mundo actual". ¿Es posible unificar tal diversidad de desafíos y necesidades en una sociedad fuertemente estratificada como la uruguaya, con diferencias locales y una realidad nacional muy distante de la que poseen la mayoría de los países del Primer Mundo?...

"...El estudio explicitó así la dificultad de aplicar homogéneamente estándares de evaluación que se elaboran preferentemente sobre la base de las expectativas de los grupos de clase media a todos los estratos sociales".

"...Por ello las pruebas estandarizadas no aportarían significativamente nada a un proceso de enseñanza aprendizaje, ya que además partirían de un imposible: 'Que todos los profesores enseñamos lo mismo y de la misma manera, en lo informativo, en lo interpretativo y en los énfasis que ponemos en estas cosas. El famoso estándar es más o menos como el mínimo común múltiplo de todos los saberes, enfoques, énfasis de un tema o de una asignatura, lo que es imposible', comentó"

Diego Sempol, Semanario BRECHA, Montevideo, 8 de agosto de 2003.

Recuadro 4

"El objetivo de este instrumento 'desarrollo de competencias' es un intento de resocialización, normalización y control social, por parte de los Estados y aparatos institucionales similares, una suerte de programación encubierta, basada en las nociones del déficit, el *handicap* y el riesgo, orientada al control ideológico y a la producción de la fuerza laboral.

Implica un juicio de valor, respecto de una cierta forma de comportamiento social, considerado como normal, esperable, deseable e incluso definidor del éxito social de un sujeto dado... El estudiante en quien busca desarrollarse estas habilidades y destrezas, se concibe como un sujeto vacío, como la caja negra en que se prefiguró su mente. Un ente sin rostro ni condiciones distintas, en una operación de claro corte político-social... Poco importa la historia individual de cada joven, su creatividad, su potencialidad personal, esas desviaciones a la norma que frecuentemente hacen a los genios. Debe desarrollar unas –predeterminadas- competencias, debe ser funcional al lugar que se le tiene destinado en la sociedad. Debe responder a los requerimientos que se preestablecieron para él. En definitiva, lo que debió ser un umbral, termina convirtiéndose en un techo...

FICHA 5

¿Debemos creerle a las evaluaciones estandarizadas externas o a las evaluaciones que realiza el docente en el aula?

> Javier –podría tener mil nombres– es un estudiante secundario convencido que solamente estudiando podrá 'ser alguien'. Javier estudio bachillerato en un liceo de la Transformación. Para Javier, el 'éxito´ en sus estudio es 'desarrollar competencias'. Las que otros decidieron que desarrolle. Esas, sin las cuales 'no es'...
>
> Javier es alguien, no es un artefacto del cual desencadenar mecanismos. Javier es un sujeto complejo, ante todo un sujeto que debe ser respetado y salvaguardado en su originalidad. No puede ser fragmentado en competencias, disociado asertivamente. La educación no puede ser otra cosa que una plataforma desde la cual desarrolle su complejidad como ser único, capaz no solamente de insertarse en un mercado laboral, sino de salirse de ese mercantilismo que lo cosifica, de pensar en otras formas de relaciones sociales, en otras formas de funcionamiento económico..."
>
> *Margarita Ferro; Competencias: Una intervención ideológica de control social; Boletín de la Asociación de Docentes de Enseñanza Secundaria, Montevideo, marzo de 2004.*

En la medida en que existe educación, y que esta no ocurre en un vacío social y político, siempre va a existir un recorte o priorización arbitraria de saberes, capacidades, habilidades, competencias o como se les quiera llamar. Eso y no otra cosa es el currículo.

En el fondo, si nadie puede decir algo sobre qué deben aprender los alumnos, nadie puede tampoco controlar ni exigir nada a los docentes. Cada docente enseña lo que le parece más apropiado y cada alumno aprende lo que puede de acuerdo a su singularidad, creatividad y potencialidades individuales.
Es interesante mostrar que esta suerte de individualismo o localismo exacerbado se parece mucho a las visiones del mundo más liberales.

Las propuestas del tipo de las incluidas en los Recuadros 3 y 4 terminan deslegitimando el rol del Estado en la formación de los ciudadanos y coincidiendo con posturas propias del otro extremo del espectro ideológico, como la incluida en el Recuadro 5, que justamente busca minimizar el rol de la educación pública y el papel del Estado en la educación.

Recuadro 5

"Ello nos conduciría a una situación en la cual el Estado estaría descalificado como maestro, pero –al mismo tiempo- tendría que ocuparse de la cuestión educacional. ¿De qué manera? No enseñando, por descontado, pero sí asegurándose de que alguien debidamente calificado enseñaría a aquéllos que se hallen en la edad de aprender.

Y esa seguridad la procuraría el Estado valiéndose de su doble condición de generador válido de normas coactivas –capaz, por tanto, de obligar a los hogares a observar el comportamiento deseado– a la vez que de titular de la soberanía fiscal, con la consiguiente facultad de extraer recursos pecuniarios de la comunidad, vía impuestos, y utilizarlos para allegárselos a las familias que no dispusiesen de ellos, vía subsidios.

Quedaría aún por determinar quiénes serían los educadores habilitados, y cómo se asignarían a cada uno de ellos los potenciales educandos. Para resolver una y otra cosa la comunidad se valdría de la institución a que cotidianamente recurre para satisfacer las necesidades de sus individuos o familias mediante la aplicación de recursos escasos, vale decir, como es obvio, al mercado. Al mercado de educación, dicho más precisamente, compuesto de una oferta de servicios docentes, integrada por todos aquellos que, en ejercicio de la libertad de trabajo, se sintiesen con la vocación de prestarlos, y una demanda, compuesta por los hogares, en ejercicio de la libertad de educación, munidos, a efectos de ejercitarla, de recursos propios o allegados por la autoridad pública....

"Volviendo al tema de la enseñanza, me parece oportuno cerrar este artículo con una transcripción de John Stuart Mill, tomada de su famoso opúsculo 'On Liberty'. Mill ha estado ponderando las virtudes de la individualidad de los caracteres y los patrones de conducta para el bien social; después de lo cual, agrega que ello 'implica, por iguales razones, la suprema importancia de la diversidad en la educación. Una educación general a cargo del Estado no es más que un artificio para hacer que cada individuo sea una réplica perfecta de los demás; y, puesto que el molde en el cual va a ser fundido será el que le plazca a quien tenga el poder,... ello impone un despotismo sobre las mentes, que por tendencia natural se extenderá a un despotismo sobre los cuerpos".

Ramón Díaz, EL OBSERVADOR, Montevideo, 12 de enero de 2002.

El punto, en todo caso, es quién decide qué es lo que debe enseñarse: el Estado a través de las autoridades legítimas; los profesores individualmente o como corporación; las familias a través de la libre elección de escuelas en el marco de un "mercado educativo"... ¿Quién puede legítimamente definir qué deben aprender niños y jóvenes?

FICHA 5

¿Debemos creerle a las evaluaciones estandarizadas externas o a las evaluaciones que realiza el docente en el aula?

Probablemente no haya una respuesta única a esta pregunta.

Las evaluaciones estandarizadas implican que el Estado tiene un papel central en la definición de la respuesta a esa pregunta, a través de la producción de información sobre el grado en que los conocimientos y competencias fundamentales están siendo logradas en el sistema educativo.

Ello implica que tal definición debe existir, sea bajo la forma de currículo nacional o bajo otras formas.

Las posturas que niegan este papel al Estado en el fondo parten de un supuesto romántico-rousseauniano: el individuo es un ser sin condicionamientos sociales, libre y creativo. Se desconoce que sin un "umbral" o piso de competencias, sin una acumulación cultural básica, nadie puede construirse como ser libre[1]. Una minoría de la sociedad recibe de su propia familia este "umbral" de competencias, pero la gran mayoría no. Por tanto, dejar librado a la originalidad individual, a la diversidad de contextos locales o a los mecanismos de mercado el logro de las mismas, es condenar a las grandes mayorías a una posición de subordinación social.

Lo anterior no significa que el Estado deba decidirlo todo. Por supuesto, debe existir espacio para la autonomía, para la diversidad y para aprendizajes localmente decididos.

Las evaluaciones estandarizadas no pretenden entrar en el terreno de la diversidad y no se proponen evaluar todo lo que los alumnos aprenden, sino simplemente aquello que haya sido definido como base de formación común al conjunto de los ciudadanos (lo cual sin duda remite al problema de quién y cómo elabora el currículo, de qué proporción de este debe ser común a todos y qué espacio queda para las decisiones locales).

El problema central de las argumentaciones contra las evaluaciones estandarizadas es la tesis de que no puede haber un referente común que defina los aprendizajes fundamentales para todos los alumnos de un país.

Al mismo tiempo, la postura ultra liberal expuesta en el Recuadro 5 refleja también una contradicción: si el Estado solo debe exigir lo básico, y cuanto menos mejor, entonces las evaluaciones estandarizadas de la calidad de los centros y la publicación de *rankings* de escuelas como mecanismo para generar un mercado competitivo carecen de todo sentido.

1) Como mínimo es necesario que todos dominen un conjunto de competencias para poder vivir dentro –o voluntariamente fuera– del ámbito cultural dominante o más extendido. Es necesario poder "manejarse" con la cultura dominante, aun cuando no sea la propiamente abrazada.

Si cada centro es totalmente autónomo, entonces no es posible establecer un sistema central de control de calidad que oriente a los padres en la elección de escuela, dado que cada escuela enseñaría cosas distintas. La elección de los padres debería basarse en la propaganda de cada escuela sobre qué es lo que ofrece como distintivo, pero no habría un marco de conocimientos universalmente aceptado que pudiese ser objeto de evaluación.

Finalmente, merece también un espacio de reflexión la "satanización" del término "competencias". Satanizar ciertos términos es un recurso bastante común en los debates educativos, generalmente construidos a través de falsas dicotomías (educación nueva *vs.* tradicional o crítica *vs.* tecnicista; evaluar procesos vs. evaluar resultados; cualitativo *vs.* cuantitativo; etc.). Lo que hace el artículo citado en el Recuadro 4 es discutir en abstracto el término competencia y atribuirle una serie de propiedades "malignas".

En lugar de ello, habría que analizar qué competencias fueron definidas para una determinada evaluación y discutir si son relevantes o no, si ayudan al desarrollo de las potencialidades individuales o no, si abren puertas y posibilidades al individuo o no. A la inversa de lo que plantea la autora, el enfoque de competencias justamente busca enfatizar la capacidad del individuo de seguir aprendiendo, de no verse predestinado a un único tipo de trabajo o carrera para toda su vida. Muchas de las competencias que propone la OCDE para el estudio internacional PISA tienen relación justamente con la capacidad de reflexión y análisis crítico de los jóvenes.

El tema de las competencias es analizado con más detalle en la Ficha 7.

Síntesis final

Los principales argumentos contrarios a las evaluaciones estandarizadas pueden ser resumidos en las siguientes aseveraciones:

- Las evaluaciones estandarizadas evalúan "resultados" pero no evalúan "procesos".
- Las evaluaciones estandarizadas evalúan memorización de contenidos y habilidades simples, pero no pueden evaluar habilidades complejas.
- Las evaluaciones estandarizadas son cuantitativas y responden a una visión positivista de la realidad que pretende medir todo, en tanto las evaluaciones que realiza el profesor en el aula son cualitativas

y son las que verdaderamente pueden captar el proceso de aprendizaje de un alumno.

- Las evaluaciones estandarizadas evalúan apenas unos conocimientos básicos de Lenguaje y Matemática pero dejan de lado el resto de los aprendizajes.
- Las evaluaciones estandarizadas dejan de lado las actitudes y los valores como objetivo fundamental de la labor educativa.
- Las evaluaciones estandarizadas pretenden estandarizar las mentes, promover el "pensamiento único" y quitar libertad de enseñanza al profesor.
- Las evaluaciones estandarizadas no respetan la diversidad de aprendizajes que se dan en diversos contextos ni la diversidad de los alumnos. Pretenden que todos aprendan lo mismo y, por tanto, que todos los profesores enseñen lo mismo.

La mayoría de las afirmaciones anteriores contienen una parte de verdad y una parte de falsedad. Varios aspectos ya han sido discutidos a lo largo de la Ficha.

La dicotomía procesos-resultados es falsa. Las evaluaciones estandarizadas pueden aportar mucho a la comprensión de los procesos de trabajo de los alumnos en algunas áreas, si bien es cierto que el estudio en profundidad de los procesos cognitivos de los alumnos requiere de abordajes cualitativos. Algo similar ocurre con la segunda de las afirmaciones señaladas. En la Ficha 6 se muestra que las pruebas estandarizadas pueden evaluar diversidad de grados de complejidad de las competencias de los alumnos, dependiendo del enfoque conceptual y del tipo de actividades que se proponga a los alumnos.

Las contraposiciones cuantitativo-cualitativa y externa-interna han sido discutidas suficientemente en esta Ficha.

Es cierto que las evaluaciones estandarizadas no pueden evaluar todo lo deseable como objetivo de la educación –ni pretenden hacerlo–. Son generalmente pruebas de lápiz y papel dirigidas a evaluar algunos de los aspectos centrales de la enseñanza, pero no cubren la totalidad de los objetivos y propósitos educativos.

Hay disciplinas importantes que no suelen ser objeto de evaluación. La educación se propone además formar a los niños y jóvenes en actitudes y valores, cuya evaluación es muy compleja –es discutible que sea posible y deseable hacerlo–. Hay aspectos de carácter local o contextual que son parte central de la enseñanza pero no pueden formar parte de una evaluación estandarizada a nivel nacional.

Por estas razones el término "calidad", que generalmente se adjudica a las evaluaciones estandarizadas, puede ser excesivo, en la medida en que hay mucho más en la calidad de la educación que no puede ser contemplado por las evaluaciones estandarizadas. Es importante decir explícitamente que lo que se evalúa es una parte de los objetivos educacionales, una parte importante, pero que no se está abarcando todo lo importante ni se pretende que los centros educativos se dediquen exclusivamente a trabajar en torno a aquello que es evaluado.

Esto lleva a que se deban extremar los cuidados en relación a muchos de los usos propuestos para los resultados de las evaluaciones estandarizadas, en particular en lo que hace a vincular consecuencias "fuertes" a las mismas. Es cierto que muchas veces el modo en que se difunden y utilizan los resultados induce a reducir la educación a aquello que es objeto de evaluación estandarizada, aspectos que son tratados con más detalle en la Ficha 11.

Parte del proceso de consolidación de una cultura de la evaluación es aprender a ver los resultados de las evaluaciones estandarizadas en sus justos términos, es decir, como información necesaria pero no suficiente, que permite describir la situación del sistema educativo y orientar el trabajo de los docentes, a partir de un recorte o parcela de conocimientos y competencias que han sido definidas como fundamentales para todos los estudiantes y que son necesarias para el desarrollo de sus capacidades en otras áreas formativas.

Por cierto, un supuesto "fuerte" en el enfoque de esta Ficha es que esto último, la definición de un recorte de conocimientos y competencias fundamentales exigibles a todos los estudiantes es posible, necesario e imprescindible. El punto fue abundantemente discutido a lo largo de la Ficha.

Para finalizar, partiendo del supuesto anterior, es preciso enfatizar que la "estandarización" en las evaluaciones a gran escala consiste sencillamente en garantizar que los resultados sean comparables, para lo cual es necesario que las pruebas sean comunes para todos, al igual que los procedimientos de aplicación y corrección.

¿QUÉ EVALÚA ESTA PRUEBA?

Distintos tipos de actividades en las evaluaciones estandarizadas

Las Fichas 6 y 7 pretenden mostrar con ejemplos cómo distintas pruebas para una misma disciplina pueden evaluar cosas muy distintas. La Ficha 6 lo hace a través de ejemplos del tipo de actividades que son propuestas a los alumnos en las pruebas, en tanto que la Ficha 7 lo hace en un nivel de mayor abstracción, mostrando los distintos enfoques conceptuales sobre los aprendizajes que subyacen a las actividades.

La intención de estas Fichas es alertar al lector para que, antes de fijar su atención en los "números" de las evaluaciones, se detenga a indagar qué fue evaluado, qué tipo de ítemes fueron empleados y qué grado de complejidad tuvieron las tareas que los alumnos debieron enfrentar.

Lamentablemente, muchas veces se utilizan los datos de las evaluaciones sin hacer antes estas preguntas, probablemente porque el lector da por supuesto que lo evaluado es lo realmente importante. Sin embargo, no debería darlo por supuesto. Muchas veces, como lo señalan con razón los críticos de las evaluaciones estandarizadas, estas evalúan competencias muy simples o nada más que la capacidad de memorizar y reproducir datos y procedimientos rutinarios.

Sin embargo, algunos sectores de la prensa suelen dejarse llevar casi exclusivamente por la fascinación por los números, con titulares del tipo *"Tabla de posiciones por comuna"* o *"Los Colegios Top de la región metropolitana"* y otros por el estilo, omitiendo muchas veces informar al público sobre qué son capaces de hacer los alumnos y qué es lo que no han aprendido.

Del mismo modo, muchos economistas realizan complejos análisis estadísticos, sin preguntarse de qué hablan los números de las evaluaciones. Dan por supuesto que hablan de aprendizajes relevantes, pero esto no siempre es así.

FICHA 6

Los "números" solo pueden tener algún significado si uno comprende qué tipo de aprendizajes fueron evaluados y desde qué concepción. En esta Ficha, este aspecto será abordado desde las actividades que las pruebas proponen a los alumnos, en Lenguaje y en Matemática, con el fin de mostrar la diversidad de niveles de complejidad. En la Ficha 7 se analizará las concepciones que subyacen a los distintos tipos de actividades.

1. Ejemplos de actividades de pruebas de Lenguaje

A continuación se incluye varios ejemplos de actividades de Lenguaje de pruebas estandarizadas. En cada caso se indica el origen de la actividad. La inclusión de actividades aisladas no implica ningún tipo de valoración positiva o negativa sobre el conjunto de la prueba o de la evaluación de la que fue tomada cada actividad.

Simplemente se pretende mostrar que el espectro de propuestas que se hace a los estudiantes es muy variado y que algunas actividades son débiles desde el punto de vista de lo que requieren de los alumnos, en tanto otras son más complejas y exigen más del alumno en términos cognitivos.

En general, las actividades de Lenguaje están referidas a uno o varios textos. Si bien resultaría interesante analizar los tipos de texto que se utilizan en las pruebas, los mismos no se incluyen por razones de espacio.

Lenguaje / 6° grado de Primaria — Ejemplo N° 1

Fuente: Esta es una actividad que iba a integrar una prueba censal en 1995 en Uruguay, que finalmente fue cancelada. Nunca fue aplicada. Publicada en: ANEP/Unidad de Medición de Resultados Educativos (2000); Evaluaciones Nacionales de Aprendizajes en Educación Primaria en el Uruguay. 1995-1999. Montevideo, Uruguay.

¿Qué voz del verbo «oír» falta en esta oración?
Alicia ..... que su madre la llamaba.

A) Huyó
B) Oró
C) Hoyo
D) Oyó

Lenguaje / 3° y 4° grado de Primaria — Ejemplo N° 2

Fuente: Esta es una actividad que integró la prueba aplicada por el Laboratorio Latinoamericano de Evaluación de la Calidad Educativa de la OREALC/UNESCO en 1997 (Forma A, ítem 4). La actividad pretende evaluar la "práctica metalingüística", específicamente, la capacidad para usar marcas de concordancia gramatical. El 74% de los niños la respondió correctamente. Es uno de los ítemes liberados, disponibles en Internet: http://www.simce.cl/doc/Preguntas_Primer_Estudio_LLECE_Lenguaje.pdf

El colibrí ..... hermosos colores.

A) tienen
B) tengo
C) tiene
D) tienes

Lenguaje / 3° y 4° grados de Primaria — Ejemplo N° 3

Fuente: Esta es una actividad que integró la prueba aplicada por el Laboratorio Latinoamericano de Evaluación de la Calidad Educativa de la OREALC/UNESCO en 1997 (Forma B, item 7). La actividad pretende evaluar la "comprensión lectora", específicamente, la capacidad para reconocer información precisa de un texto. El 69% de los niños la respondió correctamente. Es uno de los ítemes liberados, disponibles en Internet: http://www.simce.cl/doc/Preguntas_Primer_Estudio_LLECE_Lenguaje.pdf

La "caza fotográfica" consiste en .....

A) matar animales.
B) fotografiar animales.
C) encerrar animales.
D) alimentar animales.

FICHA 6

Ejemplo Nº 4 **Lenguaje / 3er. grado de Primaria**

En la pancarta azul se dice: "como las de Castilla en el mes de abril"; la palabra "las" se refiere a

A) Las épocas.

B) Las huertas.

C) Las poblaciones.

D) Las aguas.

Fuente: Secretaría de Educación de la Alcaldía Mayor de Bogotá D.C. (2001); Compendio Resultados. Evaluación de Competencias Básicas en Lenguaje, Matemática y Ciencias Naturales. Grados tercero, quinto, séptimo y noveno. Bogotá, Colombia

Ejemplo Nº 5 **Lenguaje / 3er. y 4º grado de Primaria**

Lee atentamente la página de este periódico y luego contesta las siguientes preguntas:

A) El diccionario define la palabra calzada de las siguientes maneras:
1. Camino pavimentado y ancho
2. Parte comprendida entre dos 'veredas'
3. En las carreteras, parte central por donde circulan los vehículos.
¿Cuál se adecua a este texto, la 1, la 2 o la 3?

B) ¿Cuál es la fecha de la publicación?

C) ¿En qué fecha se produjo el accidente?

D) ¿En qué momento del día se produjo el accidente?

E) ¿Cuál fue la causa del accidente?

Fuente: ANEP/Unidad de Medición de Resultados Educativos (2001); La prueba y otros aportes. Evaluación Autónoma de Aprendizajes. Enseñanza Primaria. 4º. año. Marzo 2001. Montevideo, Uruguay.

Lenguaje / 7° grado de Primaria **Ejemplo N° 6**

¿Cuál es una opinión y no un dato en el contenido de los avisos clasificados?

A) Arriendo para empresas.

B) Listo para la entrega.

C) Excelente casa.

D) Parqueadero y bodega.

¿Cuál es un dato y no una opinión en el contenido de los avisos clasificados?

A) Preciosa vista.

B) Excelente precio.

C) Sector La Carolina.

D) Acabados de primera.

Fuente: Estas dos actividades integraron la prueba aplicada por el Ministerio de Educación y Cultura de Ecuador en 1996 (APRENDO). Las actividades pretenden evaluar la capacidad de los alumnos para "diferenciar los hechos y las opiniones que contiene un texto". El 37% de los niños respondió correctamente a la primera pregunta y el 25% a la segunda. Publicados en: Análisis de las pruebas APRENDO 1996 y de sus Resultados. Lenguaje y Comunicación. Séptimo Año de Educación Básica. Ministerio de Educación y Cultura – EB/PRODEC (1998); Quito, Ecuador.

Lenguaje / 9° grado de Educación Basica **Ejemplo N° 7**

En el verso: *dar la vida y el alma a un desengaño,* ¿qué recurso se emplea?

A) Hipérbole

B) Símil

C) Epíteto

D) Humanización

Fuente: Esta actividad integró la prueba aplicada por el Ministerio de Educación de Venezuela en 1998 (SINEA). La actividad pretende evaluar "nociones lingüísticas", específicamente la capacidad de los alumnos para "identificar recursos literarios". El 33% de los alumnos respondió correctamente a la pregunta. Publicada en: Informe para el Docente. 9°. grado. Ministerio de Educación/SINEA (1999); Caracas, Venezuela.

FICHA 6

Ejemplo Nº 8 **Lectura / Jóvenes de 15 años escolarizados**

GRAFFITI

El estímulo de esta unidad –que no ha sido incluido por razones de espacio- consistía en dos cartas publicadas en Internet, una a favor y otra en contra de los grafiti. Han sido clasificadas como argumentativas en la medida en que exponían propuestas e intentaban persuadir al lector sobre un punto de vista.

Pregunta 1

El propósito de cada una de estas cartas es:

A) Explicar lo que son los *graffitis*.
B) Dar una opinión sobre los *graffitis*.
C) Demostrar la popularidad de los *graffitis*.
D) Decir a la gente cuánto cuesta borrar los *graffitis*.

Pregunta 2

¿Por qué Sofía hace referencia a la publicidad?

Esta es una pregunta "abierta", en la que el alumno debía construir su respuesta. Para contestar la pregunta correctamente, el estudiante debía reconocer que se había establecido una comparación entre la publicidad y el *graffitis*. La respuesta debía ser consistente con la idea de que la publicidad es una forma legal de *graffitis* o debía reconocer que la referencia a la publicidad era una estrategia para defender al *graffitis*.

Pregunta 3

¿Con cuál de las autoras de las cartas estarías de acuerdo? Explica tu respuesta con tus propias palabras para referirte a lo que se dice en una o en las dos cartas.

En esta pregunta, también de carácter "abierto", los estudiantes debían comparar lo que se afirma en los textos con sus propias opiniones sobre el tema. Para responderla era necesario que el estudiante comprendiera ampliamente al menos una de las cartas. Se consideraron correctas las respuestas que explicaban la opinión del estudiante y a la vez se referían al contenido de una o ambas cartas. Podrían referirse tanto a la

Fuente: Estas preguntas integraron la prueba aplicada por OCDE en más de 40 países en el año 2000/2001 (PISA). Publicado originalmente en inglés en Knowledge and Skills for Life. First Results from PISA 2000; OECD, 2001. Adaptado de la traducción incluida en el Informe Nacional de Perú, Una aproximación a la alfabetización lectora de los estudiantes peruanos de 15 años; UMC, 2003.

opinión general del autor como a detalles específicos de sus argumentos. Podrían parafrasear las cartas pero no se admitía como válido copiar partes íntegras de las cartas.

Pregunta 4

Se puede hablar sobre lo que dice una carta (su contenido).
Se puede hablar sobre la forma en que una carta está escrita (su estilo).
Sin tener en cuenta con qué carta estás de acuerdo, ¿cuál piensas tú que es la mejor carta? Explica tu respuesta refiriéndote a la forma en que una o las dos cartas están escritas.
Esta tarea evaluaba la reflexión y evaluación de la forma de un texto, tarea para la cual los lectores necesitaban hacer uso de su propio conocimiento de lo que constituye un buen escrito. Se esperaba que los estudiantes expliquen su opinión haciendo referencia al estilo, tonos, estructura y/o estrategias argumentativas de una o ambas cartas. Algunas respuestas típicas que obtuvieron crédito total fueron: "la carta de Olga fue efectiva por la manera directa como se dirigió a los artistas del *graffitis*", "En mi opinión, la segunda carta es mejor porque tiene preguntas que te involucran, haciéndote sentir que estás en una discusión más que en una conferencia".

La selección de actividades incluida en las páginas anteriores pretende ilustrar la diversidad de posibilidades existente para evaluar aprendizajes en Lenguaje.

Los tres primeros ejemplos corresponden a actividades que pueden ser respondidas usando el sentido común para descartar alternativas algo absurdas, antes que por la posesión de ciertos conocimientos o competencias. Son actividades que solo podrían tener sentido en una evaluación para individuos cuya lengua materna no fuera el castellano.

Un problema común a dichas actividades es que utilizan el formato de opinión múltiple para evaluar competencias que no se prestan para dicho formato, dado que no es posible construir alternativas de respuestas razonables y plausibles.

En el ejemplo 2 parece poco apropiado intentar evaluar la capacidad para usar marcas metalingüísticas mediante una pregunta de completamiento con elección múltiple entre

alternativas dadas. Esta capacidad debería ser evaluada en la producción escrita, o bien como pregunta de respuesta construida (es decir, dejando que el alumno complete la frase sin ofrecerle alternativas). Algo parecido ocurre en los ejemplos 1 y 3.

La actividad propuesta en el ejemplo 1 es particularmente absurda, sobre todo si se tiene en cuenta que estaba dirigida a alumnos de 6º año de escuela. Sin embargo, esta actividad cumplía con todos los requisitos estadísticos como para ser incluida en una prueba[1]. Con este ejemplo se pretende llamar la atención en cuanto a que no es suficiente con que un "ítem" cumpla con ciertos requisitos estadísticos para ser validado; debe además pasar por un análisis riguroso de sus propiedades "didácticas".

Los ejemplos 4 y 6, por el contrario, muestran actividades de elección múltiple que resultan relevantes, dado que las alternativas que se ofrecen al alumno forman parte del texto con el que está trabajando y son en mayor o menor medida plausibles. En el ejemplo 4 se pretende evaluar si el alumno reconoce en la práctica de la lectura el papel de un pronombre. En el ejemplo 6 se busca evaluar si el alumno es capaz de distinguir entre datos y opiniones, en este caso en un aviso.

El ejemplo 5 muestra una combinación de actividades de opción múltiple y de carácter "abierto" o "respuesta construida", dirigidas a alumnos de 3er y/o 4º grado de Primaria. Las actividades están dirigidas a evaluar la comprensión de una noticia periodística. En este caso la pregunta de opción múltiple corresponde a un contexto real, dado que las alternativas han sido tomadas de un diccionario. Las preguntas abiertas están dirigidas a evaluar la capacidad del alumno de identificar información en el texto, en algunos casos información más fácil de identificar y, en otros, información que requiere cierto grado de interpretación por parte del alumno.

La actividad incluida en el ejemplo 7 corresponde al 9º grado, es decir, a alumnos de 14-15 años de edad que están próximos a finalizar la escolaridad obligatoria. La pregunta que el lector debe hacerse en este caso es si el tipo de conocimiento evaluado por el ítem es tan relevante como para formar parte de una prueba nacional. Obviamente es materia discutible. El punto es qué se espera como base de conocimientos y competencias fundamentales para todos los alumnos al finalizar un determinado ciclo educativo.

En contraste con lo anterior, el ejemplo 8 muestra cómo se evalúa la competencia lectora de los jóvenes de 15 años en el Programa Internacional PISA de la OCDE. Como se

1) Normalmente los ítemes son probados en una prueba piloto, a partir de lo cual se obtiene para cada uno de ellos una serie de propiedades estadísticas relacionadas con su nivel de dificultad, su poder de diferenciación entre alumnos mejores y peores, etc.

puede apreciar, las preguntas están más orientadas hacia la capacidad de los alumnos de comprender y reflexionar sobre el lenguaje escrito que a conocimientos específicos como en el ejemplo 7.

Por otra parte, se puede observar que en PISA se trabaja con un conjunto de preguntas vinculadas a una situación disparadora (en este caso dos cartas sobre los *graffitis*), a partir de la cual se utilizan diversos formatos de pregunta, muchos de ellos de respuesta construida.

2. Ejemplos de actividades de Matemática

El presente apartado tiene la misma finalidad que el anterior, pero en este caso para el área de la Matemática.

La actividad propuesta en el ejemplo 9 ilustra un tipo de seudoproblema matemático, muy común en muchas pruebas, así como en muchas prácticas de enseñanza en las aulas. La situación es bastante absurda y difícilmente en alguna ocasión un niño deba realizar un cálculo como el que se le pide para saber cuántos caramelos le regalaron. Si lo que se pretende es evaluar la capacidad del niño para realizar ciertas operaciones con números, sería mejor planteárselas directamente como tales.

Es llamativo, además, que para el nivel de 6º grado se emplee solo 3 alternativas de respuesta, dado que ello implica una probabilidad de acierto al azar del 33%.

La actividad incluida en el ejemplo 10 puede en principio aparecer como irrelevante. Difícilmente alguien necesite realizar una conversión como la propuesta. Sin embargo, asumiendo que los diferentes sistemas de numeración son trabajados por los alumnos en clase, la realización de la actividad implica un nivel de abstracción importante y pone en juego varias capacidades y conocimientos. Según se explica en el Informe del cual fue tomada la actividad, para resolverla el alumno debe poner en juego los siguientes elementos:

"Recordar los símbolos de los números romanos y su valor en el Sistema Decimal:
I = 1 V = 5 X = 10 L = 50
C = 100 D = 500 M = 1.000".

"El valor de cada símbolo es siempre el mismo sin importar la posición que ocupa".

"Ningún símbolo se puede repetir más de 3 veces seguidas. L, D y M no se repiten".

"Una letra colocada a la izquierda de otra de mayor valor, le resta a esta su valor".

"Una letra colocada a la derecha de otra de igual o menor valor le suma a esta su valor".

Según se puede apreciar, si bien la actividad no tiene un contexto real, es sumamente exigente desde el punto de vista de sus exigencias cognitivas. Por cierto, se puede argumentar que así planteada la actividad puede no resultar motivadora para muchos alumnos.

Ejemplo Nº 9 **Matemática / 6º grado de Primaria**

Fuente: Ministerio de Educación y Culto / Sistema Nacional de Evaluación del Proceso Educativo (SNEPE) (1997); Primer Informe de Resultados. Sexto Grado. Asunción, Paraguay.

A) 39 caramelos
B) 30 caramelos
C) 13 caramelos

Ejemplo Nº 10 **Matemática / 6º grado de Primaria**

Fuente: Esta actividad integró la prueba aplicada por el Ministerio de Educación de Venezuela en 1998 (SINEA). Publicada en: Informe para el Docente. 6º grado. Ministerio de Educación de Venezuela, SINEA, 1999.

¿Cómo se escribe en numeración romana, el número 1998?

A) MCMXCVIII
B) MXMCCVIII
C) MCMXVIII
D) MCMCVIII

Matemática / 6° grado de Primaria **Ejemplo N° 11**

¿Cuál es la capacidad de un envase de 1 kilolitro, expresado en decilitros?

A) 1.000 decilitros

B) 10 000 decilitros

C) 100 decilitros

D) 10 decilitros

Fuente: Esta actividad integró la prueba aplicada por el Ministerio de Educación de Venezuela en 1998 (SINEA). Publicada en: Informe para el Docente. 6° grado. Ministerio de Educación de Venezuela, SINEA, 1999.

A un niño le regalan caramelos. Si la cantidad de caramelos corresponde al triple de 15 menos el triple de 2, el niño recibirá:
Las actividades incluidas en los ejemplos 11 y 12 están ambas referidas a conversión de unidades de medida. Sin embargo, son muy diferentes entre sí.

La actividad propuesta en el ejemplo 11 carece, en primer término, de un contexto de aplicación. Simplemente se pide al alumno que realice la conversión. Pero su principal debilidad es que la conversión que solicita refiere a unidades de medida cuyo uso social es inexistente. Es muy raro que algún envase de algún producto exprese el volumen de su contenido en kilolitros o decilitros. Este tipo de unidades de medida son de uso casi exclusivamente escolar, pero no son utilizadas normalmente en la vida real ni en la ciencia.

La actividad incluida en el ejemplo 12 tiene varias diferencias importantes. En primer término, refiere a unidades de medida de uso social: kilos, toneladas, metros y centímetros. En segundo lugar, la actividad propone un problema que responde a una situación propia del mundo real. Finalmente, la actividad es compleja, en la medida en que el alumno debe manejar dos variables simultáneamente (peso y altura), al tiempo que realizar las conversiones de unidades de medida.

FICHA 6

Ejemplo Nº 12 — Matemática / 6º grado de Primaria

Fuente: Esta actividad integró la prueba aplicada por la UMRE en Uruguay en 1996. Aparece publicada en: Material Informativo para Docente. Manual de Interpretación de la Prueba de Matemática. Administración Nacional de Educación Pública (1996). Montevideo,

Cuatro camiones desean pasar por un puente, pero un letrero dice:

Peso máximo: 8.500 kg
Altura máxima: 2,85 m

¿Cuál de estos camiones puede pasar?

	PESO	ALTURA
A)	8,7 t	279 cm
B)	8,5 t	3 m 84cm
C)	7,98 t.	280 cm
D)	8.400 kg	2,9 m.

Ejemplo Nº 13 — Matemática / 7º grado de Educación Básica

Fuente: Esta actividad integró la prueba aplicada por el Ministerio de Educación y Cultura de Ecuador en 1996 (APRENDO). Aparece publicada en: Análisis de las pruebas APRENDO 1996 y de sus Resultados. Matemática. Séptimo Año de Educación Básica. Ministerio de Educación y Cultura – EB/PRODEC (1998); Quito, Ecuador.

En agua salada el sonido recorre 1.400 metros por segundo. Si las ondas sonoras tardan 3,5 segundos en llegar del submarino al buzo y tardan 5 segundos en llegar del mismo submarino al barco, ¿cuál es la distancia entre el buzo y el barco?

(Nota: la versión original incluye un dibujo esquematizando la situación planteada)

A) 2.100 m
B) 4.900 m
C) 7.000 m
D) 11.900 m

El ejemplo 13 incluye otra actividad, dirigida al igual que las anteriores a alumnos que están finalizando la escuela primaria, en la que también se logra proponer al alumno una situación real que requiere ser traducida a términos matemáticos para ser resuelta a través del cálculo.

Las actividades propuestas en los ejemplos 14 y 15 están dirigidas a alumnos que finalizan la educación obligatoria. Nuevamente, se puede apreciar dos enfoques muy diferentes.

En el primer caso se prioriza la evaluación de una técnica operatoria compleja, que difícilmente una persona necesite realizar en nuestros días , y en caso de necesitarlo, sin duda recurrirá a una calculadora,[2].

En el ejemplo 15, por el contrario, se puede apreciar nuevamente el enfoque de las pruebas PISA. La actividad está situada en un contexto razonablemente plausible en la vida real, en el cual el alumno debe emplear sus conocimientos sobre proporcionalidad.

Matemática / 9° grado de Educación Básica **Ejemplo Nº 14**

¿Cuál es el resultado de racionalizar la expresión ?

$$\frac{9}{3\sqrt{2^5}}$$

A) $\frac{\sqrt{2}}{2}$

B) $\frac{3\sqrt{2}}{8}$

C) $\frac{\sqrt{2}}{8}$

D) $\frac{3\sqrt{2}}{2}$

Fuente: Esta actividad integró la prueba aplicada por el Ministerio de Educación de Venezuela en 1998. Aparece publicada en: Informe para el Docente. 9º. grado. Ministerio de Educación/SINEA (1999); Caracas, Venezuela.

2) Por cierto, es posible argumentar legítimamente a favor de las virtudes de resolver operaciones de gran complejidad sin la ayuda de la calculadora.

FICHA 6

Ejemplo Nº 15 | **Matemática / Jóvenes de 15 años escolarizados**

PAGO POR SUPERFICIE

Las personas que viven en un edificio de apartamentos deciden comprar el edificio en forma conjunta. Para ello reunirán el dinero de modo tal que cada uno deberá pagar una cantidad proporcional al tamaño de su apartamento.
Por ejemplo, un hombre que vive en un apartamento que ocupa un quinto de la superficie total de todos los apartamentos, deberá pagar un quinto del precio total del edificio.

Pregunta 1

Marca con un círculo cada una de las siguientes afirmaciones para indicar si es correcta o incorrecta.

Pregunta 2

Afirmación	Correcto / Incorrecto
La persona que vive en el apartamento más grande deberá pagar más dinero por cada metro cuadrado de su apartamento que la persona que vive en el apartamento más chico.	Correcto / Incorrecto
Si conocemos las superficies de dos apartamentos y el precio de uno de ellos, entonces podemos calcular el precio del otro.	Correcto / Incorrecto
Si conocemos el precio del edificio y cuánto va a pagar cada comprador, entonces es posible calcular la superficie de cada uno de los apartamentos.	Correcto / Incorrecto
Si el precio total del edificio fuese rebajado en un 10%, entonces cada uno de los compradores deberá pagar un 10% menos por su apartamento.	Correcto / Incorrecto

(Nota: solo se considera correcta la respuesta a la pregunta si el estudiante marca apropiadamente las cuatro afirmaciones: Incorrecto / Correcto / Incorrecto / Correcto).

Fuente: Estas preguntas son parte del estudio PISA de la OCDE, publicadas originalmente en inglés en The PISA 2003 Assessment Framework. Mathematics, Reading, Science and Problem Solving. Conocimientos y destrezas; OECD, 2003. Traducción propia.

Los formatos de pregunta son variados. En la primera pregunta se usa lo que se denomina "opción múltiple compleja". El alumno debe decidir acerca de un conjunto de afirmaciones, y la respuesta es considerada correcta únicamente si el alumno califica adecuadamente todas las afirmaciones. Esto limita las posibilidades de acierto respondiendo al azar. La segunda pregunta propuesta es de respuesta construida. El alumno debe elaborar la solución al problema que se le presenta y mostrar el proceso mediante el cual lo ha resuelto.

El edificio tiene 3 apartamentos. El apartamento 1, que es el más grande, tiene una superficie total de 95m^2. Los apartamentos 2 y 3 tienen respectivamente 85m^2 y 70m^2 de superficie. El precio de venta del edificio es 300.000 zeds.
¿Cuánto debe pagar el comprador del apartamento 2?
Muestra tus cálculos.

3. Diversos formatos de pruebas e ítemes

Según se ha podido apreciar, en las evaluaciones estandarizadas es posible emplear diversos tipos de actividades de evaluación. Cada una tiene algunas ventajas y algunos problemas. A continuación se presenta la tipología que PISA utiliza para describir los diferentes tipos de actividades[3].

3.1 Actividades de opción múltiple

Son las más comunes en las pruebas estandarizadas. Se presenta al alumno una consigna, pregunta, situación o problema y se le solicita que elija, entre 3, 4 ó 5 alternativas cuál es la respuesta adecuada.

Este tipo de actividades tiene dos ventajas principales, de tipo práctico. Primero, por estar las respuestas precodificadas, no es necesario corregirlas, con lo cual se reduce al mínimo los problemas de confiabilidad derivados de la intervención de correctores. Segundo, la digitalización de los datos se realiza rápidamente por lectura óptica.

Las actividades de opción múltiple pueden ser más o menos sencillas o complejas, según

3) Organization for Economic Cooperation and Development (OECD) (2003); The PISA 2003 Assessment Framework – Mathematics, Reading, Science and Problem Solving Knowledge and Skills. OECD, París.

hemos intentando ilustrar a través de los ejemplos. No todas las actividades de opción múltiple son elementales o requieren únicamente de la memorización de datos o hechos. Pero, al mismo tiempo, muchas competencias no son evaluables con este tipo de actividades, pues no es posible establecer alternativas plausibles. Los ejemplos 1 y 2 son paradigmáticos de competencias que requerirían de la construcción de la respuesta por el alumno.

Las actividades de opción múltiple también tienen como debilidad el hecho de que el estudiante puede responder eligiendo su respuesta al azar (si hay cuatro alternativas tiene un 25% de posibilidades de acertar la respuesta correcta) o bien descartando alternativas (lo cual, por un lado, es una forma de razonamiento en sí misma, siempre y cuando las alternativas sean plausibles, lo que no siempre sucede).

Este último es un aspecto central de las actividades de opción múltiple: las alternativas no correctas deberían responder a posibles errores de los alumnos. En el ejemplo 13, la opción C implica que el alumno sumó los tiempos en lugar de restarlos, en tanto que la opción D implica que el alumno calculó la distancia directa del barco al submarino y no desde el buzo, como pide la consigna. En ambos casos, los distractores aportan información sobre el error del alumno. En cambio la opción D, en el ejemplo 3 no tiene ningún viso de plausibilidad.

3.2 Actividades de opción múltiple complejas

Están constituidas por una serie de afirmaciones, normalmente entre 4 y 6, ante cada una de las cuales el alumno debe indicar si se trata de una afirmación correcta o incorrecta (o verdadera o falsa) en relación a la situación planteada.

Las actividades del tipo verdadero/falso no deberían ser utilizadas en forma aislada, dado que, en ese caso, la probabilidad de acierto al azar es del 50%. En cambio, cuando se plantean como una serie de afirmaciones relacionadas con una misma situación, es posible lograr una actividad cuya resolución es más compleja, al tiempo que la probabilidad de acierto por azar disminuye considerablemente.

Por ejemplo, si hay cuatro afirmaciones como en la pregunta 1 del ejemplo 15, la probabilidad de acierto al azar de las cuatro afirmaciones se reduce al 6,25% (0,5 × 0,5 × 0,5 × 0,5 = 0,0625 × 100 = 6,25%).

Este tipo de actividad permite proponer al alumno un desafío de resolución más compleja, manteniendo las virtudes de las preguntas de opción múltiple en cuanto a confiabilidad de la codificación y economía en la digitalización de los datos.

3.3 Actividades de respuesta "construida" pero "cerrada"

En muchas situaciones es imprescindible solicitar al alumno que produzca –o construya– la respuesta, en lugar de seleccionarla entre alternativas.

Las preguntas B, C y D en el ejemplo 5 ilustran este tipo de pregunta. Se las denomina "cerradas" porque, si bien el alumno debe producir la respuesta, hay una sola respuesta correcta que se expresa de manera corta e inequívoca. Esto hace que este tipo de pregunta también tenga alta confiabilidad y no requiera de un dispositivo de codificación sofisticado. Simplemente es necesario constatar si la respuesta dada por el alumno es la adecuada.

De todos modos, implica un costo adicional en relación a las de elección múltiple, en la medida en que es necesaria la intervención de un codificador que distinga entre respuestas apropiadas a inapropiadas.

3.4 Actividades de respuesta construida "abierta" y breve

La característica de este tipo de actividades es que, además de que el alumno debe producir su respuesta, la misma no está predeterminada, sino que existen diversidad de respuestas apropiadas y, eventualmente, grados de "corrección" de la respuesta, es decir, puede haber respuestas parcialmente adecuadas o correctas.

Este tipo de actividades permite evaluar en mayor profundidad capacidades más complejas en el alumno, lo que las hace relevantes y necesarias. La pregunta E en el ejemplo 5, la 2 en el ejemplo 8 y la 2 en el ejemplo 15 son ilustrativas de este tipo de actividades.

Su implementación es más costosa, en primer lugar porque, al igual que en el caso anterior, requieren de la intervención de un codificador. Pero en el caso de las preguntas de respuesta abierta, el costo de implementación se incrementa porque es necesario garantizar la confiabilidad de la codificación.

Se requiere elaborar un buen manual de codificación, entrenar intensivamente a un equipo de correctores para que codifiquen de la misma manera las respuestas de los alumnos, y montar un dispositivo de control de la confiabilidad de las codificaciones asignadas, es decir, controlar que los correctores corrijan de acuerdo al manual y que todos ellos lo apliquen de la misma manera.

Normalmente esto se efectúa haciendo que distintos correctores codifiquen por separado un mismo conjunto de trabajos de los alumnos, sin conocer las codificaciones asignadas por los demás correctores. Luego se digitan y comparan los códigos asignados y se analiza estadísticamente el grado en que los códigos asignados por distintos correctores a un mismo trabajo son coincidentes.

3.5 Actividades de respuesta construida "abierta" y extendida

Este tipo de actividad es similar al anterior. Difieren en cuanto a la extensión –y por lo tanto complejidad– de la respuesta requerida al alumno, lo que también hace más compleja la codificación.

Este tipo de preguntas se utiliza cuando se quiere ver el proceso de resolución de una situación que el alumno ha seguido –os pasos y decisiones por los que transitó– o cuando se espera que el alumno sea capaz de formular un argumento o reflexión, o justificar sus decisiones o puntos de vista. Las preguntas 3 y 4 del ejemplo 8 ilustran este tipo de actividad.

3.6 Evaluación de la producción escrita

La evaluación de la competencia de los alumnos para expresarse por escrito es un problema aparte, sumamente complejo, pero que merece ser mencionado en este lugar.

La producción escrita no forma parte de las evaluaciones internacionales. Fue incluida en la evaluación LLECE pero no se llegó a producir un reporte al respecto. Si bien suele ser evaluada en las evaluaciones nacionales, no siempre es luego procesada, analizada y reportada.

Si la codificación de preguntas de respuesta abierta y extendida es compleja y costosa, mucho más lo es la corrección y codificación de producciones escritas de los alumnos. Es

también difícil construir una consigna apropiada, que estimule a los alumnos a escribir y a hacerlo en lenguaje estándar.

No es este el lugar para detenerse a analizar el tema. Simplemente se considera conveniente señalarlo, dado que la capacidad para expresarse por escrito es de fuerte relevancia social, es imprescindible para la continuación de estudios y para muchos contextos laborales, así como un medio de consolidación del pensamiento. Por ello es importante que la evaluación de esta competencia no sea omitida, a pesar de su complejidad y costo.

Síntesis final

A lo largo de esta Ficha el lector pudo conocer diversas actividades de pruebas estandarizadas y ver cómo las mismas requieren de los alumnos distintos tipos y niveles de competencia. De allí la importancia de analizar el tipo de actividades propuestas a los alumnos antes de ocuparse de los números de una evaluación. Los números solo tienen sentido en función de las actividades propuestas. Las actividades responden a un marco conceptual ("referente") y a una serie de decisiones sobre qué se quiere evaluar. Analizar estos marcos conceptuales es más complejo, pero este análisis también debe ser realizado por el usuario inteligente de las evaluaciones. Por eso, aquí se optó por comenzar por ver la diversidad de actividades, cuyas diferencias son más fácilmente perceptibles que las relacionadas con los marcos conceptuales. A este otro tema está dedicada la Ficha 7.

El análisis de la diversidad de actividades de evaluación intenta también ayudar al lector a comprender el análisis de costos y beneficios que es necesario realizar a la hora de diseñar una evaluación. Si se quiere que la evaluación aborde capacidades complejas de los alumnos, es necesario tener presente que ello implica costos y tiempos importantes en el proceso de corrección y codificación, tanto mayores cuanto mayores sean los tamaños de las muestras que se pretende alcanzar.

Al mismo tiempo, es importante señalar que uno de los problemas de realizar evaluaciones de carácter censal –por ejemplo cuando se pretende tener resultados para todos los establecimientos educativos de un país o para todos los docentes– es que, o bien los costos se vuelven prohibitivos, o bien la evaluación deberá omitir el uso de actividades abiertas y, por tanto, la evaluación de las competencias más complejas, lo cual puede tener un efecto contraproducente en cuanto envía una señal al sistema educativo y en cuanto lo que se evalúa siempre tiene una influencia directa en lo que se enseña.

FICHA
6

¿QUÉ EVALÚA ESTA PRUEBA II ?

Contenidos currículo y competencias

Además de analizar el tipo de actividades incluidas en la prueba, la lectura inteligente de los resultados de una evaluación requiere conocer cómo fue definido conceptualmente aquello que se buscó evaluar. De hecho, las diferencias entre las actividades se derivan de diferencias en las definiciones conceptuales.

Esta tarea debería ser realizada por el lector antes de dirigir su atención hacia los datos numéricos. Sin embargo, es una labor compleja porque requiere de conocimientos especializados que el lector puede no poseer. En realidad, los responsables de la evaluación deberían preocuparse por hacer esto comprensible para sus usuarios.

Dada la importancia del tema, en esta Ficha se intentará explicar las diferencias, no fácilmente perceptibles para el lector no especializado, entre una evaluación centrada en competencias y otra centrada en contenidos curriculares. Se explicará también las distintas formas de relación entre evaluación estandarizada y currículo. Para ello se utilizará casos específicos de pruebas nacionales e internacionales aplicadas en América Latina.

La presentación de qué fue lo evaluado aparece de distintas maneras en los reportes de evaluaciones. Las principales diferencias son:

a) El lugar del reporte en que aparecen estas definiciones. En algunos reportes aparecen al inicio, para explicar al lector sobre qué versan los datos que se le presentará luego. En tanto, en otros reportes aparecen como anexo, al final del reporte, con lo cual el énfasis se pone principalmente en los datos numéricos.

b) El contenido de estas definiciones. En algunos casos el mismo se limita a listas de contenidos curriculares. En otros se profundiza en una descripción de lo que se esperaba que los alumnos fuesen capaces de hacer y, más allá, en la descripción de niveles de desempeño a través de los cuales se informa sobre en qué medida los alumnos dominan cada competencia evaluada.

FICHA 7

Los siguientes ejemplos, que por razones de espacio limitamos al área del Lenguaje, ilustran estas diferencias.

Dos ejemplos de evaluación de Lenguaje al final de la educación media básica

A continuación se contraponen dos enfoques completamente opuestos para la evaluación en el área de Lenguaje al final de la educación obligatoria:

un enfoque orientado a la evaluación de contenidos curriculares de 9° grado (SINEA) en Venezuela;

un enfoque orientado a la evaluación de competencias de Lectura (PISA).

En el ejemplo 1 se transcribe la "Tabla de Especificaciones" de la evaluación de Lenguaje en 9° grado realizada por el SINEA en Venezuela en 1998. Esta tabla es la que sustenta ítemes como el incluido en el ejemplo 7 de la Ficha 6.

El reporte del SINEA comienza indicando al lector que la prueba de Lenguaje está sustentada en el programa de la asignatura "Castellano y Literatura", el cual propone los siguientes "objetivos finales" para el ciclo 7°-9° grado:

"Expresar juicios, opiniones, ideas, sentimientos y creencias de manera clara y coherente en cualquier situación comunicativa.

"Demostrar que utiliza destrezas de lectura crítica en la presentación de los trabajos de investigación, en el estudio independiente y en el análisis de textos diversos.

"Demostrar que percibe las obras literarias como fuente de placer, de información y como medio de enriquecimiento personal.

"Producir textos escritos de carácter funcional y artístico, donde se refleje su actitud crítica y exprese su creatividad y sensibilidad".

Sin embargo, estos objetivos no son considerados en el análisis de los resultados. Las pruebas no dan cuenta de ellos.

Tabla de Especificaciones de Lengua de 9° grado – SINEA, Venezuela **Ejemplo I**

Área de Contenido	Objetivos Específicos	Nivel de razonamiento
	3.1. Tipo de Narrador	C.I.
	3.1. Organización de planos narrativos	C.C.
	3.1. Personajes	C.I.
	3.1. Secuencias	C.C.
	3.1. Realidad y fantasía	C.C.
Comprensión de la Lengua Escrita	2.2. Idea principal	C.I.
	2.2. Significado por contexto	C.I.
	2.2. Vocabulario	C.I.
	2.2. Idea principal	C.I.
	2.2. Causa-efecto	C.I.
	2.2. Consecuencia	C.I.
	2.2. Generalización	C.I.
	2.2. Conclusiones	C.I.
	2.3. Hecho y opinión	C.C.
	2.3. Comparación y contraste de opiniones	C.C.
	2.1. Localización de información	C.L.
	2.3. Técnicas de publicidad y propaganda	C.C.
	2.3. Detección de prejuicios y estereotipos	C.C.
	2.2. Paráfrasis	C.I.
	2.2. Resumen	C.I.
Nociones Lingüísticas	3.2. Recurso semántico	R
	3.2. Tipo de verso	R
	3.2. Recurso semántico	R
	3.2. Tipo de rima	R

Fuente: Ministerio de Educación / SINEA (1999); Informe para el Docente. 9no. grado. Caracas, Venezuela.

Dos páginas más adelante de dicha definición de objetivos se indica que, a partir de la revisión de los programas vigentes, se seleccionaron *"aquellos objetivos evaluables mediante pruebas objetivas. Para el caso de Lengua, la prueba se dividió en 3 grandes tópicos: Nociones Lingüísticas, Comprensión de la Lengua Escrita y Producción Escrita"*. Se aclara, además, que no se tuvo en cuenta los aspectos relativos a "hablar" y "escuchar", dado el carácter escrito de la prueba.

Finalmente se indica que *"se clasificaron los objetivos según los niveles de procesamiento cognoscitivo. Para ello se utilizaron diferentes taxonomías o conjuntos de criterios de clasificación. Para Nociones Lingüísticas, los niveles utilizados fueron: reconocimiento (R), clasificación (C), análisis (A) y aplicación de conceptos (A.C.). En Comprensión de la Lengua Escrita, los niveles fueron: comprensión literal (C.L.), comprensión inferencial (C.I.) y comprensión crítica (C.C.)"*, y se aclara que producción escrita, por su complejidad, será incluida en un informe posterior. No se brinda ninguna definición sobre el significado de cada una de los "niveles de procesamiento cognoscitivo" mencionados.

Inmediatamente, se pasa a informar los resultados a través de puntuaciones promedio para el conjunto de la prueba y para los tópicos, es decir, los resultados se abren para Comprensión de la Lengua Escrita y para Nociones Lingüísticas, pero no para los *"niveles de procesamiento cognoscitivo"*.

En un anexo del reporte se encuentra la "Tabla de Especificaciones" del ejemplo 1 (la tabla original incluye información relativa a qué ítem estaba destinado a cada objetivo, cuál fue su nivel de dificultad y cuál era la alternativa correcta).

Según se puede apreciar, la Tabla de Especificaciones es un listado de objetivos curriculares. No contempla la totalidad de los "niveles de procesamiento cognoscitivo" (por ejemplo, en Nociones Lingüísticas solo se evalúa Reconocimiento).

El informe en general no permite hacerse una idea acerca de en qué situación están los alumnos en relación a los objetivos finales estipulados para el ciclo, que aparecen como interesantes y relevantes.

En resumen, el informe no ofrece lo que promete y el lector no llega a hacerse una idea cabal de qué son capaces de hacer los alumnos en materia de lectura.

Recuadro 1

"Tabla de Especificaciones"

Es un instrumento para la elaboración de las pruebas. En él se consignan en forma esquemática los conocimientos, contenidos, objetivos, competencias (se emplean diferentes denominaciones, que explicaremos más adelante) que serán objeto de evaluación.

Incluye además la indicación acerca de qué ítemes o actividades de la prueba corresponden a cada contenido u objetivo (por razones de espacio no hemos incluido este aspecto en los ejemplos propuestos).

De esta manera, la Tabla de Especificaciones permite apreciar qué es lo que pretendía evaluar cada ítem de la prueba, permite garantizar que sean cubiertos todos los aspectos relevantes del referente, y permite apreciar el peso en cantidad de ítemes que tiene cada aspecto.

En contraste con el ejemplo 1, la Evaluación Internacional PISA evalúa la competencia lectora de los jóvenes de 15 años, definida como **"la capacidad para comprender, utilizar y reflexionar sobre textos escritos, con la finalidad de lograr los propios objetivos, desarrollar el conocimiento y potencial personal, y participar en la sociedad".**

Para ello, PISA considera cinco procesos principales para la competencia lectora:

- la comprensión global del texto;
- la obtención de información específica;
- la elaboración de una interpretación del texto;
- la reflexión sobre el contenido del texto;
- la reflexión sobre la forma y estructura del texto.

El reporte inicial de PISA incluye un capítulo entero destinado a explicar el significado de estas definiciones, así como los niveles de desempeño construidos, antes de reportar algún dato estadístico.

En el ejemplo 2, se puede apreciar el modo en que están definidos los tres niveles de desempeño que emplea PISA para reportar los resultados en Lectura. Estas escalas, además, están ilustradas en el informe con ejemplos de las actividades que son capaces de realizar los alumnos en cada nivel. Recién después, el reporte incluye los porcentajes de alumnos de cada país que fueron ubicados en cada uno de los niveles de desempeño.

Las actividades de prueba incluidas en el ejemplo 8 de la Ficha 6 corresponden a este marco conceptual. De las preguntas incluidas en dicho ejemplo, las preguntas 1 y 2 corresponden a la escala de interpretación de textos, en tanto las preguntas 3 y 4 corresponden a la escala de reflexión y evaluación de textos.

La comparación entre los ejemplos 1 y 2 permite apreciar dos enfoques diferentes para la evaluación de la lectura en alumnos próximos a finalizar la educación obligatoria, así como dos modos diferentes de informar acerca de los resultados.

2. Ejemplos de evaluación de Lenguaje en la educación primaria

Los ejemplos 3, 4 y 5 pretenden ilustrar diferencias similares en la evaluación del lenguaje en alumnos de los primeros años de la educación primaria.

El ejemplo 3 incluye la Tabla de Especificaciones de la evaluación de Lenguaje realizada por el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), tal como aparece en un anexo del Primer Informe publicado en 1998.

Este anexo es toda la información que se brinda al lector acerca de qué fue evaluado en Lenguaje en ese Primer Informe. En el texto principal únicamente se indica al lector que, para elaborar la prueba, se construyó una "matriz de objetivos curriculares" a partir del análisis de "lo que se enseñaba en Lenguaje en los cuatro primeros años del Primer Ciclo de Educación General Básica o Primaria", y que dicha matriz de objetivos curriculares fue aprobada por los países participantes.

Ejemplo 2

Niveles de desempeño en competencia lectora en PISA

Nivel	Comprensión y obtención de información	Interpretación de los textos	Reflexión y evaluación de los textos
5	Ubicar y posiblemente ordenar secuencialmente o combinar múltiples fragmentos de información profundamente incrustada, alguna de la cual puede encontrarse fuera del cuerpo principal del texto. Inferir qué información del texto es relevante para la realización de la tarea. Manejar información altamente verosímil y/o extensa, cuyos contenidos compiten, es decir, son todos plausibles.	Construir el significado de lenguaje cargado de matices o sutilezas, o demostrar una comprensión total y detallada de un texto.	Evaluar críticamente o establecer hipótesis, haciendo uso de conocimiento especializado. Manejar conceptos contrarios a lo que podría esperar el lector, y desarrollar una comprensión en profundidad de textos largos o complejos.
4	Ubicar y posiblemente ordenar secuencialmente o combinar múltiples fragmentos de información incrustada, para lo cual se requiere cumplir con múltiples criterios, en un texto de contexto o formato poco familiares. Inferir qué información del texto es relevante para la realización de la tarea.	Desarrollar un alto nivel de inferencia basada en el texto, para comprender y aplicar categorías en un contexto poco familiar, y construir el significado de una sección de texto tomando en cuenta el texto como un todo. Tratar con ambigüedades, ideas que son contrarias a lo esperado e ideas que son enunciadas en forma negativa.	Emplear conocimiento formal o público para establecer hipótesis acerca de un texto o evaluarlo críticamente. Mostrar una comprensión precisa de textos largos o complejos.

Fuente: Publicado originalmente en inglés en Knowledge and Skills for Life. First Results from PISA 2000; OECD, 2001. Adaptado de la traducción incluida en el Informe Nacional de Perú, Una aproximación a la alfabetización lectora de los estudiantes peruanos de 15 años; UMC, 2003.

3	Ubicar y, en algunos casos, reconocer la relación entre fragmentos de información, para lo cual se requiere cumplir con múltiples criterios. Manejar información cuyos contenidos compiten.	Integrar varias partes de un texto con el fin de identificar una idea central, comprender una relación o construir el significado de una palabra o una frase. Comparar, contrastar o categorizar tomando varios criterios en cuenta. Manejar información cuyos contenidos compiten.	Realizar conexiones o comparaciones, dar explicaciones o evaluar una característica del texto. Demostrar una comprensión detallada del texto en relación con conocimientos familiares y cotidianos, o hacer uso de conocimientos menos comunes.
2	Ubicar uno o más fragmentos de información que deben cumplir con múltiples criterios. Tratar con información cuyos contenidos compiten.	Identificar la idea central de un texto, comprender relaciones, formular o aplicar categorías simples o construir significados dentro de una parte limitada del texto cuando la información no es prominente y se requiere hacer inferencias de bajo nivel.	Realizar comparaciones o conexiones entre el texto y el conocimiento exterior, o explicar una característica del texto haciendo uso de la experiencia y actitudes personales.
1	Ubicar uno o más fragmentos independientes de información explícita, satisfaciendo un solo criterio y sin otra información que compita.	Reconocer el tema central o el propósito de un autor en un texto sobre un tema familiar, cuando la información requerida en el texto es prominente.	Realizar una conexión simple entre la información del texto y el conocimiento común y cotidiano.

Ejemplo 3

Tabla de Especificaciones de Lenguaje para 3er y 4º grado de Primaria
Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE)

Tópico: Comprensión Lectora
Identificar tipos de texto
Reconocer la función de un texto
Distinguir el emisor/destinatario de un texto
Identificar el significado de la tipografía (tipo de letra, tamaño)
Identificar el mensaje de un texto
Reconocer información específica del texto
Distinguir la secuencia temporal y causal explícita
Identificar el vocabulario en relación al sentido del texto

Tópico: Práctica Metalingüística
Usar marcas de concordancia gramatical (número, género y persona)
Identificar la función de palabras (sustantivo, adjetivo y verbo)
Usar grafías v-b, g-j, s-c-z
Usar la mayúscula adecuadamente
Usar la puntuación al nivel de párrafo y oración
Cortar las sílabas al final del renglón

Fuente: UNESCO-OREALC/Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (1998); Primer Estudio Internacional Comparativo sobre Lenguaje, Matemática y Factores Asociados en Tercero y Cuarto Grado. Santiago, Chile. No se incluye producción escrita dado que no forma parte del Informe.

En un Segundo Informe, publicado en octubre de 2000, se realiza un intento por construir niveles de desempeño de los alumnos. En este Segundo Informe sí se dedica un capítulo a explicar al lector este aspecto y se señala lo siguiente:

"El análisis por niveles de desempeño, homologable al análisis de competencias, permite reconocer las tendencias de lo que un estudiante, o un grupo de ellos, puede o no realizar e informa cómo se manifiestan en niños y niñas los distintos grados de las competencias que se enseñan, dando una visión del estado de la educación respecto a su calidad y equidad".

A partir de esta premisa, se informa al lector acerca del significado de tres grandes niveles de desempeño en Lenguaje y luego se reporta la proporción de alumnos que alcanzó

FICHA 7

Ejemplo 4

Fuente: UNESCO-OREALC/Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (2000); Primer Estudio Internacional Comparativo sobre Lenguaje, Matemática y Factores Asociados en Tercero y Cuarto Grado de la Educación Básica. Segundo Informe. Santiago, Chile.

Niveles de Desempeño en Lenguaje para 3er y 4º grado de Primaria - LLECE	
Nivel I Lectura literal primaria	Es el nivel más básico y simple de lectura e implica el reconocimiento de estructuras explícitas del nivel local: identificar los actores de un relato, los fragmentos claves en la argumentación y las relaciones explícitamente planteadas.
Nivel II Lectura de carácter literal en modo de paráfrasis	Hay aquí un grado mayor de complejidad en la lectura, que requiere una traducción de las palabras regulada por el sentido literal del texto. Hay preguntas que piden dar cuenta del texto con otras palabras, sin que sea necesaria una interpretación profunda de él.
Nivel III Lectura de carácter inferencial	En este nivel se llenan espacios vacíos del texto, se explicitan los supuestos sobre los que está estructurado, se vinculan proposiciones a nivel micro y macrotextual y se identifican distintas formas de relaciones implícitas en el texto. Aquí las preguntas exigen relacionar parte del texto en torno a un tema parcial y reconocer algunas siluetas textuales.

cada nivel en cada país. El ejemplo 4 recoge estas definiciones. Según se puede apreciar, los niveles están referidos únicamente al "tópico" definido como "comprensión lectora" en la tabla de especificaciones, pero no contemplan el "tópico" "práctica metalingüística".

Nótese la diferencia entre el ejemplo 3 y el ejemplo 4, ambos correspondientes a una misma evaluación. Mientras en el primer caso la única información que se ofrece al lector sobre lo que fue evaluado es la Tabla de Especificaciones empleada para construir la prueba –una lista de habilidades– en el segundo caso se le ofrece una explicación más elaborada y global sobre la competencia lectora de los alumnos.

En el ejemplo 5 se muestra un modo algo distinto de evaluar y reportar sobre el Lenguaje de alumnos de los primeros grados de Primaria. En este caso se trata de una evaluación de niños de preescolar (5 años de edad) y 1º y 2º grado de primaria, realizada en Uruguay,

Niveles de desempeño en "construcción de significado en la lectura" – Uruguay	Ejemplo 5

Nivel	Desempeño
1.	**Interpreta solamente por la imagen.** El alumno interpreta la idea global o parte del texto solamente a partir de los elementos icónicos que aparecen en el mismo pero no logra interpretar la parte verbal.
2.	**Interpreta algunas palabras y enunciados.** Puede interpretar algunas palabras y/o enunciados, aunque no pueda fundamentar su opinión o si lo hace es en forma inadecuada. El niño no reconoce el tema global del texto o cuando se le hacen preguntas sobre el mismo no ofrece respuestas adecuadas.
3.	**Interpreta algunos enunciados o párrafos.** Es capaz de interpretar algunos enunciados o párrafos en forma coherente, aunque aún no logre captar el tema central del texto. A veces reconoce, señalándola, alguna información explícita que se le solicita puntualmente.
4.	**Interpreta en forma global.** Interpreta el texto en forma global a partir de indicios verbales que aparecen en el mismo. Responde las preguntas que se le hacen acerca de la lectura en forma coherente. Reconoce la mayoría de la información explícita y/o es capaz de ampliarla a partir de otros datos que el niño tiene acerca del texto.
5.	**Realiza una buena síntesis del texto.** Realiza una buena síntesis del texto pudiendo relacionar elementos explícitos que aparecen en distintas partes del mismo. Por otra parte es capaz de vincular estos datos con información que no está explicitada, infiriéndola a partir de los datos solicitados o de los conocimientos previos que posee (aunque lo haga en forma parcialmente adecuada).

Fuente: Administración Nacional de Educación Pública/Gerencia de Investigación y Evaluación (2002); Los Niveles de Desempeño al Inicio de la Educación Primaria. Estudio de las competencias lingüísticas y matemáticas. Montevideo, Uruguay.

mediante pruebas individuales aplicadas en el marco de una entrevista con cada niño. En el área de la Lengua fueron evaluadas cinco grandes competencias:

- Oralidad;
- Oralización de la lectura;
- Construcción de significado en la lectura;
- Producción de textos escritos;
- Reflexiones sobre el lenguaje.

Cada una de estas competencias es luego descrita a través de cinco niveles de desempeño.

En el ejemplo 5 se consigna la descripción de la competencia "construcción del significado en la lectura". Como se puede observar, en este caso se establece niveles de desempeño para cada competencia, en lugar de hacerlo para el conjunto de la prueba, lo que permite una percepción más afinada y menos abstracta de las capacidades de los alumnos.

Luego de explicar el significado de cada una de las competencias y de los niveles de desempeño correspondientes a cada una de ellas, el Informe de esta evaluación reporta la proporción de niños de cada edad o grado que quedó ubicado en cada nivel de desempeño.

Los ejemplos 6 y 7 son otras dos muestras de enfoques diferentes, uno orientado a la evaluación de competencias –Colombia– y otro orientado a la evaluación de corte curricular -–Ecuador—.

En el caso de Colombia, los niveles están referidos globalmente al desempeño en Lenguaje, no están especificados para competencias específicas. Luego se reporta qué porcentaje de los alumnos quedó ubicado en cada nivel de desempeño.

En el caso de Ecuador los contenidos curriculares son traducidos a "destrezas específicas" que dan lugar a los ítemes. Luego se reporta qué proporción de los alumnos domina cada "destreza específica".

3. Currículo, contenidos, competencias y evaluación

De lo presentado hasta el momento en esta Ficha el lector debería aprender, en primer término, a observar en un reporte en qué grado y de qué manera se le informa acerca de qué se quiso evaluar, antes de aportar datos estadísticos.

Ejemplo 6

Niveles en Lenguaje – Secretaría de Educación de Bogotá, Colombia

Nivel de competencia	Desempeño evaluado
1.	**Reconocimiento y distinción de códigos** • Reconocer características básicas del lenguaje escrito como la convencionalidad y arbitrariedad de los signos y reglas que conforman el sistema de escritura • Reconocer los elementos básicos de una situación de comunicación: quién habla a quién, de qué modo habla, cuáles son los roles de los participantes en una comunicación • Identificar relaciones, semejanzas y diferencias entre el lenguaje de la imagen y el lenguaje verbal • Reconocer los significados de las palabras y frases del lenguaje coloquial y cotidiano
2.	**Uso comprensivo** • Inferir el significado de lo que se dice o escribe en relación con un tema o un campo de ideas • Caracterizar las semejanzas y diferencias en distintos tipos de textos • Utilizar y analizar categorías del sistema lingüístico (conectores, pronombres, adverbios, signos de puntuación) para comprender fenómenos textuales y de comunicación • Analizar las intenciones de quienes participan en la comunicación y el papel que juegan en la misma
3.	**Explicación del uso y posicionamiento crítico** • Reconstruir los mundos posibles de los textos, sus contextos y épocas representados en ellos, y sus componentes ideológicos y socioculturales • Comprender y explicar las intenciones comunicativas de los textos y la forma como se organiza el contenido de los mismos • Realizar lecturas en el modo crítico en las que el lector fija una posición o punto de vista y da cuenta de procesos persuasivos y manipulatorios • Establecer relaciones entre el contenido de un texto y el de otros textos, y entre lo que el texto le dice al lector y lo que él ya sabe

Fuente: Secretaría de Educación de la Alcaldía Mayor de Bogotá D.C. (2001); Compendio Resultados. Evaluación de Competencias Básicas en Lenguaje, Matemática y Ciencias Naturales. Grados tercero, quinto, séptimo y noveno. Bogotá, Colombia.

Algunos reportes le informan con claridad a qué refieren los números que luego encontrará en el informe, en tanto otros simplemente relatan cómo fueron elaboradas las pruebas y remiten al lector a una tabla de especificaciones en un anexo.

En segundo término, el lector debería poder distinguir entre una evaluación orientada a contenidos u objetivos curriculares y una evaluación orientada a competencias.

El primer caso admite al menos dos variantes. Puede tratarse de una evaluación que solo se proponga evaluar memorización de contenidos, hechos y datos, o puede tratarse de una evaluación que define destrezas u objetivos específicos que establecen lo que el alumno debería poder "hacer" con los contenidos (esto último es lo más común actualmente).

Si bien esta segunda variante, la más común, se asemeja a una evaluación de competencias en cuanto se propone evaluar lo que los alumnos son capaces de hacer con el conocimiento, la diferencia radica en que cada destreza u objetivo aparece como un fin en sí mismo, aislado de los demás.

La evaluación orientada a competencias, en cambio, considera a estas últimas como capacidades complejas que no pueden ser aprehendidas directamente a través de un ítem, pero que pueden ser expresadas en su complejidad a través de la descripción de niveles de desempeño. Se requiere de una operación conceptual para construir una visión acerca de la competencia a partir de los ítemes o actividades propuestos en la prueba.

En algunos casos se considera como una sola competencia global a la habilidad en el lenguaje (como en los ejemplos 4 y 6), en tanto que en otros casos se busca una descripción de competencias más específicas que constituyen la competencia lingüística (como en los ejemplos 2 y 5).

Finalmente, el lector debería ser capaz de apreciar si el propósito central de la evaluación es conocer si los alumnos dominan el currículo o si los alumnos son capaces de utilizar lo que han aprendido en contextos y situaciones propias de la vida real.

En el primer caso, la evaluación estará más autocontenida en los límites del currículo y del tipo de tareas y problemas típicos del ámbito escolar. En el segundo caso, la evaluación, si bien tendrá en cuenta lo que se enseña a los alumnos a través del currículo, estará más

Lenguaje y Comunicación – 7° Año – Matriz de Contenidos y Destrezas APRENDO, Ecuador **Ejemplo 7**

Áreas Temáticas	Contenidos	Destrezas Generales	Destrezas Específicas
LECTURA	COMPRENSIÓN LECTORA 1. LITERALIDAD 1.1. Elementos/ detalles/ datos 1.2.Secuencia	01. Reconocer los contenidos explícitos del texto.	01.01. Identificar elementos explícitos del texto: personajes, objetos, características, tiempo, escenarios y datos (informativos, científicos y estadísticos). 01.02. Distinguir las principales acciones o acontecimientos que arman el texto y el orden en que ellos suceden.
	2. REORGANIZACIÓN 2.1. Comparación 2.2. Clasificación 2.3. Causa-efecto 2.4. Hecho-opinión 2.5. Relaciones pronominales	02. Reconocer la organización del texto y establecer las relaciones explícitas que existen entre sus elementos.	02.01. Comparar dos elementos de un texto para identificar una semejanza o una diferencia. 02.02. Clasificar elementos mediante un criterio dado en el texto o propuesto por el evaluador. 02.03. Distinguir causa-efecto. 02.04. Diferenciar los hechos y las opiniones que contiene el texto. 02.05. Establecer las relaciones pronominales que contiene el texto.
	3. INFERENCIA 3.1. Del tema o idea principal 3.2. De ideas o significados implícitos 3.3. De conclusiones	03. Formular inferencias elementales a partir del texto.	03.01. Inferir el tema o idea principal que plantea el texto. 03.02. Inferir el significado de palabras y oraciones a partir del contexto. 03.03. Derivar conclusiones a partir del texto.

Fuente: Análisis de las pruebas APRENDO 1996 y de sus Resultados. Lenguaje y Comunicación. Séptimo Año de Educación Básica. Ministerio de Educación y Cultura – EB/PRODEC (1998); Quito, Ecuador.

preocupada de proponer situaciones propias del mundo real en que es necesario hacer uso de los conocimientos y capacidades aprendidos.

Esto se verá reflejado, por un lado, en el tipo de actividades que se propone a los alumnos en la prueba y, por otro lado, en un modo de reportar los resultados más preocupado por las competencias de los alumnos que por el dominio del currículo.

Recuadro 2

Currículo

Es el conjunto de conocimientos, habilidades, valores y experiencias que han sido seleccionadas para la formación de los estudiantes en la educación formal

Currículo Intencionado

Es el currículo que aparece formalmente establecido en los planes y programas de estudio y en el que se supone todos los alumnos tienen oportunidad de participar y aprender.

Currículo Implementado

Es el currículo al que realmente los alumnos están expuestos y tienen la oportunidad de aprender. Normalmente no coincide con el anterior, dado que los profesores y las instituciones educativas hacen sus propias selecciones sobre qué enseñar y qué exigir a los alumnos.

Currículo Aprendido

Es la parte del currículo que los alumnos efectivamente han logrado aprender a partir de lo que se les ha enseñado. No necesariamente coincide con el anterior, debidoa que muchas cosas "se enseñan", o más bien se dictan, pero en realidad no son comprendidas ni aprendidas por los alumnos.

Currículo Evaluado

Es la parte del currículo que forma parte de la evaluación. Intenta mostrar qué aspectos del currículo intencionado se ha convertido en currículo aprendido por los alumnos. Sin embargo, debe tenerse siempre presente que ninguna evaluación puede abarcar todo el currículo. Por tanto, hay aprendizajes logrados por los alumnos que no están siendo evaluados. Del mismo modo, hay aspectos relevantes del currículo intencionado y del currículo implementado que no están siendo evaluados.
Por eso debe manejarse siempre con cuidado la equiparación entre resultados de una evaluación y calidad de la educación. Las evaluaciones estandarizadas solo dan cuenta de una parte de lo que la educación se propone lograr.

Recuadro 3

Contenidos

Son todos aquellos elementos de aprendizaje que forman parte de los programas escolares. Tradicionalmente el término se refería a temas o conocimientos, pero luego se amplió la definición hacia tres tipos de contenidos.

Conceptuales	Procedimentales	Actitudinales
Son aquellos contenidos del currículo que expresan los conocimientos que los alumnos deben estudiar y comprender	Son aquellos contenidos que expresan habilidades, procedimientos prácticos, algoritmos, destrezas que los alumnos deben aprender y ser capaces de llevar a cabo por sí mismos.	Son aquellos contenidos que expresan las actitudes, valores y creencias que se espera que los alumnos desarrollen en las instituciones educativas.

A través de esta distinción se buscó ampliar el espectro de propósitos de la educación reflejados en el currículo, para que éste no se limitara a un listado de temas que los alumnos debían memorizar.

En este contexto, la relación entre evaluación y currículo deviene compleja. Por un lado, resulta obvio que la evaluación debe estar alineada al currículo, en la medida en que éste expresa lo que la sociedad ha definido como deseable para el aprendizaje y la formación de los alumnos.

Al mismo tiempo, esto depende de la calidad del currículo. Muchos currículos están formulados en términos de listados de contenidos; otros lo están en términos muy generales, que requieren de la construcción de formulaciones más específicas para poder construir instrumentos de evaluación. Finalmente, hay casos de países en que el currículo es obsoleto, por lo cual la evaluación puede ser un instrumento para propiciar el cambio curricular. En estos casos, el alineamiento entre evaluación y currículo no es deseable, porque significa reforzar un currículo obsoleto (siempre y cuando las pruebas no tengan consecuencias para los alumnos, ya que, si las tuvieran, no sería legítimo evaluar a los alumnos sobre cosas que no les han sido enseñadas).

FICHA 7

Recuadro 4

Competencias

"Una competencia es definida como la habilidad para enfrentar en forma satisfactoria demandas complejas en un contexto particular, mediante la movilización de recursos psicológicos cognitivos y no cognitivos".

Esta definición implica al menos **tres énfasis** relevantes:

a. Un énfasis central sobre ***"los resultados que el individuo logra a través de acciones, decisiones o formas de comportarse, respecto a demandas externas relacionadas con su profesión u ocupación, su rol social o su proyecto personal".*** Este énfasis pone a las competencias en relación a la capacidad del individuo para responder a los desafíos complejos que encuentra en la vida real.

b. Un énfasis en ***la estructura mental interna del individuo que involucra "conocimientos, habilidades cognitivas, habilidades prácticas, actitudes, emociones, valores y ética, motivación".*** Este aspecto es central para evitar reducir la competencia meramente a una lista de 'habilidades para' hacer ciertas cosas.

c. Un énfasis en **la dependencia respecto del contexto,** en consonancia con las teorías del aprendizaje situado. ***"Los individuos no actúan en un vacío social.*** *Las acciones siempre tienen lugar en un ambiente sociocultural, en un contexto estructurado en múltiples campos (el político, el del trabajo, la salud, la familia), cada uno de ellos consistente de un conjunto estructurado de posiciones sociales organizadas dinámicamente alrededor de un conjunto dado de intereses y desafíos sociales".*

Se asume, además, que **las competencias se aprenden y, por tanto, son enseñables.** De este modo, deben ser diferenciadas del sistema de habilidades primarias que son innatas y no aprendidas.

Las competencias ***implican un continuo que va de niveles más bajos a niveles más altos, de acuerdo a la dificultad y complejidad de los desafíos que el individuo es capaz de asumir y resolver en forma satisfactoria***. Por tanto,*"cada vez que se realizan juicios de valor sobre competencia (por ejemplo en las evaluaciones), el problema no es saber si un individuo posee o no posee un competencia particular o un componente, sino más bien determinar en qué lugar del continuo de más bajo a más alto se ubica el desempeño del individuo".*

Traducido y adaptado de: RYCHEN, D.S. & SALGANIK, L.H. (editoras) (2003); ***Key competencies for a successful Life and a Well-Functioning Society*** *(capítulo 2, "A holistic model of competence"). OECD / Hogrefe&Huber.*

Síntesis final

TIMSS vs. PISA

Evaluación orientada al currículo vs. evaluación orientada a competencias

A lo largo de esta Ficha se centró la atención en las diferencias entre la evaluación de contenidos curriculares y la evaluación de competencias. Junto con la Ficha 6, se busca alertar al lector acerca de la importancia de analizar **qué está siendo evaluado** antes de ocuparse de los resultados numéricos.

El caso de las evaluaciones internacionales TIMSS y PISA sirve para ilustrar y resumir las diferencias de enfoque.

En el año 2004, ambos estudios presentaron resultados de Matemática y Ciencias: en el caso de PISA para los alumnos de 15 años, en el caso de TIMSS para los alumnos de 4°, 8° y 12° grado.

Muchos países han participado en ambas evaluaciones y es normal que un mismo país aparezca en situaciones distintas en una y otra evaluación. Por ejemplo, en Matemáticas de 8° grado, en TIMSS 2003, Nueva Zelanda (494 puntos) aparece con un promedio inferior al de los Estados Unidos (504 puntos), en tanto que en PISA 2003 Nueva Zelanda (523 puntos) queda ubicada muy por encima de los Estados Unidos (483 puntos).

La razón es bastante simple.

El "referente" de TIMSS está constituido por los currículos de los países participantes. Como base para la elaboración de las pruebas se analizaron los currículos de los países y se construyó un marco curricular común y la correspondiente tabla de especificaciones. El tipo de actividades de prueba que propone TIMSS es bastante cercano al tipo de actividades empleadas en la enseñanza y en la evaluación al interior del sistema escolar.

El "referente" de PISA está constituido por una conceptualización de los conocimientos y competencias necesarios para desempeñarse en la vida adulta, tal como ya se vio para el caso de la evaluación de Lectura, y se hace similarmente para Matemática y Ciencias–. Por tanto, si bien se tiene en cuenta los currículos a la hora de seleccionar los contenidos de las pruebas, estas no buscan reflejar los currículos

FICHA 7

de los países participantes. El tipo de actividades de prueba que propone PISA intenta salir del ámbito escolar y proponer a los estudiantes situaciones y contextos propias del mundo real.

Un determinado país puede tener buenos resultados en TIMSS (si su currículo está bastante reflejado en las pruebas y si es bien enseñado por los profesores) pero tener malos resultados en PISA (si el currículo no prepara a los alumnos para el tipo de competencias y situaciones reales que evalúa PISA). Otro país puede tener buenos resultados en PISA (si sus alumnos están bien preparados para el tipo de competencias y situaciones que PISA evalúa) pero tener malos resultados en TIMSS (si su currículo nacional no está suficientemente reflejado en las pruebas TIMSS o si el mismo no es suficientemente dominado por los alumnos).

Éste es un ejemplo de la importancia que tiene para el lector conocer qué está siendo evaluado por una prueba, como paso previo para poder interpretar los resultados en forma apropiada.

FICHA 8

¿QUÉ SIGNIFICAN LOS NÚMEROS DE LAS EVALUACIONES? (I)

Elementos básicos para comprender los datos estadísticos

Las Fichas 8 y 9 tienen como objetivo ayudar al lector a comprender el significado de los distintos tipos de "datos numéricos" que se utiliza para reportar los resultados de las evaluaciones estandarizadas.

Para ello se intentará explicar, de manera accesible, una serie de conceptos imprescindibles para comprender los datos que se incluyen en los reportes.

En esta Ficha se comienza por una explicación básica acerca de los dos grandes modelos existentes para la construcción de pruebas estandarizadas y el procesamiento de sus resultados: la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). Cada uno de estos modelos produce "datos numéricos" que tienen distintos significados y es imprescindible que el usuario de evaluaciones estandarizadas tenga una información básica al respecto.

En segundo lugar, la Ficha explica la diferencia entre promedios y distribución de frecuencias, dado que los resultados de las evaluaciones suelen aparecer bajo una u otra de estas formas.

En el resto de esta Ficha se muestran y explican ejemplos de reportes de resultados utilizan promedios producidos tanto a partir de la TCT como a partir de la TRI. En la Ficha 9 se mostrarán ejemplos de reporte de resultados que emplean la distribución de frecuencias, también con TCT y TRI.

Finalmente, la Ficha explica un concepto estadístico importante, el de "intervalo de confianza" de las mediciones. Este concepto implica tener en cuenta el margen de error que toda medición tiene (recuérdese lo dicho en la Ficha 4 acerca de la confiabilidad) a la hora de interpretar las diferencias de resultados entre países, escuelas u otro tipo de entidades.

FICHA 8

1. Los puntajes en la Teoría Clásica de los Tests y en la Teoría de Respuesta al Ítem

1.1. Teoría Clásica de los Test (TCT)

Tradicionalmente, las pruebas estandarizadas son elaboradas y calificadas de modo tal que cada ítem o actividad propuesta a los alumnos vale 1 punto si la respuesta es correcta y 0 si es incorrecta.

Por lo tanto, todos los ítemes tienen el mismo valor en los puntajes de los alumnos, independientemente de su grado de dificultad. Tanto un ítem fácil como un ítem muy difícil valen 1 punto.

En este marco, el puntaje de un alumno es muy fácil de calcular y de comprender. Si la prueba tiene 36 preguntas y un alumno responde correctamente 24, ese será su puntaje. La prueba tiene un puntaje mínimo ('0'), en el caso en que el alumno no responda correctamente ninguna pregunta, y un puntaje máximo, que es igual al total de preguntas de la prueba (36 en este ejemplo).

Como consecuencia, si un alumno responde correctamente 10 preguntas, entre ellas las cinco más complejas, y otro alumno responde correctamente 10 preguntas, pero ninguna de las cinco más complejas, ambos obtendrán 10 puntos en la prueba, aun cuando uno haya demostrado capacidad para responder aspectos más complejos o difíciles.

Este modelo de trabajo se denomina "Teoría Clásica de los Tests" (TCT) y es la metodología más antigua y más usada en los países de la región. Si bien es más sencilla de comprender, tiene varias limitaciones. En particular, no permite considerar la dificultad y complejidad de las tareas a la hora de establecer los puntajes y no permite establecer comparaciones en el tiempo suficientemente estables y precisas.

1.2. Teoría de Respuesta al Ítem (TRI)

La Teoría de Respuesta al Ítem (TRI) es un desarrollo matemático más sofisticado para la generación de puntajes en pruebas estandarizadas.

Parte del supuesto de que existe en cada alumno una capacidad relativa a lo que la prueba evalúa (generalmente llamada "rasgo latente"), y que dicha capacidad determina la probabilidad de que el alumno responda correctamente a cada pregunta de la prueba, según la dificultad de las mismas.

Por su complejidad matemática, es bastante difícil explicar y comprender cómo se calculan los puntajes, dado que esto se hace utilizando un modelo matemático sofisticado y un software específico de procesamiento. Pero lo importante es que el lector comprenda qué significan y qué no.

Los puntajes de TRI por lo general se expresan en una escala que tiene una media de 500 puntos[1]. Los valores de la escala pueden variar entre alrededor de 100 y 900 puntos. Pero la escala TRI no tiene un cero absoluto, como en la TCT, y tampoco un puntaje máximo (que en la TCT es el total de actividades de la prueba). El mínimo y el máximo lo determina cada aplicación. Lo que se hace con la TRI es "centrar" la escala en el promedio de habilidad de la población estudiada. Por lo tanto, el punto de referencia de la escala es la media.

Si un alumno tiene un puntaje cercano a los 500 puntos, eso significa que su nivel de capacidad está en el promedio de la población evaluada. La escala se construye de tal modo que dos tercios de los alumnos se ubican entre 400 y 600 puntos. Si un alumno tiene un puntaje de 650, eso significa que tiene una capacidad muy superior al promedio. En cambio, un puntaje de 300 puntos significa que está entre los alumnos de peor desempeño[2].

La escala de puntajes representa dos cosas a la vez: la capacidad de los individuos y, la dificultad de las preguntas de la prueba. Esto significa que una pregunta con un puntaje asociado de 500 puntos es de dificultad intermedia. Una pregunta que tiene asociado un puntaje de 700 puntos es difícil –solo la responden los alumnos más capaces– y una pregunta con un puntaje asociado de 300 puntos es fácil –todos los alumnos, aun los menos capaces, tienen posibilidades de responderla correctamente[3].

Para poder interpretar mejor la escala de puntajes, normalmente se ofrece al lector una descripción de lo que son capaces de hacer los alumnos que se ubican en distintos puntos de la misma (véase la Figura 1).

1) Tanto TIMSS como PISA utilizan 500 puntos como eje de la escala, pero también es posible usar 250 puntos (como en el caso del SIMCE en Chile y del SAEB en Brasil) u otros. La decisión de qué cifra utilizar como media es arbitraria.

2) Es importante comprender que 500 puntos no significa 500 respuestas correctas como en la TCT (esta interpretación fue realizada por algunos medios de prensa en México en algún momento). Cada alumno responde entre 30 y 50 preguntas y a partir de sus respuestas se estima su puntaje a través del modelo matemático.

3) Cada pregunta tiene asociada un puntaje en función de su dificultad. Este puntaje corresponde al nivel de capacidad de los alumnos que tienen un 50% de probabilidad de responder correctamente la pregunta.

Con la TRI no todas las preguntas tienen el mismo peso. Dos alumnos pueden haber respondido correctamente la misma cantidad de preguntas, pero el que respondió preguntas más complejas obtendrá un puntaje más alto.

Una de las virtudes de la TRI es que permite estimar la capacidad de los alumnos independientemente de la versión o formato de la prueba que se aplicó o del grupo que fue evaluado. Esto significa que no todos los alumnos deben rendir la misma prueba. Basta con que haya un conjunto de preguntas en común para que se pueda estimar el puntaje de los alumnos en la misma escala.

Lo mismo ocurre con evaluaciones sucesivas en el tiempo. Basta con que dos de estas tengan un conjunto de ítemes en común –denominados ítemes de anclaje– para poder estimar los puntajes de los alumnos de la segunda evaluación en la misma escala que la primera.

Por ejemplo, 500 puntos fue la media de la OCDE en Lectura en PISA 2000. En 2003 se usaron parte de los ítemes del 2000 que fueron conservados como confidenciales. De este modo, los puntajes de Lectura de 2003 se expresaron en la escala del 2000. Por lo tanto, el promedio de 2003 ya no fue 500. Lo hubiese sido si no hubieran cambiado las competencias de los jóvenes. Fue 494, lo que indica que los resultados empeoraron levemente respecto al año 2000.

2. Promedios y distribución de frecuencias

Independientemente del modelo de pruebas con que se trabaje, desde el punto de vista estadístico existen dos modos principales de presentar los resultados: a través de promedios o mediante frecuencias relativas.

2.1. Promedios

El promedio o media es uno de las medidas estadísticas más utilizadas para describir lo que caracteriza a un determinado grupo. Su cálculo es bien sencillo: simplemente se suma la cantidad de valores y se divide el resultado obtenido entre el número correspondiente a la cantidad de casos.

Figura 1

Mapa de actividades de Lectura en PISA 2000

Alumnos	Puntaje escala TRI	Actividades
Alumnos de más alto desempeño. Tienen al menos un 50% de probabilidad de responder correctamente a las actividades correspondientes a su puntaje. Tienen una probabilidad más alta de responder a las preguntas de puntajes inferiores.	822	HIPOTETIZAR sobre un fenómeno inesperado tomando en cuenta conocimiento externo junto con toda la información relevante de una TABLA COMPLEJA, en un tema relativamente poco familiar.
	727	ANALIZAR varios casos descritos y VINCULARLOS a categorías dadas en un DIAGRAMA DE ÁRBOL, en el cual parte de la información relevante se encuentra en notas al pie de página.
	705	HIPOTETIZAR sobre un fenómeno inesperado tomando en cuenta conocimiento externo junto con parte de la información relevante de una TABLA COMPLEJA, en un tema relativamente poco familiar
	652	EVALUAR el final de una NARRACIÓN LARGA en relación con su tema implícito.
	645	RELACIONAR MATICES DEL LENGUAJE en una NARRACIÓN LARGA con el tema principal, en presencia de ideas contradictorias.
	631	LOCALIZAR información en un DIAGRAMA DE ÁRBOL utilizando información de una nota al pie de página.
Alumnos de desempeño en torno al promedio de la población. Tienen al menos un 50% de probabilidad de responder correctamente a las actividades correspondientes a su puntaje. Tienen una probabilidad menor de responder a las preguntas de puntajes más altos. Tienen una probabilidad mayor de responder a las preguntas de puntajes más bajos.	600	HIPOTETIZAR acerca de una decisión del autor relacionando la evidencia proporcionada en una gráfica con múltiples presentaciones, con el tema principal inferido.
	581	COMPARAR Y EVALUAR los estilos de dos CARTAS abiertas.
	567	EVALUAR el final de una NARRACIÓN LARGA en relación con la trama.
	542	INFERIR UNA RELACIÓN ANALÓGICA entre dos fenómenos discutidos en una CARTA abierta.
	540	IDENTIFICAR la fecha inicial implícita en una GRÁFICA.
	537	CONECTAR evidencia de una NARRACIÓN LARGA con conceptos personales, con el fin de justificar puntos de vista opuestos.
	508	INFERIR LA RELACIÓN entre DOS PRESENTACIONES GRÁFICAS con distintas convenciones.
	485	LOCALIZAR información numérica en un DIAGRAMA DE ÁRBOL.
	480	CONECTAR evidencia de una NARRACIÓN LARGA con conceptos personales, con el fin de justificar un único punto de vista.
	478	LOCALIZAR Y COMBINAR información en una GRÁFICA DE LÍNEA y su introducción, para identificar un dato faltante.
	477	COMPRENDER la estructura de un DIAGRAMA DE ÁRBOL.
	473	RELACIONAR casos concretos con categorías presentadas en un DIAGRAMA DE ÁRBOL, cuando parte de la información relevante está en notas al pie de página.
	447	INTERPRETAR información de un único párrafo, para comprender el escenario de una NARRACIÓN.
	421	IDENTIFICAR el PROPÓSITO común de DOS TEXTOS CORTOS.
	405	LOCALIZAR elementos de información explícita en un TEXTO que contiene organizadores fuertes.
Alumnos de peor desempeño. Solo pueden responder preguntas de puntajes bajos. Casi nula probabilidad de responder preguntas de puntajes superiores.	397	INFERIR la IDEA PRINCIPAL de una GRÁFICA DE BARRAS simple, a partir de su título.
	392	LOCALIZAR un elemento de información literal en un TEXTO con una estructura textual clara.
	367	LOCALIZAR información explícita en una sección especificada de una NARRACIÓN corta.
	356	RECONOCER EL TEMA de un artículo con subtítulos claros y considerable redundancia.

Fuente: Elaboración propia a partir de Reading for Change: Performance and Engagement across Countries (OECD, 2002b)

FICHA 8

Recuadro 1

La comparación de resultados en el tiempo

Una de las informaciones que tanto los tomadores de decisiones como la opinión pública demandan a las evaluaciones estandarizadas es la relativa a cómo evolucionan los aprendizajes de los estudiantes a lo largo de los años. Sin embargo, no todos los sistemas proporcionan esta información en forma apropiada. Lo primero que se debe garantizar es que se evaluaron los mismos contenidos y competencias. Como resulta obvio, si se modifica aquello que fue evaluado, los resultados pueden cambiar por ese motivo, no necesariamente porque los alumnos estén aprendiendo mejor lo que había sido evaluado inicialmente. Si se desea medir el cambio, no se debe cambiar el instrumento de medición.

Un primer ejemplo de este problema fue constatado en Argentina a través de una investigación realizada por Silvina Larripa[4]. El trabajo muestra que, en Matemática, mientras en 1995 la competencia más simple ("Reconocimiento") tenía en la prueba un peso del 28% y la competencia más compleja ("Resolver problemas") tenía un peso del 32%, en 1998 dichos pesos cambiaron respectivamente a 45% y 20%. El trabajo muestra también que para evaluar comprensión de lectura en 1995 se utilizó un único texto de una extensión de 264 palabras, en tanto en 1998 se utilizaron dos textos, uno de 521 palabras y el otro de 122. Cuando se producen cambios de este tipo en la estructura de las pruebas es muy difícil que se pueda establecer comparaciones válidas.

Un segundo ejemplo es lo ocurrido en PISA en relación con Matemática. Como en el año 2000 fueron evaluadas solamente dos subáreas de contenidos ("Espacio y Forma" y "Cambio y Relaciones"), a las que en el año 2003 se agregó otras dos subáreas[5] ("Cantidad" y "Probabilidad e Incertidumbre"), no es posible establecer una comparación global de los resultados en Matemática entre PISA 2000 y PISA 2003. La comparación fue realizada únicamente para las subáreas evaluadas en ambos ciclos.

Trabajando con TCT es posible establecer comparaciones en el tiempo siempre y cuando las pruebas tengan la misma cantidad de preguntas, evalúen las mismas competencias y contenidos con los mismos pesos relativos, y el conjunto de actividades de la prueba tenga la misma dificultad promedio medida en una generación independiente de las que están siendo comparadas. Estas pruebas se denominan "formas equivalentes" de prueba.

La TRI ofrece mucho mayor flexibilidad y precisión para establecer las comparaciones en el tiempo. Se pueden utilizar pruebas diferentes, a condición de que exista un conjunto común que se mantiene en secreto (ítemes de anclaje). Esto permite estimar los puntajes y hacerlos comparables. A este proceso se le denomina técnicamente "equiparación de puntajes".

En cualquier caso, siempre que se establecen comparaciones con resultados de evaluaciones realizadas anteriormente, el lector debe buscar en el reporte la información relativa a los recaudos técnicos tomados para garantizar la comparabilidad.

4) Larripa, S., 2003; El Sistema Nacional de Evaluación de la Calidad Educativa (SINEC): acerca de la comparabilidad de sus resultados. Argentina, 1995-2000. Tesis de Maestría, Universidad de San Andrés, Escuela de Educación.

5) En 2003 Matemática fue el foco principal de PISA, lo que permitió ampliar el espectro de contenidos evaluados.

Por ejemplo, si se desea describir el resultado de una escuela en una prueba estandarizada a través de la media o promedio de la escuela, se suman los puntajes obtenidos por los alumnos y se divide el resultado obtenido entre la cantidad de alumnos.

Este promedio de la escuela no debe ser confundido con los promedios individuales que se utiliza muchas veces para calificar a los alumnos. Estos últimos resultan de sumar las calificaciones o notas que el alumno fue obteniendo a lo largo del año y se divide esa suma entre la cantidad de calificaciones.

El promedio o media tiene la ventaja de que permite una comparación rápida entre resultados de diferentes grupos (que pueden corresponder a los alumnos de una escuela, provincia o país) y saber cuál es más alto y cuál es más bajo.

Sin embargo, como contrapartida de la sencillez, los promedios tienen la siguiente debilidad fundamental: pueden ocultar situaciones diferentes dentro de cada grupo.

Veamos qué significa esto con un ejemplo. Los gráficos incluidos en la Figura 2 corresponden a dos escuelas, A y B, ambas con 110 estudiantes, que rindieron una misma prueba de Matemática cuyo puntaje máximo era 32 puntos.

Ambas escuelas tienen un promedio de 21,35 puntos. No obstante, a pesar de que los promedios son iguales, la "distribución" de los puntajes es bien diferente entre una y otra escuela.

En la escuela A la mayoría de los alumnos se ubica muy cerca de la media. Esto significa que los resultados en la escuela son bastante parejos. No hay alumnos con resultados muy bajos ni muy altos.

En la escuela B, en cambio, los resultados tienen mayor "dispersión", es decir, hay más alumnos en los extremos, con puntajes bajos y altos.

Por lo tanto, a pesar de que ambas escuelas tienen el mismo promedio en Matemática, los resultados son más homogéneos en la escuela A y más desiguales en la escuela B.

Esta diferencia puede tener consecuencias importantes.

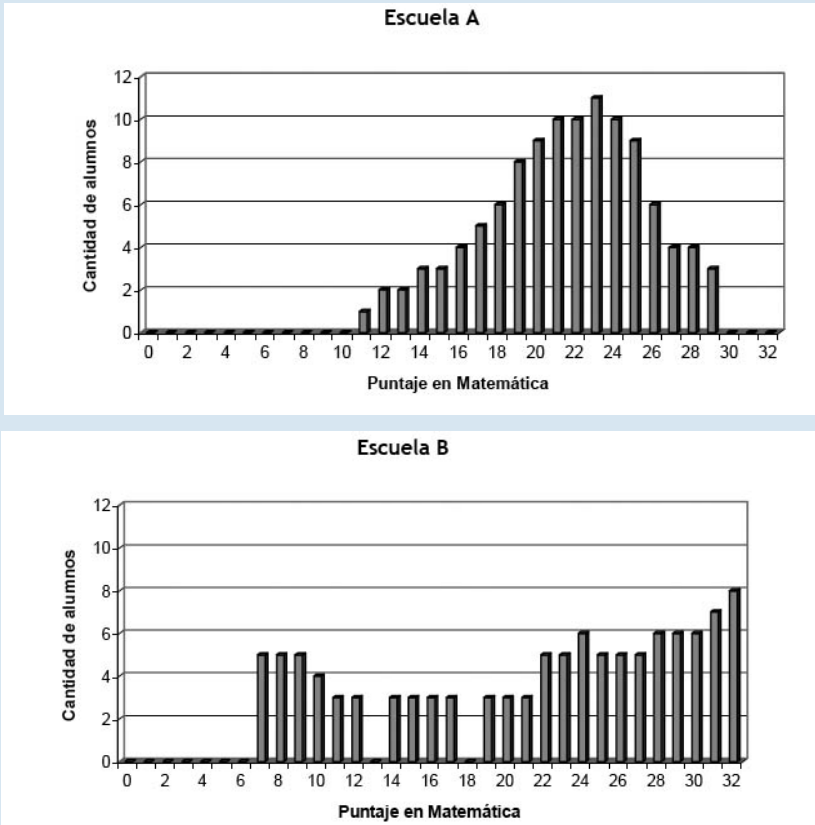
FICHA 8

Por ejemplo, si se estableciera que lograr la mitad de los puntos de la prueba (16 puntos) es un resultado satisfactorio, entonces en la escuela A 99 alumnos (el 90%) habría alcanzado dicho nivel satisfactorio, mientras que en la escuela B lo habrían logrado 79 alumnos (71,8%). La escuela A sería en este sentido "mejor" que la B (usamos deliberadamente las comillas porque, como se mostrará en la Ficha 10, esta interpretación es simplista).

Figura 2

Cantidad total de alumnos en cada escuela: 110

Promedio de los puntajes en cada escuela: 21,35



En cambio, si se considera que el resultado de un alumno es satisfactorio si res-ponde correctamente 3/4 partes de la prueba (24 puntos), en la escuela A dicho nivel es alcanzado por 36 alumnos (32,7%) y en la escuela B por 54 alumnos (49,1%). La escuela B sería "mejor" que la A.

2.2. Distribución de frecuencias

La "distribución de frecuencias" es la cantidad de casos que corresponden a cada valor, en este caso, la cantidad de alumnos que obtuvo cada uno de los puntajes posibles en la prueba.

Esto último es lo que muestran las gráficas de la Figura 2. En el eje X se representan los 32 puntos posibles que la prueba admitía, en tanto en el eje Y se representa la cantidad de alumnos que obtuvo cada puntaje (el total de alumnos en cada escuela es 110).

Se habla de "frecuencias absolutas" cuando se indica directamente la cantidad de casos –como en la Figura 2- y de "frecuencias relativas" cuando, en vez de la cantidad de casos, se indica el porcentaje que estos representan sobre el total.

Como normalmente no tiene mayor sentido reportar los porcentajes de alumnos en cada uno de los puntajes de la prueba, lo que suele hacerse es establecer "tramos" de puntaje y reportar el porcentaje de alumnos en cada "tramo".

Por ejemplo, en la Figura 2 se podrían establecer tres grandes tramos en el puntaje de la prueba de Matemática: de 0 a 12 puntos, de 13 a 24 y de 25 a 32.

En el primer tramo, la escuela A tiene 3 alumnos (2,7%), en el segundo tramo tiene 81 alumnos (73,6%) y en el tercero tiene 26 alumnos (23,6%). La escuela B tiene 25 alumnos en el primer tramo (22,7%), 37 en el segundo (33,6%) y 48 en el tercero (43,6%).

Hay diferentes modos de establecer y trabajar con estos "tramos" de puntajes, lo que será objeto de análisis en la Ficha 9.

FICHA 8

3. El reporte de resultados a través de promedios

El reporte a través de promedios puede realizarse tanto a través de puntajes de la TCT como de puntajes de TRI.

3.1. Porcentaje promedio de respuestas correctas

En la Figura 3 se presentan los promedios por provincia en la evaluación nacional realizada en Argentina en 1996. Rápidamente, uno puede constatar qué provincias tuvieron los promedios más altos y cuáles los más bajos.

El porcentaje promedio de respuestas correctas es producido a partir de la Teoría Clásica y se calcula de la siguiente manera: primero se computa el porcentaje de respuestas correctas de cada alumno (si una prueba tiene 30 preguntas y un alumno respondió correctamente 3, el porcentaje de respuestas correctas para ese alumno es 10%) y luego se calcula el promedio para todos los alumnos.

Este dato es **exactamente equivalente** al promedio de los puntajes de todos los alumnos. Para calcular el promedio de los puntajes de los alumnos, es necesario sumar todos los puntajes individuales y dividirlo entre la cantidad de alumnos.

Si el porcentaje promedio de respuestas correctas (PPRC) fue de 58,19% y la prueba tenía 30 preguntas, significa que el puntaje promedio de los alumnos fue 58,19% * 30 / 100 = 17,457 puntos. Es decir que el "porcentaje promedio de respuestas correctas" es igual que el puntaje promedio de los alumnos, calculado como porcentaje del puntaje total de la prueba.

Estos datos, al igual que los promedios de la TRI, sirven básicamente para establecer comparaciones entre entidades. El PPRC podría interpretarse en forma criterial, como indicador de que los alumnos dominan 'x' porcentaje de lo que deberían saber, solo en el caso en que los ítemes sean una buena muestra de todo lo que los alumnos deberían saber. Pero si la prueba fue construida con un enfoque normativo, eliminando los ítemes muy fáciles y muy difíciles, dicha interpretación no es válida.

ARGENTINA Porcentaje promedio de respuestas correctas por provincia **Figura 3**

3.2. Promedios en puntajes TRI

Según se explicó más arriba, los puntajes de TRI no pueden ser interpretados en términos de cantidad de preguntas respondidas correctamente, sino de probabilidad de los alumnos de responder correctamente a preguntas de distinto grado de dificultad (Figura 1 en esta Ficha).

Al igual que en el caso de la TCT, los promedios de TRI sirven principalmente para establecer comparaciones entre países (veáse la Figura 2 en la Ficha 4). En el caso de la TRI los promedios no variarán entre 0 y 100, como los porcentajes de respuestas correctas (o entre 0 y el puntaje máximo de la prueba), sino que, según se explicó antes, el punto de referencia será la media de la población evaluada, que suele ubicarse en los 500 puntos por una cuestión de conveniencia en la comunicación de los datos.

La interpretación del significado de los puntajes, como se indicó, depende de contar con una ilustración de la escala en términos de tareas, como la presentada en la Figura 1 en esta Ficha. Este tipo de descripciones suele acompañarse de ejemplos de ítemes aplicados, del estilo de los incluidos en la Ficha 6 (no se incluyen aquí por razones de espacio).

Nivel Primario - 6° Grado
Lengua

Muestra 1996

Jurisdicción	Lengua 6° Grado	
	Muestra 1996	Relación con respecto al valor medio
Capital Federal	72.39%	1.24
Mendoza	67.63%	1.16
Río Negro	66.28%	1.14
Neuquén	65.37%	1.12
La Pampa	60.58%	1.04
Chaco	60.37%	1.04
Córdoba	58.77%	1.01
Santa Cruz	58.69%	1.01
MEDIA NACIONAL	58.19%	1.00
Buenos Aires	57.88%	0.99
Santa Fe	57.87%	0.99
Entre Ríos	57.26%	0.98
Gran Buenos Aires	56.98%	0.98
Tierra del Fuego	56.87%	0.98
Misiones	55.83%	0.96
Chubut	55.41%	0.95
Salta	54.71%	0.94
Corrientes	54.12%	0.93
San Juan	53.79%	0.92
Jujuy	52.16%	0.90
La Rioja	50.14%	0.86
Tucumán	49.04%	0.84
Formosa	48.94%	0.84
Santiago del Estero	47.74%	0.82
Catamarca	46.48%	0.80

Fuente: Dirección Nacional de Evaluación, 1997. Operativo Nacional de Evaluación 1996.

3.3. La importancia de la "dispersión" de los resultados

En el apartado 2 de esta Ficha se mostró que uno de las debilidades principales que tiene reportar a través de promedios radica en que los mismos no dan cuenta de la "dispersión" interna de los resultados.

Por ejemplo, a partir de la tabla incluida en la Figura 3 de esta Ficha, una puede saber que en Córdoba el porcentaje promedio de respuestas correctas fue 58,77%, pero no puede saber si la mayoría de los alumnos estuvieron cercanos a ese resultado o si, por el contrario, hay fuertes disparidades, alumnos que lograron cerca del 100% de respuestas correctas y otros que apenas lograron un 10%. Lo mismo ocurre con los puntajes de TRI.

En la Figura 2 de la Ficha 4 el lector puede observar que el promedio de Finlandia se ubica cerca de 550 puntos, pero no puede saber si todos los alumnos están cerca de dicho puntaje o si, por el contrario, existen importantes diferencias dentro de ese país.

El tema de la "dispersión" interna de los resultados es relevante porque está vinculado a la cuestión de la equidad en el acceso a los aprendizajes.

Los reportes de resultados PISA tienen en cuenta la importancia de la información relativa a la "dispersión" que se esconde detrás de los promedios a través de un gráfico como el incluido en la Figura 4.

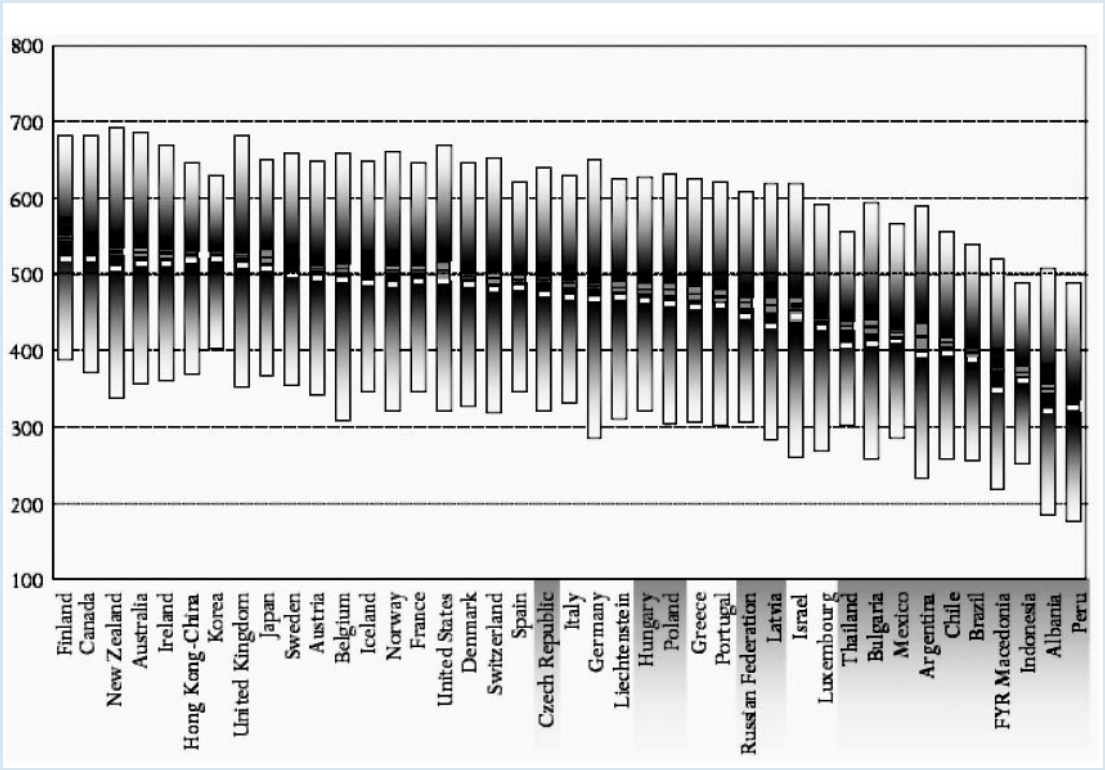
Los resultados de cada país son representados mediante una barra. Cada barra contiene varios datos referentes al país. La línea en el centro representa la media del país en cuestión. La "caja" gris representa el error estándar de medición del país (véase el próximo apartado). Las líneas negras y blancas representan los resultados de niñas y varones respectivamente. Los extremos de cada barra representan los puntajes a partir de los cuales se ubican el 5% de los alumnos de mejor (arrriba) y el 5% de los alumnos de peor desempeño (abajo).

De esta manera, el lector puede conocer no sólo el promedio del país sino la dispersión interna de sus resultados, representada por la longitud de la barra. Una barra de gran longitud indica una gran distancia entre los mejores y los peores alumnos. Una barra de reducida longitud indica que los resultados de los mejores y los peores alumnos no son tan diferentes y se acercan bastante al promedio del país.

Para comprender mejor esto último, en la Figura 5 se presenta en forma ampliada dos casos concretos extraídos de la Figura 4, los de Nueva Zelanda y Corea del Sur, países que aparecen juntos en la Figura 4 porque sus promedios son muy parecidos. Nueva Zelanda tuvo en PISA 2000 un promedio de 529 puntos en Lectura y Corea del Sur un promedio de 525.

Sin embargo, a pesar de que los promedios son muy próximos, la situación en ambos países es muy distinta.

Puntajes promedio y dispersión en Lectura, PISA 2000 **Figura 4**



Fuente: OECD/UNESCO-UIS, 2003; Literacy Skills for the World of Tomorrow: Further Results from PISA 2000.

Nueva Zelanda tiene una altísima dispersión interna en sus resultados. Sus mejores alumnos alcanzan puntajes cercanos a 700 pero, al mismo tiempo, sus peores alumnos están por debajo de los 350 puntos. Corea del Sur, en cambio, muestra una situación interna mucho más equitativa: sus mejores alumnos no tienen resultados tan destacados como los mejores de Nueva Zelanda, pero, al mismo tiempo, logra que sus peores alumnos casi no caigan por debajo de los 400 puntos.

Se puede observar, además, que en ambos países las niñas tienen mejores resultados en Lectura que los varones. Pero, nuevamente, las distancias entre ellos son bastante más importantes en Nueva Zelanda que en Corea del Sur.

De este modo, PISA reporta, junto con los promedios de los países, información relevante acerca de la equidad interna de los resultados en cada país.

4. Error estándar de medición y significación estadística de las diferencias

Otro elemento central que es necesario tener en cuenta al analizar los resultados de las evaluaciones es que, según fue explicado en la Ficha 4 al estudiar el tema de la confiabilidad, toda medida está sujeta a la posibilidad de errores de precisión.

Para controlar este problema existen procedimientos estadísticos para calcular el rango dentro del cual, con una alta probabilidad, puede ubicarse la media.

Por lo tanto, toda medida tiene asociado un determinado "error estándar". No hay resultados "exactos", sino estimaciones de valores que pueden variar dentro de ciertos rangos por encima y por debajo del valor establecido para la media.

A este rango de variación se le denomina técnicamente "margen de error" o "intervalo de confianza". Está representado en las Figuras 4 y 5 por las "cajitas" grises que rodean a la media de cada país.

¿Qué significan estas "cajitas"? Que el promedio de cada país puede no ser exactamente su media, sino que puede ser cualquier valor que se ubique dentro de la "cajita" gris. Esto

se establece "con un 95% de confianza", lo cual significa que existe una probabilidad del 95% de que la media del país sea un valor dentro del "intervalo de confianza".

¿Por qué es esto importante? Porque significa que no se debe ordenar sin más ni más a los países según sus promedios. Deben ser tenidos en cuenta también los "intervalos de confianza". Si los mismos se superponen para dos países, ello significa que, en realidad, los resultados de dichos países podrían no ser diferentes. Cualquiera de los dos podría estar

Los 7 "mejores" países de PISA 2000 en Lectura **Figura 5**

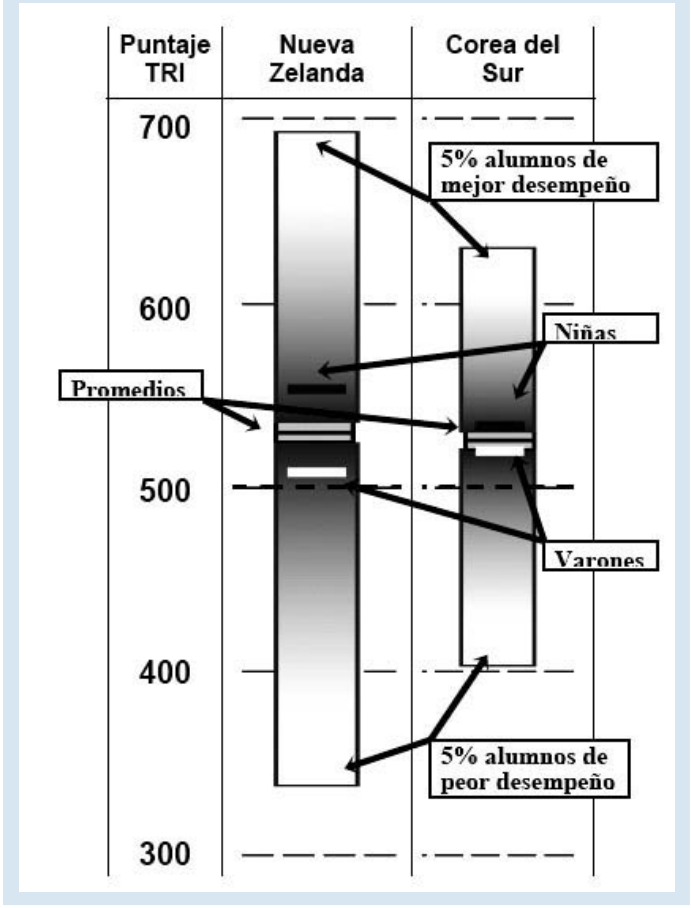
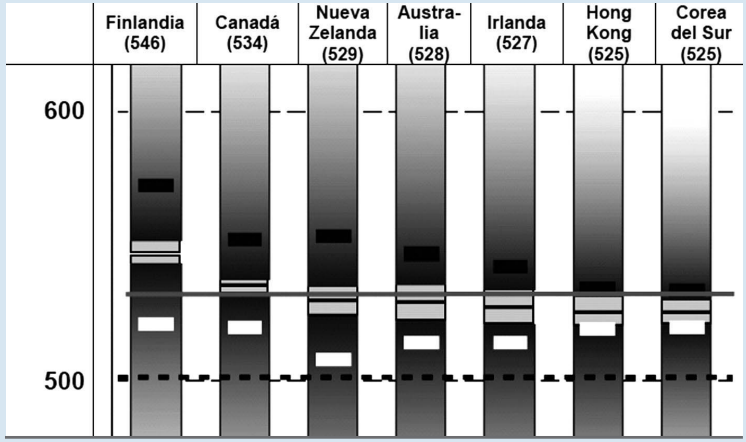


Figura 6 **Los 7 "mejores" países de PISA 2000 en Lectura**



por encima o por debajo del otro. Para que la diferencia de promedios entre dos países sea "significativa" es necesario que los "intervalos de confianza" no tengan puntos de contacto.

En la Figura 6 se amplifica la zona de la Figura 4 correspondiente a las "cajitas" de los 7 primeros países en Lectura de PISA 2000. Los números entre paréntesis son los promedios de cada país.

Según es posible apreciar, el resultado de Finlandia es superior al del resto. Su "cajita" no tiene zona de contacto con las demás. Esto quiere decir que la diferencia entre el promedio de Finlandia y los países que le siguen es "significativa". Es decir, aun considerando el margen de error, es posible afirmar que Finlandia tuvo un resultado superior a los restantes países.

Pero luego siguen cuatro países cuyos intervalos de confianza o márgenes de error se superponen: Canadá, Nueva Zelanda, Australia e Irlanda. Si bien sus promedios son diferentes, en realidad estos cuatro países no pueden ser ordenados; cualquiera de ellos podría ocupar el segundo lugar después de Finlandia.

La línea roja introducida en la Figura 6 permite apreciar que la diferencia de Canadá respecto a Corea del Sur y Hong Kong sí es significativa. Canadá obtuvo un resultado superior a estos dos países. En cambio, estos dos países no se diferencian de Nueva Zelanda, Australia e Irlanda.

El aspecto que acabamos de analizar tiene fuertes implicancias para la construcción de *rankings*, según será analizado en la Ficha 10.

Al mismo tiempo, pone de manifiesto uno de las problemas técnicos importantes de la mayoría de los reportes de resultados en América Latina: por lo general no se informa acerca de los márgenes de error.

Ello significa que, en la mayoría de los casos, no es posible saber si las diferencias reportadas son significativas. Por ejemplo, en el caso de Argentina, presentado en la Figura 3, no es posible saber cuándo la diferencia entre dos provincias es significativa y cuándo no lo es.

Por último, debe decirse que el hecho de que una diferencia de promedios sea estadísticamente significativa no quiere decir que sea importante.

Por ejemplo, acabamos de ver que la diferencia entre Canadá, por un lado, y Corea del Sur y Hong Kong, por otro, es estadísticamente significativa. Ello implica que se puede afirmar que Canadá tuvo un mejor resultado. La diferencia de promedios se ubica alrededor de los 9 puntos (podría ser algo mayor o menor en función de los márgenes de error).

Una diferencia de 9 puntos, en una escala que tiene una media de 500, es más bien modesta y no permite afirmar que la educación canadiense sea sustantivamente "mejor" que la de Corea del Sur y Hong Kong. Para hacer este tipo de juicios de valor es necesario recurrir a un conjunto de informaciones más amplia.

Por tanto, estadísticamente significativo implica que hay diferencias reales, pero no necesariamente que estas sean sustantivas y relevantes.

FICHA 8

Síntesis final

Esta Ficha, junto con la que sigue, tienen como propósito ayudar al lector a comprender mejor los datos numéricos que aparecen en los reportes de las evaluaciones estandarizadas.

Una primera distinción importante que el lector debe realizar está referida al modelo de medición desde el cual fueron construidas las pruebas y, por tanto, los resultados. Dos son los modelos que se emplean: la Teoría Clásica de los Tests y la Teoría de Respuesta al Ítem.

En la primera, los puntajes se calculan simplemente sumando el total de las respuestas correctas de cada alumno y la mayoría de los reportes informan acerca del porcentaje promedio de respuestas correctas dadas por los alumnos.

En la segunda se utiliza una escala que no tiene un "cero absoluto" ni un puntaje máximo, sino que el eje de referencia es la media de la población evaluada, que suele establecerse arbitrariamente en 500 puntos o en 250 puntos. El significado de la escala suele ilustrarse indicando qué tipo de actividades son capaces de resolver los alumnos con diferentes puntajes.

Una segunda distinción importante que el lector debe realizar es entre los promedios y las frecuencias relativas. En esta Ficha se analizaron ejemplos de reportes que emplean promedios. El principal problema que el lector debe tener presente cuando se le presentan promedios es que los mismos son una medida muy resumida que puede esconder situaciones muy distintas en términos de la diferencias internas en los resultados.

Para ello es importante que los reportes informen no solo acerca de los promedios, sino también acerca de la "dispersión" de los puntajes, que está fuertemente relacionada con la equidad o inequidad en el logro de los aprendizajes evaluados.

La Ficha destaca otros dos aspectos que el lector debe tener en cuenta al analizar resultados de evaluaciones.

Uno es el relativo a cómo se garantiza la comparabilidad de los resultados cuando se establece comparaciones entre mediciones realizadas en distintos momentos del tiempo.

El segundo es el relativo a los márgenes de error de las mediciones. Estos son importantes porque determinan si las diferencias en los resultados entre entidades son o no significativas desde el punto de vista estadístico. Todo reporte de resultados de evaluaciones estandarizadas debería incluir información relativa a los márgenes de error de los datos que se entrega.

¿QUÉ SIGNIFICAN LOS NÚMEROS DE LAS EVALUACIONES? (II)

Elementos básicos para comprender los datos estadísticos

La Ficha 9 es una continuación de la anterior, en la que se explicaron las características básicas de la Teoría Clásica de los Tests y la Teoría de Respuesta al Ítem, así como la diferencia entre el reporte a través de promedios y el reporte a través de distribución de frecuencias. Se mostró, asimismo, que los promedios aportan poca información respecto a qué son capaces de hacer los alumnos y que pueden esconder información importante respecto a la dispersión de los resultados.

La Ficha 9 está focalizada en el reporte a través de la distribución de los alumnos en categorías o niveles de desempeño. El tema central en este caso es qué son esos niveles y cómo se construyen.

Para explicar e ilustrar este aspecto se proponen diversos ejemplos tomados de reportes reales.

El tema de la Ficha hace necesario volver sobre aspectos tratados en Fichas anteriores, en especial la Ficha 3 relativa a los enfoques normativos y criteriales. En particular, se explica cómo es posible –y necesario en las evaluaciones nacionales– establecer un "estándar" o expectativa respecto a cuál es el nivel de desempeño que se espera que los alumnos alcancen al finalizar un determinado grado o nivel del sistema educativo.

1. El reporte a través de categorías de desempeño de los alumnos

En el ejemplo de las dos escuelas utilizado en la Figura 2 de la Ficha 8 se explicó que, a pesar de que ambas escuelas tenían el mismo promedio, la interpretación de sus resultados podría ser diferente según qué puntaje de la prueba fuese considerado un indicador de que el alumno había aprendido satisfactoriamente lo que se esperaba.

También se indicó que los promedios constituyen una abstracción que no da cuenta de lo

que los alumnos son capaces de hacer y que puede dar lugar a falsas impresiones en cuanto a la calidad de la educación en los países o provincias, si no se tiene en cuenta el error estándar de medición o, aun teniéndolo en cuenta, si se utiliza el promedio como indicador único de calidad.

Un modo distinto de presentar y analizar los resultados –complementario a los promedios pero imprescindible– consiste en informar cómo se distribuyen los alumnos en distintas categorías o niveles de desempeño.

Esto puede ser realizado de diferentes maneras.

En la Ficha 7 se incluyó una descripción de niveles de desempeño en una competencia, "Comprensión del significado en la lectura", en niños de preescolar, 1° y 2° grado de Primaria en Uruguay. En la Figura 1 se reitera la descripción de dichos niveles, incorporando ahora la información relativa a qué proporción de los niños de cada grado quedó ubicado en cada nivel de desempeño.

Se trata de una distribución de frecuencias relativas (expresada en porcentajes, no en cantidades de niños). Las columnas suman 100 en sentido vertical, es decir, se establece la distribución de frecuencias para cada grado (5 años, 1° y 2°) por separado.

A partir de los datos se puede constatar que en el nivel 1, que corresponde a aquellos niños que solo logran una interpretación del texto a partir de las imágenes, pero no son capaces de identificar palabras o enunciados, se encuentra el 82% de los niños de 5 años, el 32% de los niños de 1er. grado y el 7% de los niños de 2°. grado. En el otro extremo, en los niveles 4 y 5, considerados en forma conjunta, se encuentra el 4% de los niños de 5 años, el 27% de los de 1er. grado y el 63% de los de 2°.

Es importante comprender que estos porcentajes se refieren a alumnos y son completamente diferentes de los porcentajes de respuestas correctas analizados en la Ficha 8.

En la Ficha 7 también se mostró como ejemplo los "niveles de desempeño" en Lectura de los jóvenes de 15 años definidos en la evaluación internacional PISA 2000. Estos niveles constituyen uno de los aportes más significativos de PISA, porque describen toda la gama de capacidades de lectura de los jóvenes de 15 años en unos 40 países.

Figura 1

Distribución de los alumnos por niveles de desempeño en "Comprensión del significado en la lectura" al inicio de la escolaridad, según grado – Uruguay, 2001

Nivel	Desempeño	5 años	1er	2º
1	**Interpreta solamente por la imagen.** El alumno interpreta la idea global o parte del texto solamente a partir de los elementos icónicos que aparecen en el mismo, pero no logra interpretar la parte verbal.	82%	32%	7%
2	**Interpreta algunas palabras y enunciados.** Puede interpretar algunas palabras y/o enunciados, aunque no pueda fundamentar su opinión o, si lo hace, es en forma inadecuada. El niño no reconoce el tema global del texto o, cuando se le hacen preguntas sobre el mismo, no ofrece respuestas adecuadas.	10%	14%	5%
3	**Interpreta algunos enunciados o párrafos.** Es capaz de interpretar algunos enunciados o párrafos en forma coherente, aunque aún no logre captar el tema central del texto. A veces reconoce, señalándola, alguna información explícita que se le solicita puntualmente.	5%	27%	25%
4	**Interpreta en forma global.** Interpreta el texto en forma global a partir de indicios verbales que aparecen en el mismo. Responde las preguntas que se le hacen acerca de la lectura, en forma coherente. Reconoce la mayoría de la información explícita y/o es capaz de ampliarla a partir de otros datos que tiene acerca del texto.	4%	26%	52%
5	**Realiza una buena síntesis del texto.** Realiza una buena síntesis del texto pudiendo relacionar elementos explícitos que aparecen en distintas partes del mismo. Por otra parte es capaz de vincular estos datos con información que no está explicitada, infiriéndola a partir de los datos solicitados o de los conocimientos previos que posee (aunque lo haga en forma parcialmente adecuada).	0%	1%	11%
		100%	100%	100%

Fuente: Administración Nacional de Educación Pública/ Gerencia de Investigación y Evaluación (2002); Los Niveles de Desempeño al Inicio de la Educación Primaria. Estudio de las competencias lingüísticas y matemáticas. Montevideo, Uruguay.

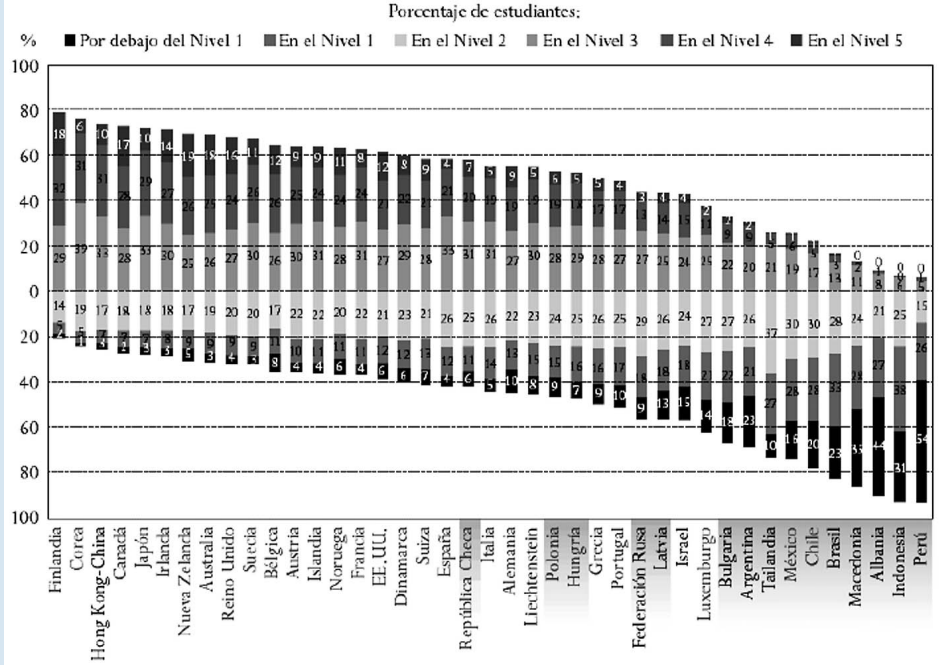
FICHA 9

Los informes de resultados de PISA incluyen, además de la información sobre promedios por país presentada en la Ficha 8, la distribución de frecuencias de los alumnos de cada país entre los distintos niveles de desempeño (véase la Figura 2).

En el Gráfico de la Figura 2 es posible apreciar qué porcentaje de alumnos tuvo cada país en cada uno de los cinco niveles de desempeño (y por debajo del nivel 1).

Los países están ordenados de acuerdo al porcentaje de alumnos que alcanzaron al menos el nivel 3. Por eso, el gráfico está estructurado en torno al valor 0 en el eje 'y'. Hacia arriba

Figura 2 **Porcentajes de alumnos por niveles de desempeño en Lectura, PISA 2000**



Fuente: base de datos del Proyecto PISA de la OCDE 2003, Cuadro 2.1a.

se indican los porcentajes de alumnos en los niveles 3, 4 y 5 y hacia abajo los porcentajes de alumnos en los niveles 2, 1 y por debajo del 1.

Obsérvese que, si bien son resultados de PISA 2000 en Lectura al igual que los presentados en la Ficha 8, el ordenamiento de países es diferente al obtenido mediante promedios. Finlandia sigue en el primer lugar, pero ahora le siguen Corea del Sur y Hong Kong. Nueva Zelanda y Australia han quedado algo más rezagados. ¿A qué se debe esto? A que si bien estos dos últimos países tienen proporciones altas de alumnos con altos niveles de desempeño en el nivel 5 (19% y 18% respectivamente), también tienen proporciones importantes de alumnos bajo el nivel 3. Corea del Sur, y Hong Kong, en cambio, tienen proporciones muy reducidas de alumnos de bajo desempeño.

Finlandia tiene la situación ideal y por eso destaca claramente en el primer lugar: porcentajes importantes de alumnos en los niveles altos y muy pocos alumnos en los niveles bajos. En el otro extremo del gráfico se puede apreciar que Brasil, Ma-cedonia, Albania, Indonesia y Perú tienen más del 80% de sus alumnos por debajo del nivel 3.

Los datos de la Figura 2, considerados en conjunto con la descripción de qué son capaces de hacer los alumnos en cada nivel, aportan más información que la mera indicación de promedios. Mientras que los promedios solo permiten comparar posiciones relativas, la Figura 2 permite saber qué proporción de los alumnos de cada país está en cada nivel de desempeño.

Un tercer ejemplo de reporte de resultados a través de distribución de frecuencias relativas de los alumnos en niveles de desempeño se consigna en la Figura 3. En este caso, se trata de un reporte de Argentina del año 1999, en el que se construyen niveles de desempeño en una prueba que responde al modelo de la TCT (la de PISA corresponde a la TRI).

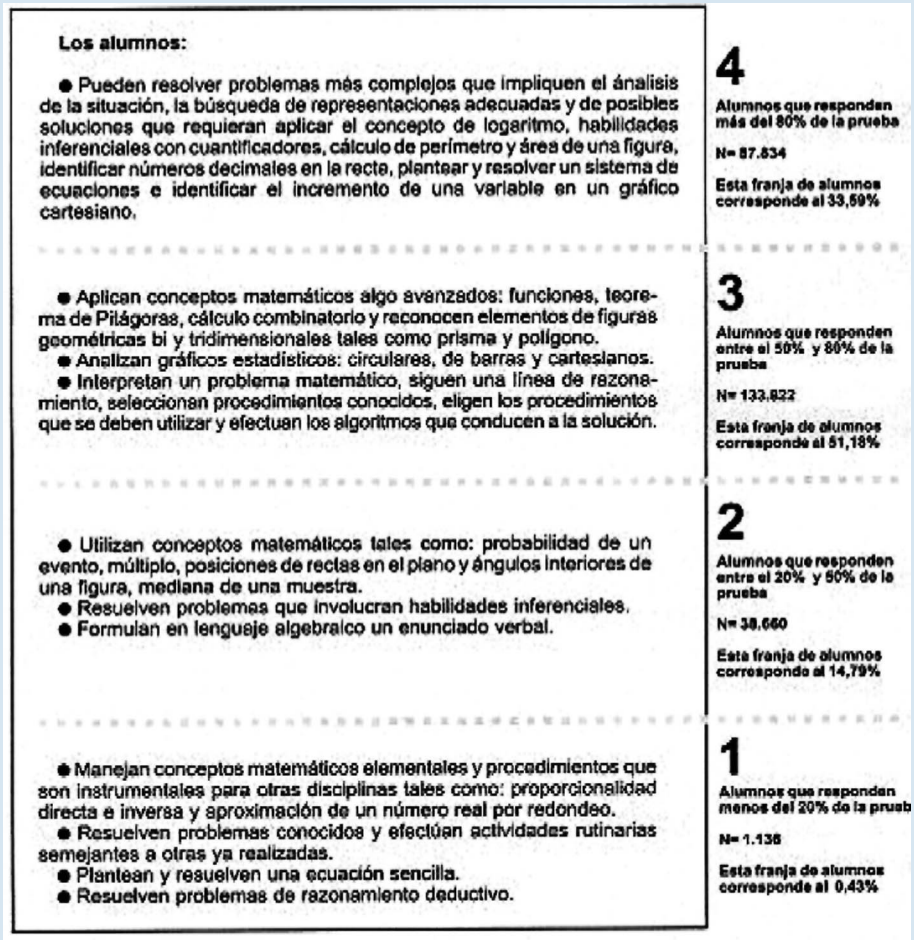
Se definen cuatro niveles de desempeño y se describe qué tipo de actividades son capaces de realizar los alumnos en cada nivel.

A la derecha de la tabla se incluyen dos informaciones relevantes. En primer lugar, se indica cuál fue el criterio para establecer los niveles. Por ejemplo, el nivel 4 se define como aquel nivel que caracteriza a los alumnos que respondieron correctamente más del 80% de las preguntas de la prueba.

FICHA 9

Figura 3

Niveles de desempeño en Matemática al final de la Educación Media – Argentina, 1999



Fuente: Ministerio de Cultura y Educación de la Nación, República Argentina. Dirección Nacional de Evaluación, 2000. III Operativo Nacional de Evaluación de Finalización del Nivel Secundario 1999.

En segundo lugar, se indica qué porcentaje del total de alumnos quedó ubicado en cada nivel. Por ejemplo, en el nivel 4 quedó ubicado el 33,59% de los alumnos. Se incluye, además, el total en números absolutos (N).

Recuadro 1

¿Cómo se definen los niveles de desempeño?

La construcción de niveles de desempeño es un modo de hacer más comprensible y relevante la información aportada por una prueba. Una prueba estandarizada entrega siempre un puntaje en una escala continua (tanto en TCT como en TRI). Transformar esa escala continua en tres, cuatro o cinco grandes categorías, que puedan ser descritas en términos de lo que los alumnos en cada una de ellas son capaces de hacer, permite que la información resultante sea más significativa y relevante.

La definición de los niveles es realizada por especialistas y/o docentes (idealmente, por una combinación de ambos). Puede ser hecha antes de la aplicación de la prueba –simplemente a partir del análisis de lo que implica resolver cada actividad de la prueba–, o después de una aplicación de la misma, –teniendo en cuenta los datos reales del desempeño de los alumnos. Normalmente se hace ambas cosas, es decir, se realiza una primera clasificación provisoria de los ítemes en niveles a partir del juicio experto de especialistas y/o docentes, que luego es revisada a partir de los resultados de una aplicación de la prueba.

Hay dos modos principales de realizar la tarea, uno que prioriza el análisis de los ítemes y otro que toma como punto de partida la distribución de los alumnos.

PISA es un ejemplo del primer caso. La definición de niveles de desempeño en PISA se realiza a partir de un mapa de ítemes como el presentado en la Figura 1 de la Ficha 8. Los especialistas analizan los ítemes y buscan en qué puntajes de la escala establecer los "puntos de corte", es decir, las fronteras entre un nivel y el siguiente. Esto se efectúa de modo que el conjunto de ítemes que quedan al interior de un nivel refleje un tipo de desempeño que tenga significado conceptual en el dominio evaluado, teniendo en cuenta las competencias que requiere resolverlos. Al mismo tiempo, se busca que todos los alumnos que quedan clasificados dentro de un nivel –según el puntaje obtenido en la prueba– tengan al menos 50% de probabilidad de responder correctamente a todos los ítemes que lo conforman. Los alumnos que están en la frontera superior del nivel tendrán una probabilidad mayor.

El caso argentino (Figura 3) y el caso del SAEB brasileño –que se presenta un poco más adelante en esta Ficha– siguen un procedimiento diferente. Primero se establecen los "puntos de corte" a partir de los puntajes de los alumnos en la prueba (por ejemplo, en el caso argentino, se define el nivel 3 como el correspondiente a alumnos que resolvieron correctamente entre el 50% y el 80% de la prueba). Luego se procede a analizar qué ítemes resolvieron correctamente la gran mayoría de los alumnos dentro de cada nivel.

En cualquiera de las dos aproximaciones el proceso es iterativo, es decir, requiere de varias revisiones mediante las cuales se busca, por un lado, que los niveles tengan sentido desde el punto de vista conceptual y, simultáneamente, que los alumnos queden clasificados adecuadamente.

Los niveles de desempeño siempre son inclusivos. Esto significa que los alumnos de los niveles superiores pueden responder no solo a las actividades correspondientes a los mismos sino que, a la vez, tienen mayor probabilidad de responder a las actividades de los niveles inferiores. En cambio, los alumnos de los niveles inferiores tienen baja probabilidad de responder correctamente preguntas que corresponden a niveles superiores –aunque ello puede ocurrir en algunos casos–.

Las fronteras entre niveles definidas por los puntos de corte siempre tienen un cierto grado de arbitrariedad. Por ejemplo, en el caso argentino seguramente hay menos distancia en términos de las competencias y conocimientos evaluados entre un alumno con 81% de respuestas correctas y uno con 79% que entre este último y un tercero con 52% de respuestas correctas. Sin embargo, el primero queda clasificado en el Nivel 4 y los otros dos en el Nivel 3. Lo mismo ocurre en PISA.

Ésta es una debilidad inevitable que no es grave cuando las pruebas no tienen consecuencias para los alumnos. Es el costo de construir una presentación más significativa de los datos. No obstante, puede constituirse en un problema serio cuando las pruebas tienen consecuencias para los alumnos, es decir, cuando del resultado de la prueba depende, por ejemplo, que el alumno apruebe o repruebe un curso. En estos casos los cuidados para establecer los "puntos de corte" son mayores y hay métodos específicos que se presentan en el recuadro 2.

2. Incorporando un criterio o expectativa de lo esperable a los niveles de desempeño

Los ejemplos de niveles de desempeño mostrados en el apartado anterior tienen un carácter únicamente descriptivo. No pretenden establecer cuál es el nivel al que todos los alumnos deberían llegar.

Por ejemplo, PISA no propone ninguno de los cinco niveles de Lectura como el nivel exigible a todos los alumnos. No tendría sentido hacer esto en una evaluación internacional, en que participan países muy diferentes. En todo caso, establecer una meta de ese tipo es una determinación que cada país puede realizar por sí mismo, de acuerdo a su realidad.

Obviamente lo deseable es que todos los alumnos alcancen los niveles de desempeño más altos y esa meta de largo plazo no debe ser nunca abandonada. Pero como meta específica de política educativa para un plazo determinado, pretender que todos los alumnos alcancen el nivel 5 no parece razonable.

Los niveles de desempeño al inicio de la escolaridad ilustrados en la Figura 1 tampoco establecen cuál debería ser el nivel al que deberían llegar los alumnos de cada grado. Simplemente describen la realidad, pero no establecen una meta ni definen una expectativa para cada edad. Lo mismo ocurre con la clasificación de niveles establecida en la Figura 3.

Ahora bien, mientras que en las evaluaciones internacionales carecería de sentido establecer una expectativa acerca del nivel al que todos los estudiantes debieran llegar, ello no es así a nivel nacional. Por el contrario, en las evaluaciones nacionales es deseable definir cuáles son los desempeños que deberían lograr todos los alumnos que finalizan un determinado ciclo educativo (el punto fue discutido en la Ficha 5).

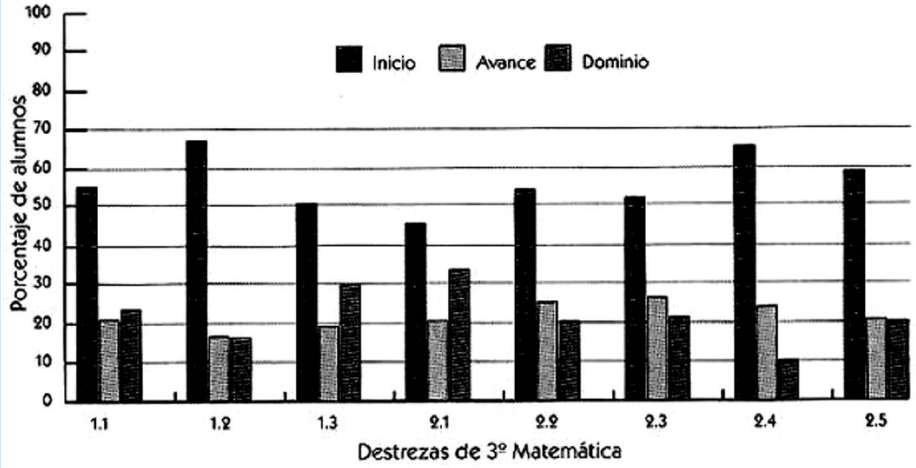
A continuación se presentan tres ejemplos de cómo esto ha sido realizado en diferentes evaluaciones nacionales en la región.

Ecuador y Costa Rica han reportado los porcentajes de alumnos que dominan cada uno de un conjunto seleccionado de objetivos curriculares –los más relevantes–. El supuesto es que todos los alumnos deberían dominar la totalidad de los objetivos curriculares seleccionados. En este sentido es que hay una meta, estándar o expectativa claramente definida.

FICHA 9

Figura 4 **Porcentajes de alumnos por niveles de logro de las destrezas de Matemática – 3º. Aprendo 1997 – Ecuador**

DESTREZAS	Inicio	Avance	Dominio
1.1 Establecer la relación de orden entre números.	55,33	21,00	23,67
1.2 Identificar la regla de formación de una sucesión.	67,05	16,56	16,39
1.3 Completar una sucesión.	50,90	19,19	29,91
2.1 Resolver adiciones y sustracciones que no requieren la destreza de llevar.	45,72	20,60	33,68
2.2 Resolver adiciones y sustracciones que requieren la destreza de llevar.	54,34	25,32	20,33
2.3 Hallar la solución de problemas que requieren una adición o una sustracción.	52,26	26,44	21,30
2.4 Hallar la solución de problemas que requieren la combinación de adiciones y sustracciones.	65,56	24,19	10,24
2.5 Estimar el resultado de problemas que requieren sumas y restas, y descubrir una relación entre números.	58,97	20,73	20,30



Fuente: Ministerio de Educación y Cultura, EB/PRODEC, Ecuador, 1998. Segunda Prueba Nacional "APRENDO 1997". Resultados Nacionales; pág. 27.

La Figura 4 muestra este modo de reportar para el caso de Ecuador. En este caso se emplearon cuatro actividades diferentes para cada "destreza" y se establecieron las siguientes categorías en relación a cada una de las destrezas:

a. *"dominio"*, constituida por los alumnos que respondieron correctamente al menos 3 de las 4 actividades;
b. *"avance"*, que corresponde al hecho de responder correctamente 2 de las 4 actividades;
c. *"inicio"*, constituida por los alumnos que respondieron correctamente una o ninguna de las 4 actividades.

Nuevamente, tenemos un caso de reporte mediante distribución de frecuencias: la información que se presenta es qué porcentaje de los alumnos se ubica en cada una de las categorías anteriores en cada destreza evaluada.

Sin embargo, estos países no definieron una expectativa de desempeño para el conjunto de la prueba. Reportan distribuciones de frecuencias por separado para cada objetivo curricular, pero a la hora de dar un resultado global de la prueba recurren al porcentaje promedio de respuestas correctas, sin establecer un parámetro de qué sería un resultado aceptable en el conjunto de la prueba.

Otra debilidad del enfoque radica en definir como dominio de un objetivo curricular resolver correctamente 3 preguntas de 4, dado que, como son muy pocos ítemes, el resultado puede verse afectado por la dificultad de algunos de ellos (ciertos objetivos pueden aparecer como más logrados que otros simplemente porque se utilizaron uno o dos ítemes más fáciles).

Además, dicha definición de dominio no aporta información alguna sobre qué son capaces de hacer los alumnos que dominan un objetivo ni qué diferencia a un alumno en estado de "avance" de los demás (salvo que responde correctamente dos preguntas).

Uruguay constituye un ejemplo diferente, que también se caracteriza por establecer una expectativa en cuanto al nivel de desempeño esperable para todos los alumnos que terminan la Educación Primaria.

Las evaluaciones nacionales se realizan en 6° grado de primaria cada tres años en Uruguay.

Para cada área evaluada se establecen tres grandes áreas de competencias y un conjunto de contenidos que fueron previamente discutidos y definidos como fundamentales para egresar del nivel. En el caso de Matemática las tres grandes competencias evaluadas son "comprensión de conceptos", "aplicación de algoritmos" y "resolución de problemas". Para el caso de Lenguaje son "comprensión de textos argumentativos", "comprensión de textos narrativos" y "reflexiones sobre el lenguaje". Competencias y contenidos aparecen explicitados en un documento marco de la evaluación.

Las pruebas están conformadas por 24 preguntas. Con anterioridad a la aplicación de las pruebas se estableció cuál sería el "punto de corte" que definiría el desempeño deseable para todos los alumnos, el cual fue fijado en 14 puntos, que corresponden a los 14 ítemes más fáciles . A este nivel se le denomina "suficiencia" y se considera como "suficientes" –o con un desempeño satisfactorio– a aquellos alumnos que alcanzan o superan dicho puntaje.

Figura 5

Resultados en Lengua y Matemática (en porcentajes de alumnos suficientes) Uruguay 1996-1999 (*)

	LENGUA		MATEMÁTICA	
Porcentaje de alumnos suficientes	1996	1999	1996	1999
	57,1	61,3	34,6	40,8
Diferencia de Resultados entre 1999 y 1996	**+ 4,2**		**+ 6,2**	
Margen de error muestral 1999	+/- 3,0		+/- 3,4	
Intervalo de Confianza de los resultados 1999	58,3 a 64,3		37,4 a 44,2	

(*) Para establecer las comparaciones entre años se trabaja con formas equivalentes de prueba, dentro de la TCT, que tienen la misma extensión, la misma estructura de competencias y contenidos, los mismos pesos internos y la misma dificultad promedio.

Fuente: ANEP/Unidad de Medición de Resultados Educativos, 1999; Evaluación Nacional de Aprendizajes en Lenguaje y Matemática. 6to. año Educación Primaria 1999. Primer Informe de Resultados.

Escala de proficiencia en Matemática - BRASIL/SAEB 1997

Figura 6

Ejemplos de desempeño	NIVEL	Resultados del SAEB 97
• *Los alumnos reconocen el valor de billetes y monedas.* • *Leen la hora en relojes digitales y analógicos y saben que una hora tiene 60 minutos.* • *Resuelven problemas sencillos de adición y sustracción con números naturales.*	**175**	*En el nivel 175 o arriba de él se encuentran:* • *56% de los alumnos del 4° grado de la Enseñanza Fundamental;* • *95% de los alumnos del 8° grado de la Enseñanza Fundamental;* • *100% de los alumnos del 3er. grado de la Enseñanza Media.*
• *Los alumnos reconocen polígonos y cuadriláteros.* • *Establecen relaciones entre los valores de cédulas y monedas y resuelven situaciones al pagar y recibir cambio, aunque todavía no saben operar con decimales.* • *Son capaces de multiplicar y dividir, así como identificar unidades, decenas y centenas.* • *Resuelven problemas que envuelven más de una operación.* • *Adicionan y substraen fracciones de un mismo denominador y conocen números naturales en la forma fraccionaria.* • *Interpretan gráficos de barras y de sector e identifican el gráfico más adecuado para representar una determinada situación.*	**250**	*En el nivel 250 o arriba de él se encuentran:* • *1% de los alumnos del 4° grado de la Enseñanza Fundamental;* • *48% de los alumnos del 8° grado de la Enseñanza Fundamental;* •*87% de los alumnos del 3er. grado de la Enseñanza Media.*
• *Los alumnos clasifican sólidos geométricos en cuerpos redondos y poliedros.* • *Interpretan resultados de medidas de longitud, masa, tiempo y capacidad.* • *Identifican, comparan y ordenan números racionales (en las formas fraccionaria y decimal) así como números enteros.* • *Interpretan lenguaje algebraico y resuelven ecuaciones y sistemas de ecuaciones de primer grado.*	**325**	**En el nivel 325 o arriba de él se encuentran:** • *8% de los alumnos del 8° grado de la Enseñanza Fundamental;* • *32% de los alumnos del 3er. grado de la Enseñanza Media.*
• *Los alumnos resuelven problemas que envuelven punto, recta, circunferencia y sus relaciones.* • *Establecen relaciones y hacen conversiones entre fracciones ordinarias y números decimales.* • *Resuelven problemas que envuelven ecuaciones e inecuaciones sencillas de primer y segundo grado y sistemas de primer grado.* • *Conocen los principios básicos de polinomios y efectúan operaciones elementales entre ellos.* • *Conocen las propiedades básicas de exponentes y logaritmos.*	**400**	**En el nivel 400 o arriba de él se encuentran:** • *5% de los alumnos del 3er. grado de la Enseñanza Media.*

Observación: Los niveles 325 y 400 muestran el dominio de habilidades y contenidos más complejos que no corresponden al 4° grado de la Enseñanza Fundamental. Es por ello que no se presentan resultados para los alumnos de este grado.

Fuente: Ministerio de Educación - Gobierno Federal. Instituto Nacional de Estudios e Investigaciones Educacionales (INEP), Brasil, 1998; ¿Cómo está la Educación Básica en Brasil (Traducción del autor).

FICHA 9

Figura 7

La definición de criterios de logro o estándares para cada ciclo de enseñanza a partir de los niveles de desempeño – SAEB / Brasil, 1997

Nivel de proficiencia - escala SAEB/97	Matemática	Lengua portuguesa	Ciencias (Física, Química y Biología)
	Ciclo y nivel de enseñanza		
	No significativo	Hacia **la mitad** del 1er. ciclo de la Enseñanza Fundamental	Hacia **la mitad** del 1er. ciclo de la Enseñanza Fundamental
	Hacia **la mitad** del 1er. ciclo de la Enseñanza Fundamental	Hacia **la final** del 1er. ciclo de la Enseñanza Fundamental	Hacia **el final** del 1er. ciclo de la Enseñanza Fundamental
	Hacia **el final** del 1er. ciclo de la Enseñanza Fundamental	Hacia **el final** del 2º ciclo de la Enseñanza Fundamental	Hacia **la mitad** del 2º ciclo de la Enseñanza Fundamental
	Hacia **el final** del 2º ciclo de la Enseñanza Fundamental	Hacia **el final** de la Enseñanza Media	Hacia **el final** del 2º ciclo de la Enseñanza Fundamental
	Hacia **el final** de la Enseñanza Media	Después **del final** de la Enseñanza Media	Hacia **el final** de la Enseñanza Media

Fuente: Ministerio de Educación - Gobierno Federal. Instituto Nacional de Estudios e Investigaciones Educacionales (INEP). SAEB 97 - Primeros Resultados

Para llegar a esta definición trabajaron docentes y especialistas que analizaron los ítemes teniendo en mente qué preguntas deberían ser capaces de responder todos los alumnos al terminar la escuela primaria (véase el recuadro 2). Simultáneamente, se mantuvieron preguntas más difíciles, que permitiesen evaluar a los alumnos de mejor desempeño. El proceso fue iterativo, dado que para que el punto de corte fuese 14 puntos para ambas disciplinas, a veces fue necesario quitar o incluir determinados ítemes.

El dato principal de los reportes de Uruguay es qué porcentaje de los alumnos logró alcanzar el nivel de "suficiencia" (véase la Figura 5).

Lo interesante de este caso es que se pone el énfasis en definir un "punto de corte" que constituye una referencia acerca de lo que todos los alumnos deberían conocer y ser capaces de hacer al finalizar la escuela primaria.

Al mismo tiempo, una limitación de los informes de Uruguay, derivada del trabajo con un número limitado de ítemes, es que no se explicita con suficiente detalle y claridad qué significado tiene ese "punto de corte", es decir, qué son capaces de hacer los alumnos que alcanzan el nivel de suficiencia.

El lector debería apreciar dos diferencias del caso de Uruguay respecto al de Argentina. En primer lugar, en la elaboración de las pruebas no se eliminan las actividades muy fáciles o muy difíciles, sino que se busca trabajar con toda la gama de dificultades. En segundo lugar, nótese la diferencia entre reportar porcentaje de alumnos que alcanzan un determinado puntaje en la prueba (Uruguay) y porcentaje promedio de respuestas correctas (Argentina) (ver Figura 3).

Los casos de Ecuador y Uruguay constituyen esfuerzos importantes por establecer expectativas definidas respecto a los aprendizajes de los alumnos. Sin embargo, el SAEB brasileño es tal vez uno de los modos de evaluar y reportar mejor logrados en la región, en la medida que combina el uso de la TRI con la definición de niveles de desempeño y con el establecimiento de "estándares" o expectativas.

Un primer análisis de cómo Brasil atribuye juicios valorativos a los niveles de desempeño resultantes de la TRI fue presentado en la Figura 7 de la Ficha 3. Complementariamente, en la Figura 6 se incluye una descripción más amplia de los niveles de desempeño en Matemática según fueron descritos en 1997.

Según fue explicado antes en esta Ficha, el procedimiento seguido consiste en establecer tramos de puntaje arbitrarios en la escala y describir qué son capaces de hacer los alumnos en cada tramo. Estas descripciones aparecen en la primera columna de la Figura 6. En la columna central se indican los tramos de la escala, al tiempo que en la columna de la derecha se informa qué proporción de los alumnos de distintos grados quedó ubicado en cada nivel de desempeño.

Un primer aspecto original del enfoque del SAEB es que emplea una misma escala de puntajes y una misma descripción de desempeños para alumnos de diferentes grados del sistema educativo.

Recuadro 2

¿Cómo se definen los "puntos de corte" o expectativas de desempeño?

Definir cuál es el nivel de desempeño aceptable en una prueba o cuál es el puntaje mínimo que un estudiante debería lograr para considerar que ha aprendido lo que se espera, es una tarea compleja cuyo resultado siempre puede ser objeto de debate, tanto en pruebas estandarizadas como en pruebas elaboradas y aplicadas por docentes en las aulas. ¿Qué tan exigente se debe ser? ¿Qué tanto es razonable esperar de los alumnos? ¿Hay una diferencia real entre quien está apenas por debajo del "punto de corte" y quien está apenas por encima?.

Sin embargo, es necesario acometer esta tarea no solo en educación, sino en múltiples áreas de la actividad humana. Por ejemplo, es necesario definir una línea de pobreza, un límite entre el nivel de colesterol en la sangre aceptable y no aceptable, o cuál es el máximo de emisiones de gas tolerables en una planta industrial. Estos tres ejemplos de "puntos de corte" presentan los mismos problemas y dificultades antes enunciadas. Inevitablemente hay un grado de arbitrariedad en su definición, que exige que dicha definición se apoye en la mayor información posible y en la opinión de expertos.

Hay diversidad de procedimientos para establecer "puntos de corte"[1] en pruebas. Todos ellos se apoyan, necesariamente, en la opinión experta de "jueces" (especialistas en las disciplinas y docentes experientes). Las herramientas estadísticas pueden servir de apoyo pero no resolver el problema, que es esencialmente valorativo. De allí que la selección y el entrenamiento de los "jueces" sea de vital importancia.

Uno de los procedimientos más conocidos y utilizados (con algunas variantes) es el de **Angoff.** Consiste en pedir a un grupo amplio de jueces que, teniendo en mente a los alumnos "aceptables" (poseedores de los conocimientos y competencias básicas requeridas), establezcan cuál sería la probabilidad de que dichos alumnos respondan correctamente a cada uno de los ítemes que conforman la prueba. Luego se calcula el promedio de probabilidades teniendo en cuenta todos los ítemes y todos los jueces. Este promedio es un porcentaje que, aplicado al puntaje total de la prueba, determina el punto de corte que distingue a alumnos aceptables de no aceptables. Por ejemplo, si el promedio de las probabilidades de responder correctamente a los ítemes establecidas por los jueces es 63% y la prueba tiene en total 32 preguntas, el punto de corte será 32 x 0,63 = 20 puntos. El procedimiento puede ser complejizado realizando dos "rondas" de trabajo, la primera en base al análisis de los ítemes y una segunda ya teniendo en cuenta resultados de la aplicación de la prueba.

Otro procedimiento, algo diferente, es el de Zeiky y Livingston. En este caso se pide a docentes competentes que clasifiquen a sus estudiantes en tres categorías: los que son competentes en la materia, los que están alrededor del límite de lo aceptable por encima y por debajo, y los que no son competentes. Luego se aplica la prueba completa a los alumnos de la categoría intermedia. La mediana obtenida de esta aplicación –es decir, el puntaje que divide a este grupo de alumnos en dos–, se utiliza como "punto de corte" para la prueba.

Cuando se ha definido previamente niveles de desempeño, el establecimiento de un "punto de corte" es más sencillo. En estos casos se trata de seleccionar cuál de los niveles es el exigible a todos los alumnos. La determinación del "punto de corte" está implícita en esta elección.

Este tipo de determinaciones nunca está exenta de debates y puede ser mejorada a partir de su discusión pública. Por ejemplo, distintos actores sociales y académicos pueden tener diferentes visiones acerca de qué deben aprender los alumnos de educación media en Matemática y de cuáles son los niveles satisfactorios a los que todos deberían llegar. Unos podrían enfatizar la Matemática conceptual y otros defender la resolución de problemas como aspecto central. Unos serán partidarios de altos niveles de exigencia, en tanto otros argumentarán que tales niveles solo son exigibles a una minoría de alumnos que cursarán carreras científicas, pero que no son adecuados para la mayoría.

De allí la importancia de establecer espacios de discusión y consulta amplios y de "abrir" las definiciones tomadas al escrutinio público.

Por información más detallada sobre estos y otros procedimientos véase Tuijnman, A. & Postlehwaite, T. (ed.), 1995; Monitoring the Standards of Education; caps. 9 y 10. Pergamon.

De todos modos, lo presentado en la Figura 6 no incluye aún una definición de cuál es el nivel de desempeño que se espera alcancen todos los alumnos que finalizan cada ciclo escolar.

Esto se hace en la Figura 7. En ella, a partir de la opinión de "jueces", se establece cuál es el nivel que deberían haber alcanzado los alumnos cuando finalizan cada uno de los principales ciclos del sistema educativo. El término "jueces" se emplea para denominar a aquellos expertos que establecen un juicio de valor respecto al nivel que los alumnos deberían alcanzar al final de cada ciclo.

Si el lector contrasta las expectativas definidas en la Figura 7 con los datos consignados en la tercera columna de la Figura 6, puede constatar que la evaluación revela serios problemas en la educación brasileña.

Por ejemplo, en la Figura 7 se establece que en Matemática el Nivel 250 debería ser alcanzado hacia el final del 1er. Ciclo de la Enseñanza Fundamental (4º grado). En la tercera columna de la Figura 6 se reporta que apenas el 11% de los alumnos de 4º grado de la Enseñanza Fundamental estaban en el nivel 250 o por encima de él.

Del mismo modo, en la Figura 7 se establece que los alumnos deberían alcanzar el nivel 400 en Matemática hacia el final de la Enseñanza Media. Sin embargo, de acuerdo a los datos reportados en la Figura 6, apenas el 5% de los alumnos de 3º de Enseñanza Media se ubican en dicho nivel.

Síntesis final

Esta Ficha, junto con la anterior, intentan orientar al lector para que esté en mejores condiciones de comprender los datos numéricos que aparecen en los reportes de las evaluaciones estandarizadas. Para ello debe tener presentes tres conceptos clave:

En primer término, observar si se trata de una escala de **Teoría Clásica** (TCT) o de **Teoría de Respuesta al Ítem** (TRI). En el primer caso, encontrará que la escala refleja la cantidad de preguntas contestadas correctamente. En el segundo se encontrará frente a una escala sin máximo ni mínimo y generalmente centrada en una media de 500 o de 250 puntos.

En segundo término, observar si está ante un **promedio** o ante una **distribución de frecuencias.** En el primer caso, puede tratarse de un promedio de puntajes de cualquiera de los dos modelos anteriores (TCT o TRI). También puede tratarse de un porcentaje promedio de respuestas correctas (en TCT). Los promedios normalmente se emplean para análisis de tipo normativo, es decir, centrados en la comparación entre entidades. Si está ante una distribución de frecuencias encontrará los porcentajes de alumnos en ciertas categorías o niveles de desempeño. Normalmente se emplean para un análisis de tipo criterial, es decir, centrado en la descripción de lo que los alumnos conocen y son capaces de hacer.

En tercer término, observar si el modo en que se reportan los resultados **incluye la definición de un estándar acerca de lo esperable o exigible a todos los alumnos, o simplemente describe diferentes categorías o niveles de desempeño.**

Los ejemplos analizados a lo largo de las Fichas 8 y 9 encajan en la conceptualización anterior de la siguiente manera:

PISA trabaja con TRI. Reporta tanto a través de promedios –que permiten ordenar a los países– como a través de distribución de frecuencias de los estudiantes en niveles de desempeño –que permiten principalmente analizar qué son capaces de hacer los alumnos, aunque también sirven para comparar entre países–. No establece un nivel exigible a todos los estudiantes.

Argentina trabaja con TCT. Reporta principalmente el porcentaje promedio de respuestas correctas que sirve para establecer comparaciones entre provincias. Construye una descripción de niveles, pero como las pruebas fueron elaboradas con un enfoque normativo (eliminando ítemes muy fáciles y muy difíciles) la descripción de desempeños es incompleta. No establece un nivel exigible a todos los estudiantes.

Ecuador trabaja con TCT. Describe los desempeños en términos de destrezas curriculares simples. Establece un nivel exigible para cada destreza por separado y reporta qué porcentaje de los alumnos alcanza el dominio de cada destreza.

Uruguay trabaja con TCT. Focaliza el reporte de resultados en el porcentaje de alumnos que alcanza un puntaje definido como "suficiencia" en la prueba, pero no hace una descripción detallada de lo que los alumnos son capaces de hacer.

Brasil trabaja con TRI. Reporta tanto promedios como porcentajes de alumnos por niveles de desempeño. Establece, además, un estándar o expectativa al definir cuál es el nivel de desempeño que deberían alcanzar los alumnos en distintos momentos de la escolaridad.

¿POR QUÉ LOS RANKINGS SON MODOS INAPROPIADOS DE VALORAR LA CALIDAD DE LAS ESCUELAS?

Hacia nuevos modos de utilizar los resultados por escuela

La Ficha 10 tiene como propósito alertar al lector acerca de la sobre-simplificación de la realidad implícita en la mayoría de los *rankings* educativos, que en general ofrecen una visión errada e incompleta de las instituciones y sistemas educativos, conduciendo a juicios inapropiados acerca de la calidad de los mismos –y muchas veces contribuyen más a deteriorar que a mejorar la calidad de la educación–.

Con este fin, se explica y ejemplifica cuatro principales tipos de problemas que presenta la mayoría de los *rankings*:

1. pretenden ordenar en función de la "calidad", cuando en realidad utilizan un único indicador (o, a lo sumo, unos pocos);
2. ofrecen una falsa impresión de precisión en el ordenamiento;
3. no tienen en cuenta que, a nivel escolar, los resultados son altamente volátiles;
4. no consideran las diferencias en la composición social del alumnado de las escuelas y en los recursos con que estas cuentan.

En función del análisis de estas debilidades, en la Ficha se propone una serie de recomendaciones sobre modos apropiados de realizar ordenamientos y comparaciones de resultados entre escuelas o jurisdicciones.

1. Una visión simplista de la realidad educativa

Para comenzar, conviene distinguir entre dos cosas bastante diferentes: una es generar información sobre qué tanto las escuelas están logrando que sus alumnos adquieran los aprendizajes esperados y otra completamente diferente es convertir eso en una tabla de posiciones tipo campeonato de fútbol, de manera simplista y con gran daño para muchas escuelas (véase el recuadro 2 más adelante en esta Ficha y el recuadro 4 en la Ficha 11).

Las Figuras 1 y 2 muestran el tratamiento que el diario *La Segunda* de Chile, en su edición del jueves 21 de noviembre de 2002, dio a los resultados de la prueba SIMCE 2001. Seguramente no es este el tipo de tratamiento que el Ministerio de Educación de Chile busca al publicar los resultados por escuela, pero la realidad es que este es el tipo de tratamiento que buena parte de los medios de prensa da a los *rankings* y que el enfoque de los medios de prensa es lo que más impacto tiene en la opinión pública y en el sector educativo.

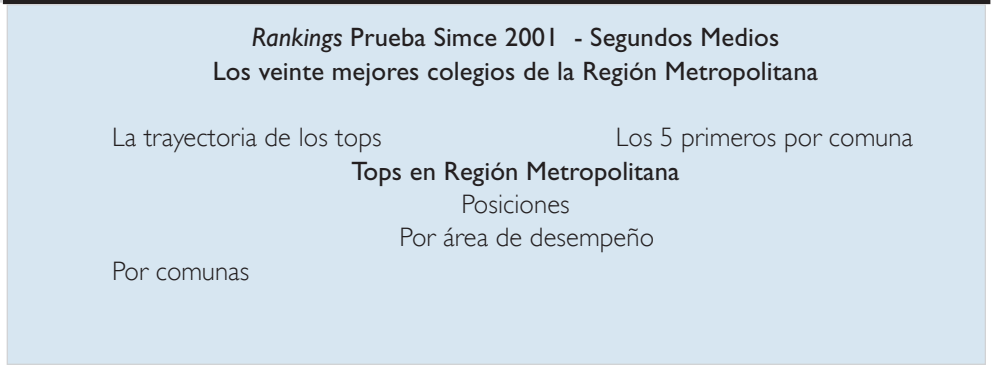
Lo primero que debe observarse son los titulares estilo campeonato de fútbol que "trivializan" la realidad educativa.

Lo segundo a observar es el hecho de que "los mejores" colegios pertenecen a los sectores privilegiados de la sociedad. 18 de 20 son colegios particulares pagados y 2 son colegios públicos tradicionales de los principales municipios de la Región Metropolitana. Parte importante de ellos son colegios bilingües que atienden preferentemente a alumnos de clase alta y media alta.

Cabe, por lo tanto, la pregunta de si se trata de los mejores colegios o de los más selectivos –tal vez sean una mezcla de ambas cosas, pero no necesariamente son los que más "agregan" a lo que los alumnos traen del hogar–.

El tercer aspecto que el lector debe notar es la inestabilidad de los resultados. El colegio que aparece en primer lugar en el 2001 ocupaba el lugar 39 en 1998. Mientras tanto el

Figura 1 **Los ranking de escuelas en los medios de prensa**



Fuente: Titulares del diario La Segunda de Chile, edición del 21 de noviembre de 2002.

Figura 2 **La trayectoria de los mejores**

Establecimiento	Comuna	Dependencia	*Ranking* 2001	*Ranking* 1998	*Ranking* 1994
Colegio Internacional Alba	Maipú	PP	1	39	-
Colegio del Sagrado Corazón Apoquindo	Las Condes	PP	2	11	47
Colegio Castelgandolfo	Padre Hurtado	PP	2	-	-
Andree English School	La Reina	PP	3	30	82
Liceo Carmela Carvajal de Prat	Providencia	M	4	10	23
Colegio Tabancura	Vitacura	PP	5	38	46
Instituto Nacional General J.M. Carrera	Santiago	M	6	9	18
Colegio Rubén Darío	La Reina	PP	7	-	-
Colegio Inglés The Grange School	La Reina	PP	8	16	20
The Southern Cross School	Las Condes	PP	8	53	98
Colegio Los Andes	Vitacura	PP	9	15	61
Colegio La Girouette	Las Condes	PP	10	6	11
Colegio Buin	Buin	PP	11	-	-
Lincoln International Academy	Lo Barnechea	PP	12	5	44
Saint Gabriel's School	Providencia	PP	12	24	53
Sociedad Colegio Alemán de Santiago	Las Condes	PP	12	-	-
Redland School	Las Condes	PP	13	49	89
Villa María Academy	Las Condes	PP	13	6	25
Colegio Cumbres	Las Condes	PP	14	20	-
Colegio Padre Hurtado y Juanita de Los Andes	Las Condes	PP	14	45	68

PP: Particular Pagado; PS: Particular Subvencionado; M: Municipal.

Fuente: Diario La Segunda de Santiago, edición del 21 de noviembre de 2002.

colegio "The Southern Cross School", de uno de los barrios más acomodados de Santiago (Las Condes), ocupó el 8º lugar en 2001, el lugar 53 en 1998 y el lugar 98 en 1994. El "Villa María Academy", en tanto, ocupó el lugar 13 en el 2001, pero había ocupado el lugar 6 en 1998 y el lugar 25 en 1994.

Ante esta comprobada variabilidad de los *rankings*, cuesta imaginar los beneficios que podría traer a los estudiantes el que sus padres optaran por seguir la lógica del mercado educativo (véase la Ficha 11), corriendo de un lado a otro de la ciudad cada tres años para llevar a sus hijos a los "mejores" colegios.

Por otro lado, según fue explicado en la Ficha 8 y volverá a serlo un poco más adelante, muchas de las diferencias de posiciones no tienen ningún significado sustantivo. Algunas obedecen a diferencias de promedios que se ubican dentro de los márgenes conocidos de error de la medición, en tanto otras solo obedecen a cambios mínimos de puntajes que alteran toda la "tabla de posiciones".

2. Las principales debilidades y problemas de los rankings

2.1. La calidad educativa no puede ser resumida en un único indicador

La "calidad" educativa es un concepto complejo que abarca múltiples dimensiones, la mayoría de las cuales no son contempladas en las evaluaciones estandarizadas.

Una escuela puede tener buenos resultados porque expulsa o no logra retener a los malos alumnos y sus tasas de deserción son altas o porque, como veremos más adelante, selecciona alumnos de un determinado sector social.

Una escuela –como una región o un sistema– puede no tener excelentes resultados en Lenguaje y Matemática pero ser excelente en cuanto al clima educativo y a la formación personal de sus estudiantes. Esto no significa que la escuela no deba lograr ciertos aprendizajes fundamentales en todos sus alumnos. Sí significa que no necesariamente debe ocupar los primeros lugares en un *ranking* para ser considerada una buena escuela. La opción de la escuela puede ser lograr lo básico en dichas asignaturas y enfatizar otros aspectos en la formación de los alumnos.

En la evaluación internacional de Lectura PISA 2000, México quedó "posicionado" por encima de Chile (véase la Figura 4 en la Ficha 8 y la Figura 2 en la Ficha 9). Una conclusión rápida –y errónea– sería que los jóvenes de 15 años en México están mejor preparados para la Lectura que en Chile. Sin embargo, los jóvenes de 15 años asistiendo a clases en 7º grado o más (que es la población evaluada por PISA) en México constituyen apenas el 51% del total, en tanto en Chile constituyen el 87%. Es razonable suponer que el resto de los jóvenes de 15 años –ya sea esté todavía en la escuela primaria o haya abandonado los estudios – tenga una capacidad de Lectura inferior a la de los alumnos evaluados. Así, es probable que los jóvenes chilenos, como conjunto, estén mejor preparados que los mexicanos, más allá del lugar obtenido en el *ranking*.

Este ejemplo muestra la necesidad de considerar varios indicadores en forma simultánea a la hora de emitir valoraciones sobre un sistema o institución educativa. No basta con tener información sobre los resultados obtenidos por los alumnos en pruebas de rendimiento: hacen falta datos de cobertura, deserción, tipo de población atendida, etc.

Otro ejemplo de esta necesidad es el siguiente.

En México, en Matemática en 6º grado de Primaria, los "mejores" estados son Sinaloa, el Distrito Federal y Aguascalientes, en ese orden. Chiapas ocupa el lugar 15, Oaxaca el 17 y Veracruz el 28 (de un total de 32 estados).

Los porcentajes de población indígena son de 3,9% en el Distrito Federal, 3,4% en Sinaloa, 0,4% en Aguascalientes, 15,3% en Veracruz, 28,4% en Chiapas y 47,8% en Oaxaca. Cuando se analizan los resultados en Matemática de las escuelas indígenas por separado del resto, Oaxaca ocupa el primer lugar, Veracruz el segundo y Chiapas el cuarto, sobre un total de 18 estados con escuelas indígenas (Sinaloa, Distrito Federal y Aguascalientes no tienen escuelas indígenas).

Cuando se analiza solo los resultados de escuelas públicas urbanas, es decir, cuando se elimina las rurales y las privadas, Sinaloa y Aguascalientes siguen ocupando el 1er. y 3er. lugar respectivamente, pero el Distrito Federal, que estaba en el segundo lugar en el *ranking* general, pasa a ocupar el lugar 13. Esto se debe, en parte, a que el peso de la educación privada es muy fuerte en la capital mexicana[1]. Al eliminar el sector privado de la comparación, la posición del Distrito Federal cae.

1) Todos los datos sobre los estados mexicanos incluidos en los párrafos anteriores están tomados de Instituto Nacional para la Evaluación de la Educación (INEE), 2003, La calidad de la Educación Básica en México. Primer Informe Anual 2003; INEE, México D.F.

En realidad uno puede "rankear" a los estados mexicanos de 10 ó 12 maneras diferentes y en cada *ranking* las posiciones son diferentes, según el indicador que se elija. Lo mismo se aplica a los ordenamientos de escuelas.

2.2. La mayoría de los rankings dan una apariencia de precisión que es falsa

La publicación de rankings organizados como listas ordenadas en función de un único puntaje induce al público y a los usuarios de las evaluaciones a error, dando la falsa impresión de un ordenamiento preciso.

Como se analizó en la Ficha 8, toda medida está sujeta a error, el error puede ser estimado, y para que la diferencia de puntajes entre dos entidades sea "estadísticamente significativa" debe ser superior al error de medición. Si no lo es, no puede afirmarse que un resultado sea mejor que el otro.

Los informes de PISA intentan explicar esta situación a través de tablas en las que se indica el rango o variedad de posiciones distintas que cada país podría ocupar si se toma en cuenta el error de medición (véase la Figura 3). Esta tabla, referida a los resultados en Ciencias, muestra que la mayoría de los países no tiene una posición única en el ordenamiento. Para empezar, no hay un "primero" en esta tabla. Tanto Japón como Corea podrían ser "primeros". Canadá puede ser el cuarto país, pero también el octavo. Estados Unidos puede ocupar cualquier lugar entre el 11° y el 21°.

Esto mismo se aplica a los ordenamientos de escuelas. Por tanto, los *rankings* que ubican a cada escuela o país en una posición única y bien definida ofrecen una imagen que no se ajusta a la realidad y que no se sustenta en los datos.

Pero además, según fue expuesto en la Ficha 8, el hecho de que una diferencia de promedios sea estadísticamente significativa no implica que represente una diferencia relevante desde el punto de vista sustantivo.

Por ejemplo, la diferencia en Lectura entre México y Chile en PISA 2000 es "estadísticamente significativa". El primero de ellos tuvo un promedio de 422 puntos y el segundo 410. ¿Qué tan importante es esta diferencia de 12 puntos? ¿Refleja una diferencia sustantiva en las competencias de los jóvenes evaluados?

Si el lector observa nuevamente la Figura 2 en la Ficha 9, en la que aparecen los porcentajes de alumnos por niveles de desempeño, constata que ambos países tienen un 1% de sus alumnos en el nivel 5, México tiene 6% en el nivel 4 y 19% en el nivel 3, en tanto Chile tiene, respectivamente, 5% y 17% de sus estudiantes en dichos niveles.

En conjunto, México tiene un 3% más de alumnos por encima del nivel 3 que Chile. Esta no parece ser una diferencia de gran importancia. En cambio, Irlanda tiene el 71% de sus jóvenes por encima del nivel 3 (una diferencia de casi 50% con relación a Chile y de 47% con relación a México). Esto sí aparece como una diferencia relevante.

En resumen, el ordenamiento en *rankings* individualizados suele dar una falsa impresión. Por este motivo, resultaría más apropiado organizar la información en grupos de países o de escuelas con resultados similares, más que en series individuales de los mismos.

Este mismo criterio debe ser aplicado cuando se comparan los resultados en el tiempo de un país, provincia o escuela. Para establecer si los cambios que se observan en los resultados son "significativos" o no, es decir, para decir que una escuela, provincia o país mejoró o empeoró sus resultados, debe tenerse en cuenta tanto la significación estadística como la significación "sustantiva".

Posiciones variables en PISA 2000 Ciencias **Figura 3**

País	Rango	
	El más alto posible	El más bajo posible
Corea	1	2
Japón	1	2
Finlandia	3	4
Reino Unido	3	7
Canadá	4	8
Nueva Zelanda	4	8
Australia	4	8
Austria	8	10
Irlanda	9	12
Suecia	9	13
República Checa	10	13
Francia	13	18
Noruega	13	18
Estados Unidos	11	21
Hungría	13	21
Islandia	14	20
Bélgica	13	21
Suiza	13	21
España	**16**	**22**
Alemania	19	23
Polonia	19	25
Dinamarca	21	25
Italia	22	25
Liechtenstein	20	26
Grecia	25	29
Rusia	26	29
Letonia	25	29
Portugal	26	29
Luxemburgo	30	30
Méjico	31	31
Brasil	32	32

Fuente: Ministerio de Educación, Cultura y Deporte, Instituto Nacional de Calidad y Evaluación (INCE), 2001; Conocimientos y destrezas para la vida. Primeros resultados del proyecto PISA 2002. Resumen de Resultados. OCDE/ INCE, Madrid.

2. 3. Mientras los resultados nacionales suelen ser estables, los resultados por escuela son altamente volátiles

Los resultados en pruebas estandarizadas aplicadas a países y estados, que involucran grandes cantidades de alumnos, suelen cambiar lentamente a lo largo de los años.

En cambio, los resultados de escuelas individuales, normalmente basados en pequeñas cantidades de alumnos que suelen oscilar entre 20 y 90 para un determinado grado, son altamente volátiles.

Es común que una escuela logre destacados resultados un año y malos o no destacados dos o tres años más tarde. Ello depende de las características propias de cada cohorte de alumnos. Como todo docente sabe por experiencia, algunos grupos están más motivados y otros menos, con algunos grupos "da gusto" trabajar, con otros grupos "no se puede", sin que se pueda explicar con precisión qué es lo que ocurre.

Por otra parte, hay fenómenos de movilidad de alumnos y docentes que también inciden en esa volatilidad de los resultados.

Pero además, cuando la atención se fija en la "posición" de una escuela en un *ranking* y no en sus logros en términos criteriales (es decir, en qué proporción sus alumnos están alcanzando niveles de desempeño aceptables o destacados), los resultados se tornan aún más volátiles, porque dependen no solo de los cambios en la escuela en cuestión, sino de los cambios en las demás escuelas. Una escuela puede "caer" 25 lugares en el *ranking* sin que sus resultados hayan cambiado, simplemente porque otras mejoraron unos pocos puntos su promedio.

En este punto es conveniente realizar una distinción muy importante: mejor no es lo mismo que bueno y peor no es lo mismo que malo.

¿Qué significa esto? Que una escuela puede ocupar el 3er. lugar y otra el lugar 50, y ser ambas malas, porque en ninguna de ellas la mayoría de los alumnos logran los aprendizajes que deben lograr. Simplemente una es un poco menos mala que la otra.

Esta distinción tiene relación con los juicios de valor normativos y criteriales que analiza la Ficha 3 y con los "estándares" o expectativas analizados en la Ficha 9. Una cosa es que una

escuela esté en mejor situación que otra en términos relativos, y otra cosa distinta es si en dicha escuela los alumnos están aprendiendo lo que se espera de ellos. Para los padres y para los docentes es más relevante conocer qué proporción de los estudiantes de una escuela está logrando niveles de desempeño aceptables, que la posición relativa de la escuela en un *ranking*. Esto último puede ser un buen recurso periodístico para vender más periódicos, pero aporta poca información relevante para los actores educativos.

2.4. La mayoría de los rankings no tiene en cuenta la composición social del alumnado

Como se mostró en las Figuras 1 y 2, los "mejores" colegios suelen ser aquellos que trabajan con alumnos provenientes de las familias más educadas y en mejor situación económica.

Esto es bastante obvio, dado que las escuelas que trabajan con los niños y jóvenes que provienen de los sectores más desfavorecidos tienen que afrontar obstáculos tales como un manejo insuficiente del lenguaje por parte de los niños, una menor preparación de las familias para apoyar al niño en las tareas escolares, la falta de material escrito y de estudio en los hogares, etc[2]. Su labor es entonces mucho más compleja que la de las escuelas de contextos sociales favorecidos, dado que en estos sectores los niños desarrollan en sus hogares capacidades que luego hacen mucho más fácil la tarea de la escuela.

Este hecho ha sido comprobado por múltiples investigaciones. El estudio PISA, por ejemplo, informa que:

> *"Como promedio en los países miembros de la OCDE, los estudiantes cuyas madres no han terminado los estudios de educación secundaria superior tienen una notoria desventaja, obteniendo puntuaciones en lectura 44 puntos menores que aquellos estudiantes cuyas madres han terminado la educación secundaria superior... El impacto de la terminación, por parte de las madres, de la educación superior (terciaria) es más débil y menos consistente a lo largo de los países... La educación de los padres está estrechamente relacionada con otros factores del entorno familiar. No obstante, cuando el resto de los factores del entorno familiar son iguales, cada año adicional de educación de los padres añade, al menos, 4,7 puntos a las puntuaciones de los estudiantes"*[3].

Nótese que esto ocurre en los países de la OCDE, en que la población tiene mayor

2) Este tema ha sido tratado con detalle en el capítulo denominado "Comparando lo incomparable: la parábola de las carreteras" en Ravela, P. 2001; ¿Cómo presentan sus resultados los sistemas de evaluación educativa en América Latina", PREAL/GRADE.

3) Ministerio de Educación, Cultura y Deporte, Instituto Nacional de Calidad y Evaluación (INCE), 2001; Conocimientos y destrezas para la vida. Primeros resultados del proyecto PISA 2002. Resumen de Resultados. OCDE/ INCE, Madrid, p. 27.

FICHA 10

Figura 4 **Matemática – 9° grado – Uruguay, 1999**

Nivel sociocultural individual de la familia del alumno	Nivel sociocultural del centro al que asiste el alumno	
	Bajo	**Alto**
Alto	58,6%	79,8%
Medio alto	51,1%	69,8%
Medio bajo	43,5%	58,7%
Bajo	40,6%	52,7%

Fuente: Administración Nacional de Educación Pública / Programa MESyFOD, 1999; Censo Nacional de Aprendizajes 1999 en 3ros años del Ciclo Básico. Primera Comunicación de Resultados. Montevideo, ANEP/ MESyFOD.

Recuadro 1

"EL EFECTO DE LOS PARES"

Estudio evalúa el efecto de los pares o compañeros sobre el logro

"El estudio del Instituto de Políticas Públicas de California con sede en San Francisco halló, por ejemplo, que los compañeros de los estudiantes tenían un efecto más consistente sobre el logro que el hecho que sus maestros tuviesen estudios avanzados…

Según cálculos de los investigadores, un estudiante de educación primaria que pasa de un grupo de pares con un nivel de logro bajo a uno con logro alto podría esperar una mejora de un 9 por ciento en las pruebas…

En menor grado, lo mismo sucedía cuando los alumnos estudiaban en aulas con alto desempeño. No obstante, los efectos compañero del aula eran más fuertes en la escuela primaria que en la secundaria, probablemente porque los estudiantes cambian de aula más a menudo en grados superiores, señalaba el estudio…

'Creemos que lo que está sucediendo aquí es que, si este año tienes una mejor cohorte de alumnos a tu alrededor, otros estudiantes te enseñan matemática, dijo Mr. Betts. `Y si los maestros de una escuela notan, por ejemplo, que la cohorte de 5° grado es inusualmente fuerte, eso puede realmente alentarlos a presionar a sus alumnos a alcanzar un nivel más alto…

Un alumnos de 5° grado que asiste a una escuela de San Diego del quintil socioeconómico más alto, por ejemplo, lee tan bien como uno de 10° grado matriculado en una escuela del grupo más pobre".

Traducido de : Debra Viadero, Education Week, 10 de setiembre de 2003.

cantidad de años de educación que en nuestra región. En esos países un nivel educativo bajo es no haber completado la secundaria (la categoría primaria incompleta no existe para ellos). Por tanto, mucho mayor es la importancia de estos factores en nuestros países, en los que la escolaridad de la población es bastante inferior a la de la OCDE.

Pero lo más importante es comprender la diferencia entre nivel sociocultural individual y grupal.

Lo que incide en los resultados no es tanto el origen social individual de cada alumno en sí mismo, sino la composición social del grupo.

Por lo general cuando los alumnos de origen social desfavorecido son una minoría que forma parte de un grupo o escuela en que la mayoría de los alumnos son de origen social más alto, se ven favorecidos por el efecto grupal. A la inversa, cuando la mayoría de los alumnos pertenecen a sectores desfavorecidos, los pocos alumnos de origen más favorecido aprenderán menos, porque el nivel de trabajo estará determinado por las posibilidades de la mayoría.

Esto ha sido comprobado a nivel internacional por el estudio PISA, que titula uno de los apartados sobre el tema de la siguiente manera: "La composición social de la población de estudiantes de una escuela es un mejor predictor del rendimiento de los alumnos que el entorno social individual". En dicho apartado se establece[4]:

"El proyecto PISA muestra, por ejemplo, que se puede esperar que dos estudiantes con las mismas características familiares que asisten a escuelas diferentes –una con un perfil social más alto y otra con uno más bajo– se sitúen más lejos uno de otro en la escala de habilidad lectora que dos estudiantes de entornos diferentes que asisten a la misma escuela. Aunque este fenómeno tiene causas complejas, subraya el enlace potencial entre la segregación social de los estudiantes en escuelas diferentes y la polarización de los estudiantes en cuanto a rendimiento..."

4) OCDE/ Ministerio de Educación, Cultura y Deporte, Instituto Nacional de Calidad y Evaluación (INCE), 2001, op.cit.; p.29

La Figura 4 ilustra este fenómeno con datos de Uruguay. En cada casilla se indica el porcentaje de alumnos que alcanzó un nivel de suficiencia en la prueba. Las casillas indican el nivel sociocultural individual del alumno y el nivel sociocultural del centro al que asiste.

Estos datos permiten constatar que, a igual nivel sociocultural individual, los resultados son marcadamente diferentes según sea el nivel sociocultural del centro. Dicho en otras palabras,

la probabilidad de alcanzar la suficiencia en Matemática para un alumno de nivel sociocultural bajo es de 40,6% si está en un centro que también es de nivel sociocultural bajo, pero sube a 52,7% si el alumno está en un centro de nivel sociocultural alto.

A la inversa, un alumno de nivel sociocultural alto tiene una probabilidad de ser suficiente en Matemática de 79,8% si está en un centro de nivel sociocultural alto, pero cae al 58,6% si se encuentra en un centro de nivel sociocultural bajo.

Este hecho ha sido constatado en investigaciones realizadas en los Estados Unidos, que han denominado al fenómeno como "el efecto compañero" o "efecto de los pares" (véase el recuadro 1).

El desconocimiento de este hecho lleva a conclusiones posiblemente inapropiadas, como parece ocurrir en el informe elaborado por Llach y otros en la República Argentina. Al comparar los resultados individuales de alumnos de origen pobre en escuelas públicas y privadas, estos últimos exhiben mejores resultados (véase la Figura 5).

Figura 5 **Puntajes promedio de Lenguaje y Matemática en 1997 en Argentina según tipo de establecimiento**

	Escuelas estatales	Escuelas privadas
Nivel Socioeconómico		
Bajo	45,2	48,7
Medio	52,9	60,2
Alto	59,6	68,7
Nivel educativo de la madre		
Primario incompleto	42,6	47,4
Primario completo	51,1	58,0
Secundario completo	56,1	66,0
Terciario completo	58,7	68,8

Fuente: Extraído de Llach, J.J. y otros, 2000; Educación para todos; DISTAL, Buenos Aires, p. 235.

Plantear el análisis de los datos de ese modo lleva a los autores a concluir que "se observa que los chicos de las escuelas primarias privadas obtienen mejores puntajes que los de las estatales para todos los niveles socioeconómicos y para todos los niveles de educación de los padres"[5].

La afirmación anterior es cierta, según los datos presentados en la Figura 5. Sin embargo, de allí no se puede derivar un juicio de valor acerca de la calidad de escuelas públicas y privadas, dado que no se tiene en cuenta el hecho de que en las escuelas públicas los alumnos de origen desfavorecido son mayoría, en tanto en las escuelas privadas constituyen una minoría. Por tanto, en dichas escuelas los alumnos de nivel socioeconómico "bajo" se ven beneficiados por el "efecto compañero" o "efecto de los pares".

Esta diferencia conceptual entre nivel socioeconómico individual y composición sociocultural de la población escolar tiene importantes implicaciones tanto para las políticas educativas como para las comparaciones de resultados entre escuelas.

En términos de políticas educativas, la consecuencia es que debería promoverse la heterogeneidad sociocultural en las escuelas y debería evitarse, en la medida de lo posible, la selección social de estudiantes por parte de los centros educativos. Esto ha sido expresado en las conclusiones de PISA 2000 en los siguientes términos[6]:

"Los resultados del proyecto PISA tienen implicaciones importantes para la política de los sistemas educativos. En algunos países se segrega a los alumnos en gran medida en términos de variables socioeconómicas, en parte debido a la propia segregación por razones de residencia y factores económicos, pero también debido a las características del sistema de escolarización. La política educativa en tales países podría intentar moderar el impacto del entorno social sobre el rendimiento de los alumnos, bien reduciendo la magnitud de la segregación basada en patrones socioeconómicos o bien asignando de modo diferencial los recursos a las escuelas".

En términos de comparaciones entre escuelas, lo anterior implica que las mismas deberían efectuarse entre escuelas que atienden poblaciones más o menos similares en cuanto al origen social de su alumnado (sin omitir una referencia a los resultados generales del país).

5) Llach, J.J. y otros, 2000; Educación para todos; DISTAL, Buenos Aires, p.236

6) OCDE/Ministerio de Educación, Cultura y Deporte, Instituto Nacional de Calidad y Evaluación (INCE), 2001, op. cit.; p. 29.

3. Dos enfoques para el uso de ordenamientos de escuelas

Más allá de los problemas descritos en las páginas anteriores, ordenar a las escuelas según sus resultados en pruebas estandarizadas puede ser útil para muchos efectos. La primera cuestión a dilucidar es cuál es el propósito de tal ordenamiento.

Existen dos grandes tipos de propósitos.

Uno es dar a publicidad esos ordenamientos como forma de generar un mercado educativo o para establecer un sistema simbólico de premios y castigos a las escuelas en función de sus resultados.

El otro es establecer programas de apoyo a las escuelas con dificultades y aprender de las escuelas con buenos resultados.

Si bien estos dos tipos de propósitos teóricamente no son incompatibles, en la práctica muchas veces lo son, porque cuando se establece un sistema de incentivos y sanciones externas, se torna difícil que los docentes se apropien de la función formativa de las evaluaciones. Más bien tienden a generarse comportamientos de rechazo, de cumplimiento burocrático de lo que las evaluaciones exigen e, incluso, conductas directamente reñidas con la ética profesional, como hacer que los malos alumnos no asistan el día de la prueba.

Por detrás de esta disyuntiva de propósitos existe una cuestión central para la política educativa contemporánea: ¿cuál es la llave del cambio en las prácticas de enseñanza de los profesores: el establecimiento de incentivos externos o la creación de oportunidades de desarrollo profesional? ¿O alguna combinación de ambos?.

Este tema será abordado en la Ficha 11. Pero cabe adelantar que la experiencia internacional muestra que los sistemas de *rankings* públicos generan más problemas que mejoras (véase el recuadro 2).

En América Latina, Chile es el único país que sistemáticamente ha publicado *rankings* de escuelas durante una década. Sin embargo, un reciente estudio de la demanda educativa en Chile muestra que en realidad los resultados del SIMCE no son usados por las familias como elemento de decisión (véase el recuadro 3).

Recuadro 2

¿Vale la pena concentrar la atención en producir *rankings* de escuelas?

"En el Reino Unido, por ejemplo, donde las políticas gubernamentales recientes han centrado la atención en la rendición de cuentas y el monitoreo de estándares, el uso de indicadores de rendimiento ha tenido un impacto notable sobre las escuelas. En ese marco ocupa un lugar destacado la implementación, desde 1990, de un currículo nacional, evaluaciones igualmente nacionales, un sistema de inspección externa… y la publicación de los puntajes promedio no ajustados obtenidos por los alumnos de cada escuela en las pruebas… Las tablas de posiciones consisten en rankings de las escuelas basados en los puntajes promedio (crudos y sin ajuste) obtenidos en las pruebas nacionales por los alumnos de 7, 11 y 14 años… La intención es que esas tablas sean utilizadas por los padres para seleccionar la escuela a la que enviarán a sus hijos…

Las consecuencias para las escuelas que reciben la etiqueta de "deficiente" o "no efectiva" pueden ser catastróficas… El impacto de los *rankings*, a partir de la experiencia reciente del Reino Unido, ha sido evidente en ataques políticos y de los medios contra las escuelas y los maestros; un currículo dominado por las pruebas…; cabildeo por parte de los directores para que se les permita aplicar criterios selectivos para la admisión de hasta el 20% de su alumnado… en algunos casos rechazo abierto de inscribir niños de bajo rendimiento…; en algunas escuelas concentración de los esfuerzos en los mejores alumnos; algunos padres de familia han llegado a cambiar su domicilio… (Rowe, 2003).

El texto de Rowe añade que incluso en donde no se publican puntajes crudos, sino medidas más refinadas de valor agregado, el uso de *rankings* hace inevitable que haya ganadores y perdedores, y una vez que una escuela es etiquetada como deficiente es difícil encontrar la manera de ayudarla en un ambiente en el que prevalece la postura de vergüenza y recriminación, que no lleva a la formulación de estrategias de mejora".

Fuente: Instituto Nacional para la Evaluación de la Educación (INEE), 2004; Plan Maestro de Desarrollo; México D.F., INEE, Anexo B.

Simultáneamente, un estudio de la OCDE sobre el sistema educativo chileno señala que la apuesta al mercado no ha tenido impactos relevantes en la mejora de los aprendizajes y que, en cambio, sí lo han tenido programas como el conocido P-900 ("Programa de las 900 escuelas"), dirigido a apoyar a las escuelas con mayores dificultades (véase el recuadro 4).

FICHA 10

¿por qué los rankings son modos inapropiados de valorar la calidad de las escuelas?

Recuadro 3

Los criterios de los padres para elegir la escuela de sus hijos en Chile

"La Escuela de Gobierno de la Universidad Adolfo Ibáñez investigó el modo en que los padres seleccionan la escuela a la que irán sus hijos en primero básico. Se encontró con resultados sorprendentes, como que los padres privilegian el ambiente por sobre la calidad, al menos la que se mide en el SIMCE y que obsesiona a los expertos...

Las razones más importantes de los padres para escoger la escuela de sus hijos fueron su cercanía, la calidad (como un concepto amplio), las referencias, la instrucción religiosa y el currículo. A diferencia de los resultados que se obtienen en otros países, menos del 1% nombró la prueba estandarizada (SIMCE) y el tamaño del curso como las más importantes.

Hay diferencias según niveles sociales. El 68% de los padres con menos educación escogió la escuela de sus hijos por razones prácticas (cercanía y gratuidad), mientras solo el 27% de los padres con educación superior lo hizo por esta razón. Las familias con hijos en escuelas particulares (subvencionadas y pagadas) ponen más énfasis en los valores y el currículo. Pero este grupo no prioriza la calidad más que el otro, aunque puede deberse a razones distintas. Tal vez los padres con más recursos dan por descontada la calidad de sus opciones, y por eso pueden "darse el lujo" de escoger según otros aspectos, como los valores.

La interpretación más positiva es que las escuelas particulares tienden a inculcar ciertos valores en sus alumnos y tienen un currículo más innovador, que les permite distinguirse de las escuelas municipales. Investigadores menos optimistas han concluido que lo que distingue básicamente a una escuela particular de una pública en Chile es su religión y su exitoso *marketing*.

En este estudio, sólo 4 de 536 padres supo (al menos aproximadamente) el resultado SIMCE de la escuela donde estudia su hijo. Esto es menos del 1% de la muestra. Un poco mejor les fue cuando se les pidió ubicar al colegio en un ranking comunal: cerca del 46% sabía si el establecimiento estaba sobre el promedio, en él o bajo él".

MarcelaAguilar - El Mercurio - La Revista Sábado – 30 de octubre de 2004 / (subrayados añadidos). http://diario.elmercurio.com/2004/10/30/el_sabado/reportajes/noticias/FBFEB518-09AD-467D-9516-6D415213C716.htm

Recuadro 4

Conclusiones del Informe OCDE sobre el sistema educativo en Chile

"El equipo de revisión comprendió que aun cuando el sistema de subvenciones y otras políticas asociadas con una visión de mercado de la educación demostraran no mejorar el aprendizaje de los estudiantes, estas serían difíciles de terminar... no fue obvio para el equipo de revisión que la teoría subyacente de aumentar la efectividad a través de la competencia funcione en la práctica...

"...las mejoras en rendimientos son mayores que las del promedio en el caso de las escuelas básicas que han sido objeto de programas focalizados, como el P-900 y el Programa Rural...

"...para mejorar la calidad de la educación y el rendimiento de los estudiantes se necesita centrar la atención a largo plazo en las destrezas y roles del profesor. La importancia del SIMCE en la medición del rendimiento de los estudiantes en la educación chilena destaca la urgencia de asegurar que todos los profesores tengan "cultura de evaluación" en la sala de clases...

"El gobierno debería pensar en proporcionar perfeccionamiento en servicio adicional a los profesores en el tema de evaluación de estudiantes , los profesores se beneficiarían con un vínculo más estrecho entre la información recogida a través del SIMCE y sus evaluaciones de la sala de clases".

OCDE, 2004; "Revisión de Políticas Nacionales de Educación". Publicado en: http://biblioteca.mineduc.cl/documento/Texto_Libro_OCDE1.pdf

Los análisis y conclusiones consignados en los recuadros anteriores constituyen evidencia de que el uso de ordenamientos de escuelas para la toma de decisiones de apoyo a las escuelas con dificultades tendría mayor impacto que su uso público como información para el "mercado educativo".

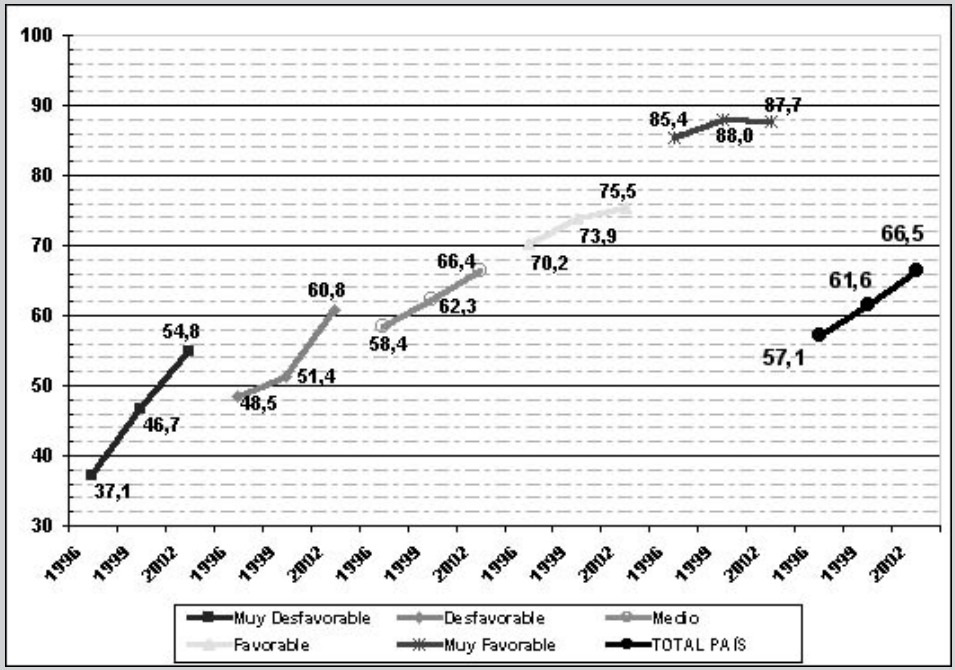
La experiencia de Uruguay apoya esta idea

En este país los resultados de cada escuela primaria, producidos en 1996, no fueron divulgados públicamente, pero se emplearon para establecer programas de compensación económica y de capacitación en servicio para los maestros de las escuelas de los contextos sociales más desfavorecidos.

Estos programas implicaron la participación voluntaria de equipos de los docentes (no de docentes individuales) de las escuelas de contextos desfavorecidos en programas de capacitación en servicio que se desarrollaban en días sábados, fuera del horario escolar, durante todo el año lectivo. Los maestros participantes percibían una compensación en su remuneración normal.

Figura 6

Porcentaje de alumnos que alcanzan el nivel de suficiencia en la Prueba de Lenguaje según contexto sociocultural de la escuela. Uruguay, 1996 – 2002.



Fuente: ANEP/Gerencia de Investigación y Evaluación, 2003; Evaluación Nacional de Aprendizajes en Lenguaje y Matemática. 6º año enseñanza primaria – 2002. Segundo Informe.

Como consecuencia de este enfoque –así como de otras políticas de apoyo a dichas escuelas y de la creación de escuelas de tiempo completo en sectores populares–, en los años subsiguientes se produjo un incremento en los porcentajes de alumnos que alcanzaron un nivel de suficiencia en las pruebas estandarizadas, incremento que estuvo marcadamente focalizado en las escuelas de los contextos más desfavorecidos (véase la Figura 6).

Según se puede apreciar en el Gráfico de la Figura 6, los resultados mejoraron notoriamente en las escuelas de contextos desfavorecidos y se mantuvieron estables en las de contextos favorables, reduciéndose de este modo la inequidad en el acceso a los aprendizajes al final de la enseñanza primaria.

Lo interesante de esta experiencia radica en que se establece un estímulo económico que no está asociado a los resultados de las pruebas sino al contexto social de las escuelas y a la voluntad de los maestros de capacitarse colectivamente para desarrollar mejor su tarea.

Síntesis final

Esta Ficha tuvo como propósito explicar al lector los problemas que presenta el uso público de *rankings* con resultados de escuelas. Se ha intentado mostrar que ordenar las escuelas en función de sus resultados puede ser útil para la toma de decisiones de política educativa, fundamentalmente en términos de inversión, apoyo y capacitación en los sectores más desfavorecidos del sistema, pero que el uso de *rankings* públicos como instrumento para introducir competencia entre las escuelas no parece favorecer la mejora del sistema educativo. Sobre estos temas se volverá en las Fichas 11 y 12. Más allá de la discusión sobre la conveniencia o inconveniencia de los *rankings*, a lo largo de la Ficha se explicó los principales problemas técnicos que tiene la elaboración de los mismos.
De los cuatro grandes tipos de problemas analizados se desprenden algunos criterios fundamentales que debieran tenerse en cuenta al efectuar comparaciones entre escuelas:

a. Utilizar diversos indicadores de calidad y no uno solo, lo que probablemente dará lugar a diversos ordenamientos, dado que unas escuelas destacarán en ciertos aspectos y otras en otros.

b. Cuando se ordena a las escuelas por un único indicador como, por ejemplo, sus resultados en una prueba de Matemática, evitar denominar "calidad" a este único indicador. Es legítimo

ordenar a las escuelas en función de un determinado resultado o indicador, pero es importante explicitar que ello no da cuenta de todo lo que las escuelas realizan.

c. Tener en cuenta la composición social del alumnado de cada escuela. Si esto no se hace, lo que se está comparando principalmente es la selección social del alumnado. Esto implica que es preferible realizar varios ordenamientos o *rankings* por separado, para escuelas de similar composición social.

d. Utilizar datos de varios años o varias mediciones para evitar ofrecer imágenes falsas que obedecen a factores aleatorios, dado que los resultados de escuelas individuales son altamente volátiles y cambiantes.

e. Controlar la significación estadística de las diferencias, tanto en la comparación entre escuelas en un mismo momento del tiempo, como en la comparación de los resultados de una misma escuela a lo largo del tiempo. Esto necesariamente llevará a relativizar las posiciones en el *ranking*, dado que cada escuela no tiene una única posición sino varias posibles. Por este motivo, es preferible trabajar con agrupamientos de escuelas –escuelas de muy buenos resultados, escuelas de resultados intermedios, escuelas de resultados inferiores, etc.– más que con un *ranking* u ordenamiento individualizado.

f. Focalizar las comparaciones en la proporción de alumnos que están logrando los aprendizajes que deberían lograr, más que en las posiciones relativas de las escuelas en función de sus promedios. Recuérdese al respecto que una escuela puede tener un promedio superior a otra, pero ambas ser malas en términos de lo que sus alumnos están logrando.

¿QUÉ ES LA RENDICIÓN DE CUENTAS?

Hacia la responsabilidad compartida

A lo largo de la última década han venido creciendo las demandas por utilizar los resultados de pruebas estandarizadas para hacer que escuelas y docentes "rindan cuentas" de su trabajo.

La Ficha 11 se propone explicar los diversos significados del término "rendición de cuentas" y las visiones contrapuestas que existen acerca de la misma. Se analizan los principales debates en torno a la rendición de cuentas y algunas experiencias relevantes que se lleva adelante en diversos lugares del mundo.

Se explica asimismo los efectos "perversos" o no deseados que algunos mecanismos de rendición de cuentas pueden desencadenar.

Al mismo tiempo, se destaca la importancia de la rendición de cuentas como mecanismo de involucramiento de la ciudadanía con la educación, de transparencia democrática y de mejora de la gestión y de la enseñanza que se brinda a los estudiantes.

Para ello se analizan las peculiaridades de la acción educativa que por un lado hacen necesaria la rendición de cuentas pero por otro exigen encararla con una lógica articuladora y no meramente como señalamiento de culpables de los problemas educativos.

Por este motivo se enfatiza la multiplicidad de actores que deberían hacerse responsables por los resultados en el sistema educativo: los directivos y docentes, pero también las autoridades centrales y locales, los técnicos, las familias y los propios alumnos.

1. El problema de fondo: la "opacidad" de la labor educativa

Por su propia naturaleza, el resultado del trabajo educativo es poco visible. No puede ser percibido directamente. Un niño que no recibe la formación apropiada recién se dará

cuenta de ello cuando sea un adulto joven y se enfrente a sus propias limitaciones. El carácter simbólico –inmaterial, intelectual y afectivo– de la labor educativa y de lo que la misma genera en los alumnos, hace difícil percibir qué está pasando.

A pequeña escala, en las escuelas, los profesores y los padres pueden construir una percepción acerca de si un alumno aprende o no aprende, o en qué grado lo hace. Pero esta percepción está determinada por las propias categorías mentales del profesor o del padre.

Un profesor mejor formado y más exigente tenderá a esperar más de sus alumnos y a considerar que no han aprendido lo suficiente. Un profesor menos formado y/o menos exigente tenderá a considerar que sus alumnos han logrado lo suficiente aun cuando desde otra perspectiva esto pueda resultar insuficiente. Ello se refleja muchas veces en discrepancias importantes cuando se conforman tribunales o jurados para examinar estudiantes.

La percepción de los padres es aún más complicada. Primero, porque está mediada por la percepción que el profesor les ofrece a partir de las calificaciones que le otorga al alumno. Si el alumno recibe una buena calificación, los padres, sobre todo los que han tenido menos trayectoria en la educación formal, tenderán a pensar que el alumno está aprendiendo aunque ello pueda no ser así. Los padres más formados, que tuvieron ellos mismos una enseñanza más exigente, tal vez pongan en duda las calificaciones que sus hijos reciben y consideren que los profesores "de ahora" son muy condescendientes.

En segundo lugar, algunos padres pueden mostrarse insatisfechos con la enseñanza, simplemente porque el modo de enseñar y los "temas" que se enseñan han cambiado –sobre todo en la educación primaria– y *"ya no se enseña como antes el sujeto y el predicado, los verbos, las reglas de ortografía..."*. Pueden no comprender qué es lo que la escuela y el maestro están intentando hacer.

Cuando se pasa de la escala micro a la gran escala, es decir, a la formulación de un juicio de valor sobre lo que se está aprendiendo en un sistema educativo, la situación se torna más compleja todavía. El deterioro o la mejora solo puede ser percibido en el mediano y largo plazo, y se requiere de dispositivos especiales de producción de información –las evaluaciones estandarizadas, entre ellos–.

En otras áreas de la vida es más sencillo evaluar la situación: uno se siente mal físicamente y consulta al médico, tiene que esperar el ómnibus durante mucho tiempo y se queja por la baja frecuencia del servicio, la luz se corta, el pan está duro y decide no comprar más en esa panadería.

En la educación todo es más complejo. Uno puede percibir lo superficial: el profesor es atento, es puntual, intenta que todos entiendan, el liceo está limpio o sucio, hay material de estudio, hay computadoras. Estos aspectos son importantes para la educación, pero no son en sí mismos la educación. Pero son visibles, por lo cual es más fácil valorarlos –y también, por esta razón muchas veces la tentación de los gobiernos es concentrarse en este tipo de elementos, que son importantes, pero que, sobre todo, se pueden mostrar al público–.

Esta intangibilidad de lo que el sistema educativo logra en las nuevas generaciones da lugar muchas veces a que las instituciones educativas en cierto modo queden eximidas de responsabilidad sobre los resultados de su tarea. No deben dar cuenta a nadie de lo que están logrando, porque lo que están logrando no es visible. Puede haber diversidad de opiniones, pero no hay bases ciertas para establecer juicios de valor sobre el quehacer educativo.

El tema de la "rendición de cuentas" o "responsabilidad por los resultados" tiene relación directa con esta característica central de la labor educativa.

2. Distintas versiones de la "rendición de cuentas" o "responsabilidad por los resultados"

La preocupación por la "rendición de cuentas" surge principalmente en el mundo anglosajón y está relacionada con la afirmación de que el público tiene derecho a saber qué es lo que está ocurriendo dentro del sistema educativo. "Rendición de cuentas" es uno de los modos de traducir el término inglés *accountability*. Otro modo de traducirlo, a nuestro juicio más apropiado, es el de "responsabilidad por los resultados"[1].

Bajo estos rótulos generales existe una diversidad de propuestas de política educativa, muchas de las cuales involucran el uso de los resultados de las pruebas estandarizadas con consecuencias "fuertes" más que con carácter formativo. A continuación se explican las principales "versiones" de la *"rendición de cuentas / responsabilidad por los resultados"*.

1) El boletín Nº 15 de la Serie "Políticas" de PREAL (Santiago, julio de 2003) está dedicado a este tema y emplea indistintamente ambas expresiones (rendición de cuentas y responsabilidad por los resultados).

2.1. Versión 1: ¿Qué se hace con el dinero destinado a la educación?

Una primera versión parte de la siguiente preocupación principal: quienes pagan impuestos tienen derecho a saber cómo se está usando el dinero que se destina a la educación –y de allí la expresión "rendir cuentas"–. En realidad se trata también de un argumento empleado por algunos gobiernos para ajustar los presupuestos educativos y para hacer más eficiente el gasto en educación.

Si bien la realidad de los países latinoamericanos en materia de presupuesto educativo es radicalmente diferente a la de los países anglosajones –por lo insuficiente de los presupuestos–, la preocupación por el correcto uso de recursos escasos y por la "rendición de cuentas" pública en relación al uso de los recursos de endeudamiento externo destinados a la educación, parece una preocupación pertinente que debe estar en la agenda educativa .

2.2 Versión 2: La educación como asunto público

En una versión más amplia del concepto, la *"rendición de cuentas/responsabilidad por los resultados"* significa mantener informada a la sociedad sobre la situación y los problemas del sector educativo, de modo que la ciudadanía se informe, se preocupe y se involucre en los asuntos educativos.

Dentro de esta versión es posible identificar dos modalidades opuestas. Algunos creen que el modo de involucrar a la sociedad es buscar que la misma ejerza presión sobre los educadores y las escuelas para que hagan mejor su trabajo. Muchos consideran que el único modo de generar cambios en el sector educativo es crear un sentido de "emergencia educativa" en la población.

Desde esta perspectiva se privilegia el uso de los resultados para destacar carencias en el aprendizaje, malos resultados, sobre todo de la educación pública, obviando generalmente la consideración del contexto de problemas sociales en que la educación desarrolla su labor.

Una modalidad alternativa consiste en involucrar a la sociedad en la educación construyendo apoyos a la misma y conciencia de que la tarea educativa es una responsabilidad común y no exclusiva de las escuelas –si bien estas tienen un rol preponderante–.

Desde esta perspectiva se privilegia el uso de los resultados para comprender mejor la

complejidad de las situaciones en que la escuela intenta enseñar y educar, para ayudar a los docentes a realizar mejor su trabajo, para construir capacidad de intervención pedagógica.

2.3. Versión 3: El mercado educativo y el control de los padres sobre las escuelas

Un modo específico de implementar un sistema de *"rendición de cuentas/responsabilidad por los resultados"* es la realización de evaluaciones censales –todas las escuelas son evaluadas en ciertos grados– y la posterior publicación en la prensa de listados o *rankings* con los resultados de cada escuela.

Mediante este procedimiento se busca que las familias, consideradas como usuarios o clientes de las escuelas, tengan información sobre la "calidad" de la educación que brinda cada escuela (el término "calidad" está entre comillas porque es discutible que los puntajes en pruebas estandarizadas constituyan por sí solos la "calidad" de la educación, si bien son un indicador importante de la misma).

De esta manera se busca hacer a las escuelas responsables ante sus clientes y posibilitar que los padres, o bien ejerzan presión sobre las escuelas para que mejoren su labor, o bien "voten con los pies" cambiando a sus hijos a otra escuela. Esto último, a su vez, contribuiría a generar competencia entre las escuelas para retener a los padres, lo cual sería otro modo de hacer que las escuelas se preocupen por mejorar sus resultados.

2.4. Versión 4: Incentivos y sanciones a los docentes y escuelas en función de los resultados de sus alumnos

Otro modo específico de implementar la *"rendición de cuentas/responsabilidad por los resultados"* es establecer incentivos y sanciones para las escuelas y/o maestros, vinculados a los resultados obtenidos por sus alumnos en pruebas estandarizadas (o también a una combinación de ellos con otros indicadores educativos). Los incentivos, en general, son económicos y pueden otorgarse con carácter individual (como en el Programa Carrera Magisterial de México) o colectivo (como en el Sistema Nacional de Incentivos Educativos -SNED- de Chile; véase el recuadro 1).

FICHA 11

Recuadro 1

El Sistema Nacional de Evaluación del Desempeño Docente en Chile

El Sistema Nacional de Evaluación del Desempeño de los Establecimientos Educacionales Subvencionados (SNED) se aplica en el país a partir de 1996. Este identifica a los colegios que reciben subvención estatal (tanto particulares subvencionados como municipales) de mejor desempeño en cada región del país.

El SNED introduce dos elementos adicionales muy importantes dentro del conjunto de reformas educativas del país: (i) posibilita una mejor información sobre el resultado educativo de los alumnos; (ii) establece un incentivo a los docentes para mejorar el resultado del proceso educativo, lo que constituye un incentivo a la oferta. El incentivo que provee el SNED – ya no derivado de la elección de colegios por parte de los alumnos y apoderados, sino asociado directamente a los docentes–, es un complemento importante del actual sistema educativo, al permitir superar problemas de falta de competencia en algunos lugares alejados del país e información insuficiente de los padres para elegir el colegio de sus hijos.

El SNED otorga una subvención por excelencia a los establecimientos cuyo desempeño califica de excelente. Esta es entregada cada dos años, en base a una selección de dichos establecimientos que, en total, representan el 25% de la matrícula regional. Está establecido que el 90% de los montos asignados debe destinarse directamente a los profesores del establecimiento de acuerdo a sus horas cronológicas, en tanto la distribución del 10% restante es definida por cada centro educacional. La subvención por desempeño de excelencia finalmente corresponde a un monto en pesos que se entrega trimestralmente a los docentes durante dos años...

El SNED es un sistema de evaluación de los docentes que tiene como componente principal los resultados académicos de los alumnos, medidos a través de la prueba SIMCE. Sin embargo, este test presenta dificultades de comparabilidad entre establecimientos, dado que sus resultados están, en parte, determinados por las características socioeconómicas de las familias de los estudiantes. Para suplir esta deficiencia y comparar establecimientos de similares características, tanto socioeconómicas, de tipo de enseñanza, como geográficas, el SNED construye "grupos homogéneos" a nivel regional.

(Ver Cuadro 1 en la página siguiente)

Fuente: Mizala, A. y Romaguera, P., 2003; Desafíos Metodológicos de los Sistemas de Evaluación e Incentivos en Educación. El caso del SNED en Chile. PREAL/GDN.

Cuadro 1: Factores e Indicadores del SNED

Factor (ponderador)	Indicador
Efectividad (37%)	- SIMCE (lenguaje y matemáticas, últimas pruebas disponibles) 4° año enseñanza básica; 8° año enseñanza básica; 2° año enseñanza media
Superación (28%)	- Diferencia promedio SIMCE (últimas dos pruebas disponibles para cada nivel) 4° año enseñanza básica; 8° año enseñanza básica; 2° año enseñanza media
Iniciativa (6%)	- Actividades e iniciativas del establecimiento, medida a través de una encuesta
Mejoramiento condiciones de trabajo (2%)	- Clasificación del establecimiento en sistema de inspección del Ministerio de Educación
Igualdad de oportunidades (22%)	- Tasas de retención y aprobación de alumnos - No existencia de prácticas discriminatorias. Entre ellas: cancelación de matrícula a alumnos repitentes; cancelación de matrícula a alumnas por razones de embarazo o maternidad y negación de matrícula a postulantes, a pesar de existir vacantes. - No existencia de sanciones indebidas sobre los alumnos. Entre otras: medidas disciplinarias por razones distintas a su comportamiento; retención de certificados de estudios y/o licencia; negar el acceso al establecimiento.
Integración de profesores y apoderados (5%)	- Actividades de información e integración en el establecimiento, medida a través de una encuesta - Opinión de los padres sobre calidad del recinto educacional, medida a través de encuesta aplicada con la prueba SIMCE

Fuente: Ministerio de Educación de Chile

Las sanciones pueden incluir el despido de maestros o la intervención o cierre de las escuelas cuyos alumnos obtienen sistemáticamente malos resultados. Pueden también incluir la denegación de recursos a los estados que no alcanzan ciertas metas, como en la experiencia en curso en los Estados Unidos (véase el recuadro 2).

Estas dos últimas "versiones" de la *"rendición de cuentas/responsabilidad por los resultados'* parten de una serie de premisas sobre los problemas de la gestión estatal. Los sistemas educativos en todo el mundo son mayoritariamente de gestión estatal, sea de tipo nacional/central, sea de tipo local (provincial, municipal, distrital). Uno de los problemas que esto acarrearía es que las instituciones del estado tenderían a responder más a los intereses de los funcionarios que a los objetivos del servicio. Por ejemplo, en la asignación de horas y

FICHA 11

Recuadro 2

La Ley "Ningún Niño Dejado Atrás" de la Administración Bush

"Haciendo de la 'rendición de cuentas/ responsabilidad por los resultados' la pieza central de la agenda educativa, el presidente George W. Bush reforzó fuertemente lo que ya era un tema central en las políticas de los estados dirigidas a mejorar la educación".

- La Ley conocida como Ningún Niño Dejado Atrás ("No Child Left Behind" – NCLB) obliga a los estados a definir "objetivos anuales de progreso adecuados", en función de sus propias evaluaciones estatales.
- Establece también que el 100% de los alumnos debe alcanzar un nivel de "proficiencia" (suficiencia) o superior para el año 2014.
- Los "objetivos anuales de progreso" deben estar establecidos de modo tal de alcanzar dicha meta en el 2014.
- Las metas anuales deben estar definidas no solo para el conjunto de los alumnos, sino para categorías específicas de alumnos: grupos raciales o étnicos importantes, alumnos con necesidades especiales, alumnos con dominio limitado del idioma inglés.
- Por lo menos el 95% de los alumnos de cada uno de estos grupos debe participar en las evaluaciones estatales.
- Las metas deben ser evaluadas a nivel de cada centro educativo, y un centro que no logre alcanzarlas por dos años consecutivos debe ser calificado como "necesitado de mejora".
- Los estados que no alcancen sus metas arriesgan perder recursos federales para la educación.

Traducido y adaptado de LINN, R. et al. (2002). Accountability Systems: Implications of Requirements of the No Child Left Behind Act of 2001. En: Educational Researcher, Volumen 31, Nº 6. AERA, agosto/setiembre de 2002.

• "Los sistemas de 'rendición de cuentas/responsabilidad por los resultados' tienen el potencial de contribuir a mejorar la calidad de la educación. Para ello deben ser diseñados de manera consistente con la evidencia y la experiencia derivadas de la investigación. Sabemos bastante acerca de las características de estos sistemas que pueden tener influencias positivas en la educación, así como acerca de las características que han resultado contraproducentes. Los sistemas de 'rendición de cuentas /responsabilidad por los resultados' deben ampliar sus definiciones acerca de qué es lo que cuenta como evidencia de éxito. Si bien las metas ambiciosas son deseables, también deben estar realistamente apoyadas en la experiencia anterior".

La Ley NCLB (2001) incluye mucho de positivo. El énfasis en el logro por parte de todos los estudiantes y la atención especial dada a los grupos de estudiantes que han tenido el rendimiento más bajo es especialmente loable. Las expectativas altas también lo son. Las metas que NCLB establece para el logro de los estudiantes serían maravillosas si pudiesen ser alcanzadas pero, desafortunadamente, son demasiado irreales, tanto que son más aptas para desmoralizar a los educadores que para inspirarlos. Si las metas de progreso anual son reforzadas, ello va a tener como resultado que muchas escuelas que están dando grandes pasos en la enseñanza a sus alumnos reciban sanciones. Esto se debe a que los progresos firmes y significativos no son reconocidos como mejoramiento bajo la Ley NCLB si no se logran las "metas anuales de progreso".

Traducido de LINN, R., (2003). Accountability: Responsabilidad y Expectativas Razonables. En: Educational Researcher, Volumen 32, Nº 7. AERA, octubre de 2003.

cargos docentes se privilegia el interés del docente más calificado, en detrimento del derecho del alumno de medios más desfavorecidos de contar con los mejores docentes.

Un segundo problema de la gestión estatal es que no existirían incentivos para hacer las cosas bien. Hay pocos controles de calidad y ningún funcionario gana o pierde nada si los resultados son buenos o malos. El sistema no suele hacer diferencias entre los profesores buenos y los que no lo son. Los sistemas de evaluación y supervisión del desempeño tienen un carácter administrativo-burocrático, más que sustantivo.

Un tercer supuesto es que, la alta conducción de los sistemas de educación pública muchas veces responde más a intereses y lógicas de tipo político partidario que educativas.

De allí que se desee introducir mecanismos como los descritos, que modifiquen las lógicas burocráticas y presionen hacia el logro de los resultados esperados.

2.5. Versión 5: Pruebas con consecuencias para los alumnos

La forma más antigua de *"rendición de cuentas/responsabilidad por los resultados"* son las pruebas con consecuencias para los alumnos, como forma de motivarlos a aprender y como forma de motivar a las escuelas y maestros a que sus alumnos logren aprobar dichas pruebas.

De hecho, esta forma de "responsabilidad por los resultados" es la más extendida. En la mayor parte de los casos se basa en pruebas elaboradas a nivel local –y en general individual– por parte de los docentes. El alumno es quien "rinde cuentas" de lo que ha estudiado y aprendido.

Si bien, no debemos olvidar que una parte central de la responsabilidad por aprender recae sobre el propio alumno, uno de los problemas es que muchas veces el profesor se desliga de su responsabilidad. El éxito o fracaso es atribuido exclusivamente al alumno, aun cuando no siempre el profesor ha hecho todo lo posible para que el alumno aprenda.

El otro problema es que no siempre los profesores evalúan de manera apropiada y justa. Asimismo, en un país puede existir una gran heterogeneidad en los niveles de exigencia de las evaluaciones elaboradas a nivel local. Por eso muchos países optan por establecer exámenes nacionales al final de algunos niveles clave de la enseñanza, en especial al final de la educación media.

El caso más típico es la prueba de bachillerato en Francia, pero muchos otros países (Brasil, Dinamarca, Holanda, Irlanda, Italia, México, Nueva Zelanda, Reino Unido, República Dominicana y Costa Rica, por mencionar algunos) establecen exámenes externos al final de la primaria, del primer ciclo de educación media y/o de la secundaria, bajo distintas modalidades.

En algunos casos las pruebas son diseñadas centralmente pero corregidas a nivel local. En otros son corregidas a nivel central, y en otros son corregidas por equipos locales integrados por docentes de la escuela y docentes externos.

En algunos casos este tipo de pruebas son obligatorias, en tanto en otros son de carácter voluntario. Existen también sistemas en que las pruebas externas tienen un peso parcial en la calificación final del alumno y el resto de la misma depende de las evaluaciones internas que se propone y corrige en cada escuela[2].

2) Por más detalles véase VALVERDE, G. (2003). La política de evaluación y currículo ante el desafío de la calidad. En: Anexo al Informe de la Comisión de Desarrollo y Uso del Sistema de Medición de la Calidad de la Educación, La Experiencia Internacional en Sistemas de Medición: Estudio de Casos. Ministerio de Educación, República de Chile.

La idea central detrás de los sistemas nacionales de pruebas con consecuencias para los alumnos es que los mismos permiten:

a. garantizar una base común de conocimientos y capacidades en todos los alumnos que terminan los niveles clave del sistema educativo;
b. establecer un fuerte incentivo para que los alumnos aprendan y los profesores enseñen lo que será evaluado en dichas pruebas;
c. que cada alumno y cada familia sepa cuál fue el resultado de cada alumno en cada escuela.

3. Efectos perversos de los sistemas de "rendición de cuentas" o de consecuencias "fuertes" para pruebas estandarizadas

Los sistemas de "rendición de cuentas/responsabilidad por los resultados", en particular aquéllos que intentan establecer presiones directas sobre distritos, escuelas y maestros para que mejoren los resultados de sus alumnos, no siempre han traído consigo los beneficios prometidos. En rigor, la evidencia empírica existente no muestra que se produzcan mejoras sostenidas en el tiempo a partir de los mismos.

Martin Carnoy, de la Universidad de Stanford, ha desarrollado recientemente un amplio trabajo de investigación sobre el impacto de los sistemas de *"rendición de cuentas/ responsabilidad por los resultados"* que involucra a los 50 estados de los Estados Unidos.

Inicialmente, el estudio mostró que los estados con sistemas de más fuertes consecuencias lograban una mejora más importante en los aprendizajes, en especial en Matemática y en 8º grado, mucho menos importante en Lengua y en 4º grado. Sin embargo, al actualizar los datos dos años más tarde y analizar cómo habían continuado las tendencias, los autores concluyeron que:

"En consecuencia, estos nuevos resultados sugieren que no existe o solo existe una relación muy débil entre las políticas de responsabilidad por los resultados en los estados y los aumentos de los puntajes de las pruebas en dichos estados a comienzos de la década del 2000. Puede que nuestros primeros resultados con respecto a los aumentos de los puntajes de matemáticas en el período 1996-2000 informados en este documento constituyan una anomalía o simplemente puede que sea difícil sostener los aumentos de los puntajes de matemáticas aun existiendo una sólida responsabilidad por los resultados"[3].

3) CARNOY, M. & LOEB, S., (2004). ¿Tiene efectos la responsabilidad externa en los indicadores educacionales de los alumnos? Un análisis entre los estados de los EE.UU. PREAL, Documento Nº29, p. 4.

FICHA 11

Recuadro 3

El caso holandés

"Holanda representa un caso intrigante de especial interés como ejemplo de un país sin la larga tradición de evaluaciones externas centralizadas (como Francia) o descentralizadas (como Alemania). Su historia en evaluación es más bien reciente, y representa una paulatina preocupación por operacionalizar criterios de calidad educativa y por monitorear su cumplimiento.

"También se destaca por el arreglo institucional de la evaluación externa. Esta se encuentra en una institución autónoma [denominada CITO] formada por el Estado holandés, responsable ante el Ministerio de Educación, pero formando una institución aparte...

Un modelo mixto y voluntario de pruebas de egreso en educación primaria

"La prueba de final de primaria de CITO es completamente voluntaria. Cada escuela es libre de usar la prueba o no. En la actualidad CITO estima que el 83 por ciento de las escuelas primarias del país participan. Estas pruebas funcionan no únicamente para certificar la educación primaria, sino como prueba de selección para los tres tipos de educación secundaria que existen actualmente en Holanda...

"La prueba de educación básica es relativamente corta y no pretende evaluar todo el currículo –consta de cuatro secciones: aritmética, lenguaje, manejo de datos y estudios ambientales (las escuelas tienen libertad de omitir la sección de estudios ambientales). Cada sección consta de 60 reactivos de selección múltiple. CITO proporciona dos tipos de informe a las escuelas relativos a su participación: un informe acerca de la escuela, que compara el desempeño promedio de los estudiantes con el promedio de todas las escuelas que participan en la prueba, y otro informe por estudiante que también compara su desempeño con respecto a su escuela y con respecto a los promedios en "escuelas similares".

Un modelo mixto y obligatorio de pruebas de egreso de educación secundaria

"El sistema de pruebas de egreso de la educación secundaria es también mixto, pero no de carácter voluntario. La evaluación externa manejada por CITO representa un 50% de la nota final de cada materia, siendo la escuela responsable de determinar un sistema interno para adjudicar el 50% de la nota que le corresponde. Los temarios para las pruebas externas, aunque implementados por CITO, son responsabilidad del Ministerio de Educación. Las escuelas pueden evaluar sus propios objetivos (no necesariamente parte del temario del Ministerio)

Fuente: VALVERDE, G. (2003). La política de evaluación y currículo ante el desafío de la calidad. En: Anexo al Informe de la Comisión de Desarrollo y Uso del Sistema de Medición de la Calidad de la Educación, La Experiencia Internacional en Sistemas de Medición: Estudio de Casos. Ministerio de Educación, República de Chile.

en la evaluación que corresponde al 50% de evaluación interna...

"Las pruebas externas se hacen durante dos semanas en mayo, y existe una prueba por cada materia de la secundaria. Los reactivos suelen ser una mezcla de preguntas de selección y preguntas abiertas. Las preguntas abiertas son corregidas por dos docentes, uno es el docente del alumno y otro docente de otra escuela. Las preguntas abiertas suelen constituir un porcentaje mayor en las pruebas que se hacen en las secundarias académicas de preparación para la universidad, siendo menos frecuentes en las pruebas para las secundarias técnico-vocacionales. Las pruebas internas siguen el calendario que fija la escuela, sin necesidad de conformarse a un calendario nacional.

Estado actual de la discusión

"...existe una aceptación general de los beneficios de las evaluaciones externas, y gran confianza en sus resultados. Aun más notable, parece haber demanda por el tipo de información que ofrece por parte del público.

"La autonomía de CITO parece haber contribuido a la superación de la resistencia original a la evaluación externa. Al recibir informes confidenciales de un organismo autónomo, percibido como una institución de gran profesionalismo, el público y las escuelas holandesas parecen dispuestos a reconocer su legitimidad.

"Recientemente es aun más notable el cambio de actitud con respecto a la evaluación externa. Al principio hubo resistencia a las pruebas externas pero en la actualidad más bien hay duda acerca de la evaluación interna. En recientes debates se ha interpelado el 50 por ciento de la nota de secundaria que se asigna mediante evaluaciones internas. Se argumenta que las pruebas externas tiene referentes de calidad educativa (fijadas por el temario del Ministerio) claros, y el significado de los mismos está abierto al escrutinio público. En cambio las pruebas internas no son tan transparentes. De este modo, se ha sugerido recientemente que los informes de educación secundaria reporten por separado los resultados en cada evaluación – cosa que hasta hoy en día no sucede".

En cambio, muchas veces estos sistemas traen consigo efectos no deseados o "efectos perversos". Por tanto, los formuladores de políticas educativas deberían ser cuidadosos antes de establecer sistemas de consecuencias fuertes asociados a las pruebas de evaluación estandarizadas. A continuación se explica brevemente cuáles son los principales tipos de "efectos perversos" a considerar.

FICHA 11

3.1. El efecto "goma de mascar" en los sistemas de mercado competitivo

Los sistemas que buscan generar competencia entre las escuelas a partir de la publicación de rankings y de la apuesta a que los padres "castiguen" a las escuelas que no tienen buenos resultados retirando a sus hijos de ellas pueden tener efectos no deseados, incongruentes con sus propósitos.

Una escuela con buenos resultados tendría, como consecuencia de ello, una alta demanda en el siguiente ciclo lectivo. Como su capacidad para recibir alumnos es limitada, tendería a seleccionar a los mejores alumnos. De este modo el director se aseguraría que sus resultados sean aún mejores en la siguiente evaluación.

Por el contrario, las escuelas cuyos resultados no sean buenos, tenderían a perder a los alumnos hijos de las familias más educadas, dado que estas contarían con más información y mayores recursos para desplazar a sus hijos. La escuela se quedaría con los alumnos de las familias menos educadas, lo cual determinaría que sus resultados empeoren en la siguiente evaluación.

De este modo se generaría una dinámica de incremento o "estiramiento" de las distancias en los resultados entre los extremos del sistema educativo. De allí la denominación "efecto goma de mascar".

Este fenómeno fue constatado por Orlando Mella para el caso chileno, que ha apostado fuertemente por este tipo de esquema de responsabilidad por los resultados desde los años 80:

> *"La distancia en los resultados entre los alumnos y alumnas de establecimientos municipalizados y particulares pagados muestra una tendencia creciente, como se puede apreciar en el gráfico 6, desde 1,02 desviaciones estándares en 1990 hasta 1,8 desviaciones estándares de diferencia en el año 2001. El indicador elegido y el método de comparación nos permiten por tanto concluir que la brecha en calidad es creciente entre la educación que reciben los más pobres y los más ricos"*[4].

4). MELLA, O., (2003). "12 años de reforma educacional en Chile. Algunas consideraciones en torno a sus efectos para reducir la inequidad". En: REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 2003, Vol. 1, No. 1. http://www.ice.deusto.es/RINACE/reice/p_vol1num1.htm

3.2. Los sistemas de incentivos motivan a los que ganan y desmoralizan a quienes pierden

Todo sistema de incentivos tiene ganadores y perdedores. Lo que muestra la evidencia empírica es que quienes ganan los incentivos se motivan y tienden a querer seguir participando en el sistema y mejorar para asegurarse de volver a ganar. Lo mismo puede ocurrir con quienes no ganan pero quedan cerca de hacerlo.

Simultáneamente, se ha constatado también que quienes quedan lejos de ganar se desmoralizan, tienden a creer que nunca lograrán alcanzar los primeros lugares y a volverse indiferentes ante el sistema. Como consecuencia, continúa sin resolverse el problema de cómo mejorar las escuelas cuyos resultados son peores.

El efecto desmoralizador puede afectar incluso a escuelas que han hecho esfuerzos importantes y han logrado mejoras, pero no las suficientes como para alcanzar las metas estipuladas (véase el recuadro 2).

3.3. Entrenar a los alumnos para responder pruebas estandarizadas

Todo sistema de evaluación estandarizada tiene como efecto dirigir una parte de la preocupación y el tiempo de enseñanza hacia lo que las pruebas evalúan.

Este efecto es mayor cuanto más importantes sean las consecuencias de las pruebas para escuelas y docentes. En sí mismo, esto no es malo. En realidad es parte de los efectos buscados cuando se evalúa. El problema se plantea cuando se vinculan consecuencias fuertes (o sistemas de *"rendición de cuentas/responsabilidad por los resultados"*) a pruebas que no son lo suficientemente complejas y abarcativas. Por ejemplo, cuando se emplea exclusivamente pruebas de opción múltiple dado su menor costo para evaluaciones censales, que son las que requiere un sistema de *"rendición de cuentas/responsabilidad por los resultados"*.

En estos casos puede ocurrir que la "mejora" educativa se limite a entrenar a los alumnos para responder pruebas de elección múltiple.

Cuando las pruebas no tienen consecuencias fuertes, es más fácil que los docentes puedan considerarlas como lo que son –indicadores de lo que los alumnos están aprendiendo– y aprender de ellas sin estar presionados a mejorar los resultados de cualquier modo.

Si las pruebas han de tener consecuencias "fuertes", deben ser suficientemente amplias y complejas, con lo cual es necesario, a la hora de tomar decisiones, considerar los costos asociados al diseño y codificación de las mismas.

3.4. La tentación o necesidad de "rebajar" los niveles de exigencia

Otro problema importante, que es necesario analizar con cuidado antes de implementar un sistema de consecuencias fuertes vinculado a pruebas nacionales, es el siguiente:

- si los niveles de exigencia de las pruebas son muy altos, se corre el riesgo de generar un fracaso generalizado entre los estudiantes o entre las escuelas que, por diversas razones (costos financieros o sociales y políticos, entre otros), obligue a reducir los niveles de exigencia;
- si los niveles de exigencia se "rebajan" para evitar el fracaso mencionado, se corre el riesgo de adaptar las expectativas al nivel promedio de la población escolar y transmitir a los docentes una señal equivocada en cuanto a qué es lo que todos los alumnos deberían aprender.

Como ejemplo de este problema véase en el recuadro 4 lo que está generando en los Estados Unidos la Ley "Ningún Niño Dejado Atrás", de la Administración Bush.

Otro ejemplo de este problema es el siguiente. En las pruebas estandarizadas de Matemática aplicadas en Uruguay en 6º grado de Primaria en 1996, apenas el 34,6% de los estudiantes alcanzó el nivel definido como de "suficiencia".

Si esa prueba hubiese tenido consecuencias de acreditación para los alumnos, el 65% debería haber reprobado el curso, lo cual hubiese sido social y políticamente insostenible.

Las pruebas externas, si son exigentes y tienen consecuencias para los alumnos, casi seguramente generarán, al menos en una primera etapa, incrementos en los niveles de reprobación y deserción.

Una alternativa para evitar este problema es experimentar con sistemas mixtos, en los que la prueba externa tiene un peso parcial en la calificación del alumno, como el caso holandés presentado en el recuadro 3.

Recuadro 4

"Los estados están rebajando las exigencias de sus pruebas para evitar sanciones"

"El dispositivo de seguridad era hermético cuando los miembros del Consejo Estatal de Educación de Tejas recibieron los resultados de la prueba piloto del nuevo examen estatal de logro educativo. Los guardias estaban en la puerta cerrada de la sala de reuniones y los miembros del Consejo debieron firmar un compromiso de confidencialidad, lo que refleja lo sensible de la situación.

"Los resultados fueron horribles' dijo uno de los miembros del Consejo'. A pocos estudiantes les fue bien. Muchos tuvieron casi ninguna respuesta correcta".

Temiendo que miles de estudiantes perderían el nuevo examen y bajarían sus calificaciones, y que cientos de escuelas podrían enfrentar sanciones bajo la ley federal "Ningún Niño Dejado Atrás", el Consejo votó reducir la cantidad de preguntas que los estudiantes debían responder correctamente para pasar el examen, de 24 a 20 sobre un total de 36, para Lectura de tercer grado.

Tejas no ha estado solo al rebajar sus estándares en las pruebas en los últimos meses. Los educadores en otros estados han estado tomando decisiones similares, buscando evitar las sanciones que la ley federal impone a las escuelas cuyos estudiantes tienen desempeño pobre en pruebas estandarizadas. Desde que el presidente Bush firmó la ley en enero de 2002, los 50 estados han presentado planes para cumplirla. Pero algunos expertos dicen que es sólo una apariencia de aceptación. Silenciosamente, dicen, los estados están haciendo lo mejor que pueden para evitar costosas sanciones.

Los estándares de Michigan han estado entre los más altos de la nación, lo que generó problemas el año pasado cuando 1.513 escuelas fueron etiquetadas bajo la ley como necesitando mejoramiento, más que en ningún otro estado. Por tanto, los funcionarios de Michigan bajaron el porcentaje de estudiantes que deben pasar la prueba estatal para certificar a una escuela como progresando adecuadamente, de 75% a 42% en las pruebas de Lengua, por ejemplo. Esto redujo el número de escuelas antes mencionadas a 216.

Colorado empleó otra táctica que resultará en menor cantidad de escuelas etiquetadas como necesitando mejoramiento. Modificó el sistema de calificación de las pruebas, juntando a los estudiantes que antes eran categorizados como 'parcialmente proficientes' a partir de sus puntajes en la prueba, con los 'proficientes'.

Fuente: Traducido por el autor, Sam Dillon, THE NEW YORK TIMES, 20 de Mayo de 2003.

FICHA 11

(...)

Bajo la ley, los estados que no cumplan arriesgan perder recursos federales para la educación. Las escuelas que fracasan varios años seguidos deben ofrecer tutoría a los alumnos de bajo rendimiento y, eventualmente, pueden ser forzadas a una reorganización total. Pero la ley deja en manos de los estados establecer sus propios estándares de éxito.

(...)

Algunos expertos también critican a la ley por requerir a los estados que lleven al 100% de sus estudiantes al nivel de proficiencia en Lectura y Matemática para el año 2014, un nivel que dicen nunca ha sido logrado en ningún estado o país.

'Las severas sanciones pueden obstaculizar el logro de la excelencia educacional', dice Robert L. Linn, profesor en la Universidad de Colorado, que ha sido el último presidente de la Asociación Americana de Investigación Educativa, "porque implícitamente empujan a los estados a diluir sus estándares de contenidos y de desempeño para reducir el riesgo de sanciones".

(...)

Richard F. Elmore, un profesor de Educación de Harvard, escribiendo en el número de primavera de la revista Education Next, llamó a la ley 'la más grande y dañina expansión en la historia del poder federal sobre el sistema educativo de la nación".

Sistemas de este tipo, como los que usa República Dominicana y próximamente El Salvador, tienen la ventaja de atenuar el impacto inicial de generación de fracasos que la prueba externa puede tener, al tiempo que permiten entregar señales fuertes a las escuelas acerca de qué es lo que todos los alumnos deberían aprender.

3.5. Aliento a conductas de corrupción

Finalmente, es necesario mencionar que todo sistema de consecuencias fuertes tiene como efecto altamente probable el de generar conductas reñidas con la ética profesional, tales como dejar a los alumnos más retrasados fuera de la prueba, ayudar a los alumnos en la prueba, etc., lo cual a su vez obliga a establecer sistemas de vigilancia y control que muchas veces contribuyen a generar un clima hostil hacia la evaluación.

4. Aspectos de concepción educativa que deben ser tenidos en cuenta

Además de los "efectos perversos" que es necesario analizar con cuidado antes de poner en práctica un sistema de *"rendición de cuentas/responsabilidad por los resultados"*, hay cuestiones conceptuales, relacionadas con las características específicas de la labor educativa, que es importante considerar.

4.1. La atribución de causa

Sam Messick, destacado especialista en evaluación, ha llamado la atención acerca de que la atribución de causa en educación es un asunto complejo. Los resultados que alcanza un alumno en una prueba dependen de múltiples factores.

Por tanto, *"es injusto hacer que alguien rinda cuentas por algo sobre lo cual no tiene control o responsabilidad.... Debemos preguntarnos, cuando un alumno fracasa, si es una falla de él, si es una falla de su profesor, si es una falla del sistema. Cuando un sistema complejo funciona pobremente, el dedo acusador suele apuntar hacia el eslabón más débil del sistema"*[5].

Las partes más débiles del sistema suelen ser, en primer lugar, los alumnos y, en segundo lugar, los docentes.

4.2. El libre albedrío de los alumnos

Un aspecto crucial a considerar, que hace de la educación una actividad radicalmente diferente de otras actividades humanas, es que la producción de los efectos buscados no depende exclusivamente de lo que haga el profesor o la escuela. En el acto de educarse está siempre de por medio el libre albedrío de cada individuo.

Algunos individuos pueden **no querer** realizar el esfuerzo que inevitablemente implica el aprender (otros pueden no estar en condiciones de realizar ese esfuerzo). Es evidente que en el presente la actividad de estudiar encuentra múltiples competidores en el entorno de los estudiantes, como la televisión e Internet.

5) MESSICK, S., (1999), "Key Issues". En: OTTOBRE, F., 1999 (editor); The Role of Measurement and Evaluation in Education Policy. UNESCO, París, Educational studies and documentes N° 69.

Este problema ocurre también en otras áreas, como la salud y la psicología.
No es común evaluar o remunerar a los médicos y psicólogos en función de si los pacientes se curan. El tratamiento prescrito por el médico puede ser enteramente apropiado, pero el paciente puede no cumplirlo y por tanto no curarse. La responsabilidad en este caso no es del médico, sino del paciente.

Lo mismo ocurre muchas veces en la educación. Los resultados no dependen únicamente de lo que el profesor haga, sino de la voluntad de aprender del niño o joven y del apoyo de la familia para que esa voluntad exista.

Por esta razón, en profesiones como la Medicina, lo que se evalúa es la buena o mala "praxis" del profesional. Del mismo modo, en educación, debería evaluarse a escuelas y docentes en función de lo que hacen con los recursos de que disponen más que en función de los resultados de sus alumnos.

Esto no implica ignorar los resultados de los alumnos, sino tomarlos como lo que son, un indicador a tener en cuenta. Si los resultados son pésimos, es un indicador de que algo puede andar mal en la escuela, del mismo modo que un médico que tiene una alta proporción de pacientes que no se cura deberá ser observado con especial atención.

4.3. Las condiciones para asumir la responsabilidad por los resultados

Un tercer elemento clave a considerar, siguiendo nuevamente el razonamiento de Sam Messick, es que para exigir responsabilidad a alguien por los resultados de lo que hace, esa persona tiene que tener los elementos básicos razonables para realizar la tarea.

De allí que se deba tener cuidado cuando se trasladan ideas o debates propios de la política educativa del primer mundo, donde las condiciones de ejercicio de la profesión docente son unas, al contexto latinoamericano, donde las condiciones son otras.

Por ejemplo, en Finlandia, el país con los mejores resultados en PISA, se invierte un 6% del PBI en educación, que además es un PBI tres o cuatro veces superior al promedio de América Latina. Un profesor gana entre 25.000 y 30.000 dólares al año y atiende a un promedio de 20 alumnos por grupo. Estas condiciones no son comparables a la precariedad que caracteriza al ejercicio de la profesión docente en América Latina en todos los sentidos:

6) *Hamilton, L. y otros, 2002; Making Sense of Test-Based Accountability in Education. California, Rand Education, pp. 117-118.*

salarios, preparación para el desempeño de la función, condiciones de vida de los alumnos, condiciones materiales de las escuelas, etc.

A nadie se la ocurriría "pedirle cuentas" a un médico que acaba de operar a un herido en una tienda de campaña, usando un cuchillo de cocina, sin los elementos mínimos imprescindibles y que, además, tal vez ni siquiera es cirujano. Del mismo modo, antes de pretender que las escuelas y docentes "rindan cuentas" de los resultados de su trabajo, deberíamos asegurarnos de que tienen los medios necesarios para realizar dicho trabajo adecuadamente.

No se puede plantear el problema de la responsabilidad por los resultados desligado del problema de los medios para hacerse cargo de esa responsabilidad: formación, oportunidades para aprender a enseñar mejor, orientación y apoyo, condiciones y materiales trabajo, condiciones de los alumnos, condiciones salariales y carga horaria, etc.

"La rendición de cuentas es un asunto de ida y vuelta: las escuelas no podrán cumplir con los estándares que la comunidad les fije si la comunidad misma no cumple con su obligación de apoyar adecuadamente a las escuelas"[6].

En este punto un aspecto central a destacar es el relativo a las oportunidades de formación continua para los docentes. De nada servirá imponer sanciones y exigir cumplimiento de las obligaciones funcionales si no se brinda a los docentes orientación técnica apropiada y oportunidades de capacitación, sin las cuales no hay mejora posible, por más incentivos que se establezcan.

FICHA 11

Síntesis final

Esta Ficha ha abordado un tema sin duda delicado y complejo. Como trasfondo del tema de la *"rendición de cuentas/responsabilidad por los resultados"* existen dos lógicas opuestas con las que el mismo puede ser encarado.

La rendición de cuentas, como se dijo antes, implica que alguien se haga responsable por el hecho de que los alumnos aprendan –o por la ausencia de aprendizaje– Lo que se busca evitar es que el sistema educativo funcione como un sistema sin responsables, que no rinde cuentas ante nadie acerca de cómo utiliza sus recursos y qué resultados obtiene.

Pero la pregunta central es: **¿quién debe hacerse responsable y cómo?**

Un modo de responderla, que denominaremos **"lógica de enfrentamiento",** consiste en que cada actor involucrado busca responsabilizar a otro.

El docente responsabiliza al alumno porque no estudia, a la familia porque no apoya y al "sistema" o a las autoridades porque no le dan los medios necesarios para desarrollar su labor. La familia responsabiliza al docente y a la escuela porque no enseñan o "no exigen como antes". Las autoridades, a través de los sistemas de *"rendición de cuentas/responsabilidad por los resultados"*, buscan responsabilizar a las escuelas y docentes y ejercer presión sobre ellos. Las autoridades centrales responsabilizan a las autoridades locales, y viceversa. Los funcionarios políticos -–que están al frente del sistema por un período limitado– responsabilizan a los funcionarios de carrera –que forman parte del aparato técnico y burocrático permanente– y viceversa.

El resultado de esta lógica es una suerte de guerra de todos contra todos en la que, paradójicamente, todos se desligan de la responsabilidad y la atribuyen a otros. Por este camino, los sistemas educativos no cambian ni mejoran.

La alternativa es construir una **"lógica de colaboración"** entre los actores.

Para hacerlo es necesario comenzar por comprender la complejidad del acto de educar.

El aprendizaje no es el resultado directo y automático de lo que hace el docente. Su labor es esencial, obviamente. Pero el hecho de que un alumno aprenda depende, además, de su acumulación cultural previa. Depende de su motivación para aprender que, en parte, debe construir el docente, pero que, además, está determinada por valores sociales (¿tiene sentido estudiar?) y por la importancia que la familia otorga a la educación. Depende del apoyo concreto que la familia pueda ofrecer al niño o adolescente en el proceso de aprendizaje. Depende del ambiente en su centro educativo y del ambiente social y cultural en relación a la educación.

Por tanto, la educación debería ser percibida como un esfuerzo colectivo, en que múltiples actores deben hacerse responsables de mejorarla:

- la sociedad en general, valorando la educación y a sus profesionales, lo cual debería verse reflejado en la consideración social hacia la profesión y en los recursos que se destinan a la educación;
- el sistema político, designando al frente de la educación a personas competentes y dando continuidad a las políticas educativas por encima de los intereses partidarios;
- las autoridades, realizando su gestión con compromiso y competencia técnica, dando cuentas públicamente de las políticas que impulsan, apoyando de manera efectiva la labor de los docentes y mejorando las condiciones en que ejercen su función;
- las escuelas y docentes, buscando los mejores modos de enseñar, actualizándose en forma permanente y preocupándose por el aprendizaje de cada alumno;
- las familias, apoyando la labor escolar de sus hijos y la labor de escuelas y docentes (más que amenazando con retirar a sus hijos de la escuela);
- los propios alumnos, realizando los esfuerzos ineludibles que requiere el aprender.

Por tanto, ***un sistema integral de "rendición de cuentas/responsabilidad por los resultados"*** que no se apoye únicamente en los eslabones más débiles de la cadena **debería, en primer término, construir una visión de la educación como emprendimiento nacional, en el que todos son responsables** de algo, en lugar de buscar responsables a los cuales señalar para premiar o castigar.

Debería, además, incluir los siguientes elementos:

1. dirección, metas claras y ambiciosas para todos;
2. apoyo a los docentes y escuelas para realizar mejor su trabajo (orientación; oportunidades para aprender; espacios de discusión y formación, oportunidades para la construcción y ensayo de alternativas; mejoras salariales; dignificación social y reconocimiento profesional);
3. información a las familias sobre lo que sus hijos deben aprender, información sobre lo que realmente están aprendiendo y orientaciones en relación a lo que ellas deben hacer para apoyar el aprendizaje de sus hijos;

FICHA 11

4. información a los alumnos sobre lo que se espera de ellos, buenas evaluaciones, orientación sobre cómo mejorar, un ambiente de aprendizaje;
5. evaluación rigurosa del cumplimiento de las responsabilidades funcionales básicas por parte de los docentes, con consecuencias fuertes;
6. evaluación del desempeño profesional docente, con fines formativos y con consecuencias de crecimiento en la carrera;
7. información sistemática sobre lo que los alumnos aprenden, así como sobre lo que los alumnos opinan acerca de la educación que están recibiendo.

¿CÓMO USAR LAS EVALUACIONES PARA MEJORAR AL EDUCACIÓN?

Los resultados como herramienta de aprendizaje y mejora

Como complemento de las Fichass 10 y 11, la Fichas 12 tiene como objetivo mostrar enfoques para el uso de los resultados de las evaluaciones estandarizadas alternativos a los *rankings* y consecuencias "fuertes".

En esta Fichas se busca mostrar experiencias dirigidas a utilizar los resultados de manera tal que contribuyan a mejorar el trabajo de escuelas y docentes y los logros educativos de los estudiantes a través del aprendizaje y el desarrollo profesional.

Para ello se vuelve sobre el tema de los usos formativos de las evaluaciones y el papel que la información resultante de ellas puede tener para cuatro audiencias principales –educadores, familias, autoridades y opinión pública– en la medida en que la información se entregue de manera adecuada, evitando las distorsiones y errores analizados en Fichass anteriores.

A modo de introducción al tema de esta Fichas, el recuadro 1 pretende destacar la importancia de los usos formativos de las evaluaciones, desde la visión del destacado especialista chileno Ernesto Schiefelbein.

1. Un enfoque alternativo en el uso de los resultados: Just For The Kids

En el Estado de Tejas, en los Estados Unidos, funciona desde 1995 una organización no gubernamental y sin fines de lucro denominada *Just for the Kids*[1].

Su objetivo es colaborar en la mejora de la educación a través del uso intensivo de los resultados de las evaluaciones para mejorar la enseñanza.

1) El sitio web de la organización es http://www.just4kids.org. El nombre de la organización juega con una doble acepción de la palabra "justo" como adjetivo vinculado al concepto de justicia o equidad, y como adverbio vinculado a las ideas de pertinencia, exactitud y/o exclusividad "sólo para los niños".

Recuadro 1

Uso de los resultados de las evaluaciones: ¿castigar o ayudar?

"Desde fines de los 60 se ha medido el nivel de aprendizaje de los alumnos para elevarlo, pero no ha sido evidente la forma de hacerlo y, de hecho, no ha aumentado. La prueba de 8° grado (1967-1971) se limitó a mostrar a los profesores los tipos de habilidades que debían estimular en sus alumnos. La PER, creada a principios de los '80, buscó determinar el nivel del servicio educativo que se ofrecía y "monitorearlo" con ayuda de los padres, que tratarían de seleccionar las mejores escuelas para sus hijos. El SIMCE midió, a partir de 1988, los rendimientos de las escuelas y trató de identificar factores que pudieran explicar las diferencias y evaluar el impacto de los programas de los municipios y del Mineduc. Sin embargo, no hay avances y conviene repensar este poderoso instrumento, que es el SIMCE, si se quiere mejorar los aprendizajes. Un rol alternativo del SIMCE sería, según otros expertos, ayudar a los profesores a enseñar mejor y revisar cuáles son los temas importantes del currículo, más que el "clasificar" rendimientos. Este segundo rol plantea diferencias nítidas con respecto al primero: ¿Seleccionar o desarrollar? ¿Poner nota o medir logro de meta? ¿Conocer el ranking o el aprendizaje de objetivos específicos (criterio)? ¿Castigar o ayudar?

"Son dos funciones diferentes (los expertos hablan de Evaluación Sumativa y Formativa, respectivamente). Si la primera no elevó hasta ahora los rendimientos, convendría probar la segunda. Esto implica fuertes cambios en el SIMCE. En efecto, para "seleccionar" o "castigar" basta calcular un puntaje total que "discrimine entre buenos y malos" y difundirlo públicamente. En este rol es fundamental no divulgar los ítemes utilizados para que mantengan su poder de discriminación (no ser conocidos por los profesores y evitar que preparen a sus alumnos para contestarlos mecánicamente). Basta calcular el puntaje de cada escuela para que los padres seleccionen entre escuelas o preparar datos e indicadores globales para que los funcionarios del Mineduc modifiquen las estrategias.

"En la segunda función, en cambio, el usuario de la información del SIMCE es el profesor en su sala de clases. En efecto, para "desarrollar" o "ayudar a mejorar" se necesita entregar información detallada a cada profesor sobre los aspectos en que cada uno de sus alumnos logró los niveles adecuados y aquellas habilidades o conocimientos que todavía no posee o no domina suficientemente. El profesor debe saber lo que contestó cada alumno en la prueba y revisar con cada uno los errores que cometió, hasta que el alumno internalice las deficiencias y reforme adecuadamente sus procesos de pensamiento. Sin información detallada el profesor no puede identificar los aspectos de su enseñanza que debe cambiar o cómo ayudar a cada estudiante...

"En resumen, una redefinición clara del objetivo de la prueba nacional de medición de la calidad de la educación debe orientar la reflexión del grupo de expertos sobre los cambios que debe tener el SIMCE..."

Fuente: Ernesto Schiefelbein, "SIMCE, ¿CASTIGAR O AYUDAR?"; publicado por el Diario La Tercera de Santiago de Chile, junio de 2003.

Para ello, la organización publica en *Internet* los resultados de las escuelas en los exámenes estatales de una manera interesante, no bajo la forma de rankings, sino ofreciendo una visión de la evolución de cada escuela e introduciendo, como punto de comparación, el desempeño de los alumnos en las mejores escuelas de medios desfavorecidos.

Dicho en sus propias palabras, "la idea es simple: ayudar a las escuelas a mejorar mostrándoles cómo otras escuelas con similar alumnado logran altos niveles de desempeño"[2]. Para ello, esta organización se dedica –además de a publicar la información en internet– a la formación en servicio de maestros, a entrenar a los equipos de conducción escolar en relación a las "mejores prácticas" y a promover el compromiso público con el mejoramiento de las escuelas públicas. También realiza investigación.

Un dato relevante que el lector debe tener en cuenta –y que hace posible esta experiencia– es que el Estado de Tejas tiene un sistema estatal de pruebas de tipo censal y criterial, es decir, que involucra a todas las escuelas y que establece categorías definidas en relación a cuáles son los niveles de desempeño aceptables para los alumnos. Estas pruebas se realizan todos los años.

Las Figuras 1 y 2 ilustran el modo en que esta organización presenta en Internet los resultados de los centros educativos.

La Figura 1 incluye, en la parte inferior de cada barra, los porcentajes de alumnos con resultado "proficiente" (satisfactorio) en la prueba estatal de Tejas (denominada TAAS).

En la parte superior de cada barra se indica el porcentaje de alumnos que simplemente "pasaron" (aprobaron la prueba) pero sin alcanzar un nivel satisfactorio.

La primera barra del gráfico corresponde a una escuela en particular, la segunda incluye sólo a los alumnos que asistieron a la escuela en forma permanente, (definidos como aquellos que tenían por lo menos 2,8 años en la misma escuela), en tanto la tercera barra corresponde a los resultados de las mejores escuelas comparables con la escuela en cuestión.

Se define como "comparables" a aquellas escuelas que tienen igual o mayor proporción de alumnos provenientes de familias desfavorecidas o con dominio limitado del idioma inglés.

2) Véase JUST FOR THE KIDS, 2001; Promising Practices. How High-Performing Schools in Texas get Results. Austin, JUST FOR THE KIDS.

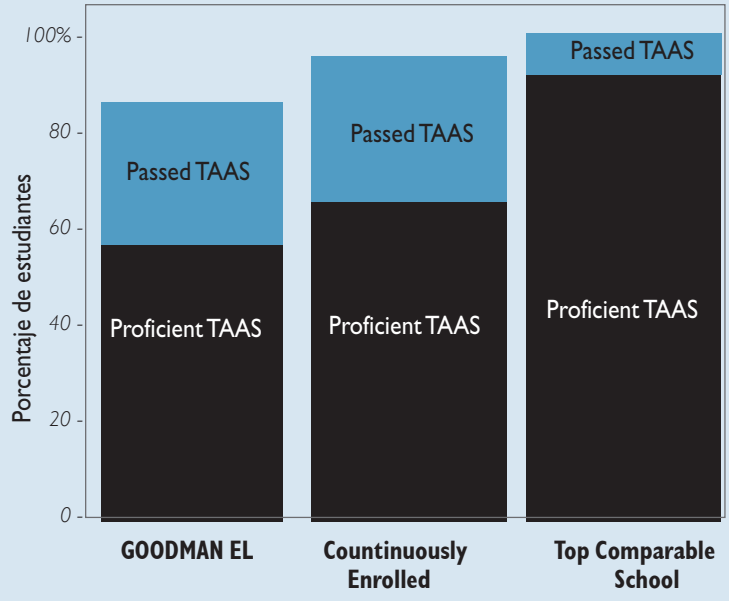
De este modo, la escuela y las familias pueden tener una idea del potencial de mejora de la escuela de sus hijos.

La Figura 2 ofrece una mirada de la evolución a lo largo de los años (eje x) del porcentaje de alumnos que lograron un desempeño satisfactorio (eje y), en una escuela en particular (la línea de abajo) y en el conjunto de las escuelas "comparables" que obtuvieron los mejores resultados (la línea de arriba). El gráfico considera únicamente a los alumnos matriculados en forma estable en la escuela.

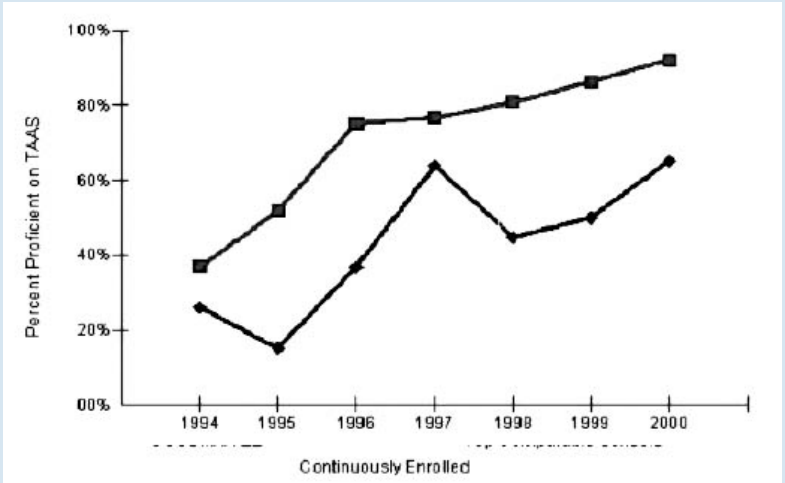
Esto permite a familias, educadores y autoridades hacer seguimiento del impacto de las iniciativas y proyectos emprendidos por las escuelas. La información se entrega para cada grado y para cada asignatura.

Figura 1 **Los resultados de una escuela y las mejores "comparables"**

Fuente: Just For The Kids, www.just4kids.org.



La evolución en el tiempo de los resultados de la escuela **Figura 2**



Fuente: Just For The Kids, www.just4kids.org.

También se publica en Internet una Fichas de cada escuela y de las mejores, en las que se indican los datos demográficos básicos que deben ser tenidos en cuenta en las comparaciones: qué porcentaje de los alumnos pertenecen a hogares con bajos ingresos; qué porcentaje de los alumnos pertenecen a grupos de inmigrantes que tienen dominio limitado del inglés; que porcentaje representan los alumnos que rindieron la prueba sobre el total de la matrícula; qué proporción de la matrícula total representan los alumnos que rindieron la prueba y son alumnos estables en la escuela.

2. ¿Qué caracteriza a escuelas con buenos resultados en sectores pobres?

Un trabajo de investigación realizado por *Just For Kids* sobre qué caracteriza a las escuelas que atienden poblaciones desfavorecidas (familias de bajos ingresos o pertenecientes a minorías étnicas) y tienen los mejores resultados, muestra que detrás de dichos resultados existen 5 elementos principales:

2.**1. Tomar la iniciativa: no dar excusas sino esforzarse por lograr que los alumnos aprendan.**

Esto significa que las escuelas se proponen metas ambiciosas y creen que todos sus alumnos pueden aprender. Los directores de estas escuelas visitan continuamente las aulas y dialogan con los maestros en relación a cómo mejorar la enseñanza.

2.2. Desarrollar y llevar adelante una estrategia clara para mejorar.

Se establece metas para cada grado y asignatura y se crea un currículo coherente para toda la escuela; se adopta prácticas de enseñanza sustentadas en investigación empírica; se reestructura la jornada o el calendario escolar para que los docentes dispongan del tiempo necesario para planificar en forma conjunta y compartir y analizar sus prácticas; se le da a los maestros el apoyo, los materiales y la capacitación que necesitan.

2.3. Evaluar continuamente el progreso de los alumnos e intervenir en forma inmediata cuando los alumnos o los maestros están en problemas.

Las evaluaciones en estas escuelas no son disparos al azar, sino que son cuidadosamente seleccionadas y aplicadas regularmente, tanto de manera formal como informal. Los maestros intervienen en forma inmediata para brindar apoyo específico a los alumnos que muestran problemas, incluyendo tutorías fuera del horario o calendario regular. También se interviene en forma inmediata cuando un maestro tiene problemas de orden conceptual o en el manejo de la clase.

2.4. Hacer que la prioridad número uno sea la enseñanza de alta calidad y la práctica basada en la investigación.

El desarrollo profesional está centrado en las prioridades de enseñanza de la escuela y la capacitación del personal es la estrategia para lograr las metas de aprendizaje. Se contrata a especialistas para observar y orientar a los docentes. Los maestros nuevos son colocados al lado de un maestro experimentado para que lo apoye en la planificación de sus clases.

2.5. Colaborar dentro y fuera de la escuela. Los maestros trabajan en forma conjunta, como grupo de resolución de problemas.

Los maestros de un mismo grado tienen momentos para planificar en común. El lenguaje escrito y la matemática son preocupación de los docentes de todas las áreas. Varias veces al año los docentes se reúnen por área o asignatura para planificar y coordinar en forma vertical (se refiere a los maestros de grados sucesivos) el trabajo de toda la escuela.

Otra investigación, dirigida a analizar el impacto del uso de estos resultados por parte de las escuelas, muestra que las escuelas que realizaron al menos uno de los siguientes tres pasos: –estudiar escuelas de muy buenos resultados, establecer metas de aprendizaje claras, realizar cambios en el programa escolar o en el desarrollo profesional de los maestros– mejoraron entre 3 y 8 puntos porcentuales la "proficiencia" en lectura, escritura y matemática en los primeros grados[3].

La experiencia de Just For The Kids es ilustrativa de varias de las reflexiones formuladas en Fichass anteriores, en particular de tres aspectos esenciales:

1. la importancia de generar y divulgar ampliamente información sobre los aprendizajes de los alumnos;
2. la importancia de hacerlo bajo un enfoque criterial;
3. la importancia de enfocar los esfuerzos hacia la creación de capacidades para enseñar, más que hacia los incentivos externos o la creación de un mercado competitivo.

Esto último incluye un aspecto de fundamental importancia que muestra la investigación realizada por esta organización. No basta con distribuir o publicar resultados; es imprescindible generar los espacios de trabajo e intercambio entre docentes para que los mismos sean analizados. La mejor publicación impresa nunca podrá sustituir el trabajo de análisis colectivo de los maestros. Si estos espacios no existen, difícilmente las evaluaciones contribuirán a mejorar la enseñanza.

3) Dougherty, C. & Collins, S. (2002); Use of the Just for the Kids Data By Texas Elementary Schools. Austin, Just for The Kids.

Recuadro 2

La experiencia del Territorio Capital de Australia

"El Territorio Capital de Australia empezó a aplicar pruebas de lenguaje y matemáticas en tercer, quinto, séptimo y noveno años en todas las escuelas gubernamentales entre 1997 y 1999. Las evaluaciones se basan en los resultados estipulados en los marcos del perfil curricular del ACT. Las capacidades de lectura, escritura y matemáticas son vistos como las piezas fundamentales de una educación exitosa, y como vitales para las oportunidades futuras de los estudiantes.

"El Programa está basado en un modelo referido a criterios y reporta la proporción de estudiantes que ha demostrado las capacidades documentadas en el perfil y, en tercer y quinto años, en los estándares nacionales.

"En el presupuesto de 1995, el gobierno del ACT anunció que se introducirían evaluaciones en dos grados de las escuelas primarias. Un grupo de referencia constituido por miembros del Foro de Consejos Escolares, el Consejo del ACT de Padres de Familia y Apoderados y de Asociaciones Ciudadanas, de la Asociación de Directores de Primaria, la Asociación de Directores de Secundaria, el Sindicato Educativo Australiano y del Departamento de Educación y Capacitación asesoró al gobierno con respecto a:

- el instrumento externo más apropiado disponible nacionalmente;
- los grados en los cuales se conducirían las evaluaciones;
- el formato para reportar datos a los padres y apoderados y al gobierno;
- arreglos para que los padres y apoderados pudieran optar por incluir o no a sus niños;
- el uso apropiado de los datos, tomando en consideración consideraciones de privacidad;
- establecer la relación entre reportar los resultados estudiantiles y los objetivos sociales de la escolaridad.

"El grupo de referencia examinó todos los instrumentos de evaluación disponibles en Australia que eran consistentes con los marcos curriculares del ACT y con el desarrollo curricular realizado en las escuelas.

"Se hizo un esfuerzo considerable por mantener una comunicación abierta con las escuelas y con la comunidad. Se mantuvo contacto regular con docentes, consejos escolares y

organizaciones de padres de familia y ciudadanas para informarles sobre el programa y obtener su apoyo desde antes de su implementación. Se usaron notas de prensa y otros medios para informar a la comunidad en general y mantener actitudes positivas de la comunidad hacia el programa.

"El grupo de referencia consideró que la privacidad de los estudiantes, docentes y escuelas individuales era un factor crítico para el éxito del programa en un sistema tan pequeño. La publicación de los resultados de las escuelas podría dar información equívoca sobre la calidad de la enseñanza y los programas escolares. Los resultados de una escuela en particular están determinados por una gama de factores que no están bajo el control de escuelas o maestros individuales. En una jurisdicción pequeña, la publicación de los resultados de las escuelas puede desestabilizar las matrículas, lo que podría generar problemas de recursos. Intentos de comparar a estudiantes individuales podría ir en detrimento de su desarrollo social y académico. Por lo tanto, se decidió que los datos no se usarían para:

- comparar individuos;
- evaluar el desempeño docente;
- comparar a maestros individuales;
- hacer comparaciones positivas o negativas sobre el desempeño de las escuelas;
- comparar escuelas gubernamentales y no gubernamentales.

"La confidencialidad en relación a este asunto estaba cubierta por la Ley de Privacidad de 1988, y por los Estándares de Gestión Pública del ACT. Se reforzó esa confidencialidad mediante el desarrollo de un Protocolo de Confidencialidad que todo individuo que tuvo acceso a los datos tuvo que firmar. Los reportes sobre las escuelas son conocidos sólo por las mismas escuelas".

Fuente: Margaret Forster, M., Masters, G. & Rowe, K., 2001; Measuring Learning Outcomes: Options and Challenges in Evaluation and Performance Monitoring. Material preparado para el Curso "Opciones Estratégicas para la Reforma Educativa" del Instituto del Banco Mundial. Australian Council for Educational Research (ACER) / World Bank Institute (WBI).

Recuadro 3

Pauta para el análisis de resultados en las escuelas

Anualmente, las escuelas tienen que elaborar respuestas a las siguientes preguntas y adjuntarlas a sus Planes de Lenguaje y Matemáticas.

1. En Lectura y Escritura, y en Sentido Numérico, de Datos y Sentido Espacial, ¿qué aspectos del currículo han tenido mayor impacto en el hecho de que su escuela haya tenido un desempeño promedio, por encima o por debajo del promedio del sistema en algunos ítems específicos de la evaluación?

   ¿Cómo está la escuela preocupándose por el desempeño escolar en

   - el nivel de grado

   - el nivel de toda la escuela?

2. Si observa a los estudiantes de su escuela que están en el 20% inferior del ACT, ¿cuántos estudiantes hay en el 5%, 10% y 15% inferior?

   Describa las intervenciones planificadas que han sido identificadas para mejorar el desempeño de estos estudiantes.

   ¿Cómo evaluarán los resultados de estas intervenciones, antes de la siguiente evaluación sistémica de los estudiantes?

3. De acuerdo con los datos de 1999, ¿cuánto crecimiento/declinación han experimentado los alumnos de quinto año desde que estaban en el tercer año?

   ¿Qué factores han causado este progreso o declinación?

4. ¿Cuán importante es la influencia del desempeño de los estudiantes de subgrupos por género, orígen lingüístico no inglés u origen indígena sobre los resultados de su escuela?

Especifique las intervenciones que se están realizando para mejorar los desempeños de cada subgrupo (v.g., Programas Individuales de Aprendizaje para Estudiantes Indígenas).

Fuente: Margaret Forster, M., Masters, G. & Rowe, K., 2001; Measuring Learning Outcomes: Options and Challenges in Evaluation and Performance Monitoring. Material preparado para el Curso "Opciones Estratégicas para la Reforma Educativa" del Instituto del Banco Mundial. Australian Council for Educational Research (ACER) / World Bank Institute (WBI).

3. Informando a familias y docentes: la experiencia del Territorio Capital de Australia (ACT)

El Territorio Capital de Australia tiene una interesante experiencia de evaluación, con un enfoque formativo (véase el recuadro 2). Dentro de este enfoque, dos modalidades de divulgación y uso de los resultados son particularmente interesantes. En primer lugar, la divulgación a los padres se realiza informándoles sobre el desempeño de su propio hijo, entregándoles una descripción detallada de los niveles de desempeño para cada competencia, al estilo de lo ilustrado anteriormente en las Fichass 6 y 7. Las Figuras 3.1 y 3.2 ilustran esta experiencia. Cada padre recibe algo así como un mapa de los desempeños en cada área evaluada y de la situación de su hijo.

Es importante poner de manifiesto la diferencia entre este enfoque y el de la información global sobre la escuela. Si bien no son enfoques incompatibles (de hecho, en el Territorio Capital de Australia estaba en estudio el modo de reportar a las familias resultados generales de la escuela en relación al resto, pero sin establecer rankings, con un enfoque bastante similar al de *Just For The Kids*), informar a los padres acerca de qué se espera que sus hijos aprendan y en qué grado sus propios hijos lo están logrando permite comprometer a los padres con el aprendizaje de sus propios hijos, los ayuda a comprender qué es lo que se está intentando enseñar y abre las puertas para que ellos puedan colaborar con la escuela apoyando a su propio hijo o hija.

Las escuelas reciben un informe que compara los resultados de sus alumnos con los de otras escuelas, en tanto los docentes reciben información sobre el resultado de cada uno de sus alumnos, y como elemento de comparación, el porcentaje de alumnos que respondió correctamente cada pregunta en todo el Territorio Capital. Con estas informaciones, cada escuela debe elaborar un informe de análisis siguiendo la pauta incluida en el recuadro 3.

4. Aportando a inspectores y maestros herramientas para evaluar

Uruguay también ha desarrollado su experiencia de evaluación a nivel nacional desde 1995 con un enfoque de tipo formativo y sin consecuencias para las escuelas.

Un primer aspecto a destacar de esta experiencia es un modo de comparar resultados entre escuelas que evitan los problemas analizados en la Fichas 10 relativos a los *rankings*. En este país, cuando las evaluaciones tienen carácter censal, cada Inspección Departamental

Figura 3.1. **Reportando a las familias (I)**

Territorio de la Capital Australiana

Programa de Evaluación
Informe estudiantil quinto grado, 1999

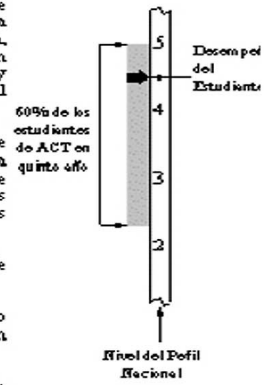
Estimados padres de familia o apoderados:

Su niño rindió recientemente algunas pruebas de lenguaje y matemáticas, junto con otros estudiantes de quinto año de las escuelas públicas del ACT, utilizando los instrumentos desarrollados por ACER denominados DART. Las pruebas han sido calificadas externamente por ACER.

Evaluación en capacidades lingüísticas
Las tareas de lectura exigían que cada niño leyera cartas, textos informativos e ilustrados y un poema, y luego responder preguntas sobre lo que habían leído. Las tareas de lectura de imágenes requerían que los niños vieran un video y respondieran algunas preguntas por escrito. Las tareas principales de Escritura pedían a los estudiantes escribir una historia o cuento y un argumento. Antes de responder por escrito a las preguntas sobre capacidad de escucha, los estudiantes oyeron una cinta de audio, mientras que sus propios maestros evaluaron presentaciones orales para medir su capacidad en lenguaje oral.

Evaluación en capacidades numéricas
Las tareas de sentido numérico exigían que cada niño manipulara y realizara cálculos con números presentados en tablas o recetas. Para algunas tareas los estudiantes podían usar una calculadora. Las tareas de sentido espacial involucraban interpretar mapas, fotos y diagramas, para demostrar su comprensión de conceptos espaciales tales como locación, forma, movimiento y orientación. Las tareas de medición y comprensión de datos implicaban medir y estimar tiempo, longitud, masa y área, y procesar e interpretar datos. Todas estas tareas estuvieron vinculadas al tema de Matemáticas en el Zoológico.

Ejemplo



Estos resultados colocan a su niño en uno de cuatro niveles de logros de los ocho niveles del Perfil Nacional del Currículo de Inglés (para kindergarten a décimo año) y del Perfil Nacional del Currículo de Matemáticas (para kindergarten a décimo año) que describen las capacidades lingüísticas y numéricas que típicamente muestran los estudiantes en ese nivel.

Las áreas sombreadas muestran el logro del 60% de los estudiantes de ACT en quinto año.

Los resultados de su niño están marcados con una ➡ para cada aspecto de lenguaje y matemáticas que ha sido evaluado. Se puede ver un ejemplo adyacente a este texto.

Es importante leer este reporte junto con otra información proporcionada por el docente de aula, ya que hay otros aspectos escolares que son de significancia para el éxito de su niño en la escuela

Fran Hinton
Oficial Principal

ACER Preparado por el Consejo Australiano de Investigación Educacional para el Gobierno de la ACT en 1999

Fuente: Margaret Forster, M., Masters, G. & Rowe, K., 2001; Measuring Learning Outcomes: Options and Challenges in Evaluation and Performance Monitoring. Material preparado para el Curso "Opciones Estratégicas para la Reforma Educativa" del Instituto del Banco Mundial. Australian Council for Educational Research (ACER) / World Bank Institute (WBI).

Reportando a las familias (II)

Figura 3.2

CAPACIDADES DE LENGUAJE QUINTO AÑO 1999

Lectura — Visualización — Escritura: contenido

Logros típicos de los estudiantes en cada nivel

Nivel 5

- Identifica el significado cultural de una imagen.
- Reconoce la relación entre un estilo de presentación y la naturaleza de la información (v.g., el formato de preguntas-respuesta en datos de entrevistas).
- Reconoce varias piezas de información relacionada de un texto largo.
- Infiere significado de lenguaje figurativo.

Nivel 4

- Interpreta una relación que no ha sido hecha explícita.
- Ordena eventos detallados de una narrativa.
- Reconoce el tema de un poema simple.
- Hace conexiones entre ilustraciones y texto escrito.
- Selecciona letras con la misma idea clave.
- Interpreta la conexión entre una ilustración y un texto.
- Selecciona varias piezas de información de una presentación compleja de texto (busca ejemplos de información solicitada).
- Encuentra evidencia para sustentar una declaración.
- Localiza información en una carta.
- Reconoce rasgos lingüísticos convencionales (v.g., guías de pronunciación).

Nivel 3

- Localiza información en una presentación compleja de texto.
- Conecta letras y materia temática en un formato de tabla.
- Ofrece una interpretación plausible de una porción de texto.
- Interpreta información ficticia.
- Interpreta una palabra inventada con referencia a una ilustración.
- Identifica la idea central en una carta (a partir de una lista dada).
- Reconoce la relación entre dos porciones de texto (v.g., interpreta el propósito de anotaciones).
- Hace conexiones entre palabras con significado similares (v.g., 'ilustrador' traza dibujos).
- Infiere información de claves obvias en una porción simple de texto.
- Infiere el significado de jerga técnica a partir de su contexto escrito y visual.
- Localiza información en una presentación compleja de texto.
- Extrae información de una presentación compleja de texto y fotografía.
- Identifica el propósito de un texto escrito o visual computarizado.
- Identifica el autor de una carta a partir de su estilo y contenido (v.g., una carta escrita por un niño).
- Infiere información implicada por el uso de palabra clave en un texto computarizado.

Nivel 5

- Explica la consistencia de una escena final con referencia al tema de una película.
- Explica el detalle de un texto en términos de su contribución a la estructura del mismo (v.g., la música con que se inicia y termina una película).

Nivel 4

- Identifica los elementos de una película que contribuyen a su tema (v.g., reconoce el uso de una parodia).
- Reconoce la audiencia objetivo de publicidad.
- Reconoce que las voces superpuestas pueden usarse para dar una perspectiva particular.
- Identifica una gama de técnicas que se usan para establecer un clima (v.g., ángulos de cámara, cámara lenta, edición).
- Interpreta literalmente las palabras usadas al final de una película.
- Justifica su propia interpretación de un texto (v.g., puede escoger referirse a personaje, trama o tema cuando interpreta la selección de recursos cinematográficos).
- Detecta inferencias como ideas implícitas en un texto.

Nivel 3

- Predice las instrucciones verbales para un procedimiento filmado, reconociendo el estilo instruccional.
- Resume una narrativa focalizándose en el tema central.
- Explica el detalle de un texto refiriéndose coherentemente a su contexto mediato.
- Explica la escena final de una película con referencia al argumento o trama.
- Explica el significado central de un recurso cinematográfico.
- Explica el propósito general de la publicidad.
- Explica un vínculo simple entre sección visual y texto oral.
- Identifica y explica la selección de una localización para una película.
- Puede extraer información contradictoria.
- Identifica la posición de la cámara en una toma.
- Entiende el propósito de una técnica cinematográfica básica.

Nivel 2

- Construye argumentos razonables para distintos puntos de vista (v.g., filmes alternativos de películas).
- Reconoce que las voces superpuestas se usan como recurso narrativo.
- Recuerda detalles de una película.
- Explica la escena final de una película sólo en términos de detalles pertinentes.
- Recuerda algunos pasos en un procedimiento.
- Infiere los motivos de un personaje.
- Identifica un elemento clave de una secuencia narrativa.
- Resume una narrativa focalizándose solamente en detalles pertinentes.
- Predice instrucciones verbales básicas para un procedimiento fílmico.
- Identifica un recurso cinematográfico central.

Nivel 5

- Escribe una narrativa desarrollada e integrada que incorpora el estímulo de manera convincente.
- Presta atención al ordenamiento temporal, un punto de vista narrativo consistente y el desarrollo de los personajes.
- Escribe un argumento coherente con evidencia considerable, aunque todavía no un caso totalmente desarrollado.
- Puede refutar un contra-argumento.
- Involucra y persuade a su audiencia.

Nivel 4

- Escribe una narrativa bien construida en términos generales que incorpora bien el estímulo y tiene una cálida tipificación de los personajes.
- Escribe un argumento coherente, focalizado en la cuestión y con algún razonamiento, quizás incorporando un reconocimiento de una posición opuesta.
- Intenta involucrar o persuadir al lector.

Nivel 3

- Mantiene el foco de la narrativa, con eventos y detalles que contribuyen a la trama de la historia y con personajes identificables.
- Usa el estímulo para la historia de manera plausible en el propio mundo de ficción.
- Escribe un argumento basado en el estímulo, con un punto de vista claro y cierta distancia crítica, y puede refutar brevemente a ambos lados.
- Considera el impacto sobre la audiencia y puede imitar modelos publicados.

Nivel 2

- Muestra comprensión del género narrativo y habilidades emergentes en lo que se refiere a construcción de trama, con los personajes emergiendo a través del discurso o la acción.
- Da un punto de vista personal y general sobre una cuestión, con varias ideas como sustento, y puede identificar "el otro lado".
- Usa el estímulo para estructurar la redacción.
- Muestra algunas conciencia de la audiencia.

Nivel 1

- Escribe una narrativa con un hilo de historia reconocible, con estructura narrativa y con una localización clara, pero con definición mínima de personajes.
- Escribe una opinión basada en involucramiento personal, con unas cuantas ideas relacionadas.
- Vincula claramente un cuento o una opinión con el estímulo ofrecido.
- Escribe con alguna coherencia, que podría no ser sostenida.
- Muestra un sentido emergente de audiencia.
- Incluye elementos de un cuento (puede ser corto), pero la narrativa carece de coherencia.
- Da una declaración simple de opinión, con alguna explicación, pero no apoya más en observaciones que en argumentos.
- Escribe un cuento o una opinión que tiene sentido para el lector.

Nivel 0

- Muestra alguna conciencia sobre la tarea.
- Escribe una respuesta muy corta o una larga y disconexa.
- No tiene control sobre la extensión de la historia, de la trama ni de los personajes.
- Escribe una historia personal en lugar de una opinión sobre un tema.
- Muestra poca, si alguna, conciencia de la audiencia.

Fuente: Margaret Forster, M., Masters, G. & Rowe, K., 2001; Measuring Learning Outcomes: Options and Challenges in Evaluation and Performance Monitoring. Material preparado para el Curso "Opciones Estratégicas para la Reforma Educativa" del Instituto del Banco Mundial. Australian Council for Educational Research (ACER) / World Bank Institute (WBI).

recibe un "mapa" de sus escuelas, como el incluido en la Figura 4.

En cada casilla se incluye grupos de escuelas, clasificadas en función de su contexto socio-cultural y de sus resultados.

Este modo de organizar la información sobre las escuelas tiene varias ventajas:

a. Evita ofrecer un ordenamiento posiblemente erróneo de las escuelas, como ocurre en el caso de los rankings (véase la Fichas 10), proponiendo una clasificación de las escuelas en cuatro grandes grupos en función de los resultados de sus alumnos.
b. Estos grupos no están definidos a partir de promedios, sino a partir del porcentaje de alumnos que alcanzaron un nivel de suficiencia en la prueba de Lengua. En este caso, se considera como escuelas con alto rendimiento a aquellas en que más de 3 cuartas partes de su alumnado alcanzaron dicho nivel. El grupo medio-alto está conformado por las escuelas cuyo porcentaje de alumnos suficientes estuvo entre la media nacional (57,5%) y el 75%.

Figura 4 **«Mapa socioacadémico» distribuido a los cuerpos inspectivos**

Resultado en la prueba de Lengua	Muy Favorable	Favorable	Medio	Desfavorable	Muy Desfavorable
Alto *75,1% a 100% de alumnos suficientes*	Escuelas:	Escuelas:	Escuelas:	Escuelas:	Escuelas:
Medio-alto *57,5% a 75% de alumnos suficientes*	Escuelas:	Escuelas:	Escuelas:	Escuelas:	Escuelas:
Medio-bajo *39,4% a 57,4% de alumnos suficientes*	Escuelas:	Escuelas:	Escuelas:	Escuelas:	Escuelas:
Bajo *0% a 39,3% de alumnos suficientes*	Escuelas:	Escuelas:	Escuelas:	Escuelas:	Escuelas:

Fuente: ANEP, Unidad de Medición de Resultados Educativos, 2000; Evaluaciones Nacionales de Aprendizajes en Educación Primaria en el Uruguay. 1995-1999. UMRE, Montevideo.

c. Los grupos se conformaron por separado, en función del contexto sociocultural de la escuela. De este modo se evitó la comparación entre escuelas con poblaciones muy diferentes. Al mismo tiempo, es posible identificar a las escuelas con muy buenos resultados en contextos desfavorecidos (la zona del "mapa" que aparece en gris).

Uno de los aspectos más interesantes de la experiencia de Uruguay en materia de evaluaciones estandarizadas ha sido la denominada "**evaluación muestral con aplicación autónoma en el conjunto de las escuelas**".

Esta estrategia busca mantener el involucramiento de todas las escuelas en la evaluación nacional, sin incurrir en el costo de una evaluación censal. Lo que se hace es evaluar en forma controlada desde la unidad central una muestra representativa pero pequeña de escuelas, para luego distribuir las pruebas al conjunto de todas las escuelas del país, junto con manuales para la aplicación y corrección de las pruebas en forma autónoma por parte de las escuelas.

Cada escuela participa en forma voluntaria y los resultados de su aplicación son manejados únicamente a nivel de la escuela. La escuela recibe luego los resultados nacionales y por contexto sociocultural, con una columna en blanco para ubicar allí los resultados de la propia escuela, como forma de compararse con los resultados nacionales y con los resultados de escuelas de similar población.

A modo de ejemplo, todas las escuelas del país recibieron las pruebas para los niños de 5 años, 1º y 2º grados (aquéllas que utilizaron los niveles de desempeño presentados en el ejemplo 5 de la Fichas 7), de modo que pudieron aplicarlas a sus alumnos y compararse con los resultados nacionales.

La experiencia ha mostrado que los maestros encuentran mucho más enriquecedor este ejercicio de evaluación autónoma que cualquier resultado nacional de una prueba en la que no participan directamente. De acuerdo a la investigación realizada en el marco de PREAL (Kaztmann, R., Aristimuño, A. B. Monteiro, L., 2003; **¿Cómo se usa y qué impacto tiene la información empírica sobre las evaluaciones nacionales de aprendizajes en el mejoramiento de la Educación Primaria pública en Uruguay?**; PREAL/GDN) en una encuesta representativa a nivel nacional alrededor del 80% de los maestros declara que participan de la aplicación autónoma de las pruebas.

Otra ventaja de este enfoque es que permite trabajar con pruebas de respuesta construida en lugar de pruebas de elección múltiple, dado que la muestra que debe ser corregida centralmente es relativamente pequeña en comparación con un censo.

5. Información para las autoridades y la opinión pública: un desafío complejo

Si bien su tratamiento exhaustivo excede los límites de esta Fichas, es necesario mencionar otras dos audiencias clave de los resultados de las evaluaciones, además de los educadores y las familias: las autoridades y la opinión pública, en este último caso, a través de los medios de prensa.

En relación a las autoridades, es preciso señalar la importancia de que las unidades de evaluación desarrollen su capacidad para entregar a las autoridades resúmenes ejecutivos que presenten con claridad lo principal de la información producida, así como análisis de situación que den una visión más amplia de los problemas educativos que la evaluación pone de manifiesto.

Esto último seguramente requiere de la discusión abierta de los resultados con diversos actores, dado que de los resultados no surge automáticamente el diagnóstico ni los lineamientos de trabajo. Es necesario un trabajo de investigación y de reflexión y análisis, que necesariamente debe involucrar a diversos actores.

Para las autoridades es importante, además, la información relativa a los factores que tienen incidencia sobre los resultados. Este aspecto es tratado en la Fichas 13.

En relación a la opinión pública, el problema central es que normalmente ni las unidades de evaluación ni los ministerios cuentan con recursos para desarrollar una campaña de tipo informativo, razón por la cual el modo en que los resultados se divulgan queda librado a los enfoques que la prensa quiera dar. Y, lamentablemente, la experiencia demuestra que, en general, predomina en la prensa un abordaje de tipo sensacionalista, que busca generar escándalo como modo de vender, más que la contribución a la información y al debate ciudadano.

Este aspecto ha sido investigado en el informe ¿Cómo aparecen los resultados de las evaluaciones educativas en la prensa?, disponible en el sitio web del Grupo de Trabajo de Estándares y Evaluación de PREAL (http://www.preal.org/Grupo2.asp?Id_Grupo=3).

En ese trabajo se analizan los tratamientos que la prensa da a los resultados de las evaluaciones, los errores más comunes que se cometen y recomendaciones para mejorar la forma de reportar.

Síntesis final

A lo largo de esta Fichas se intentó ilustrar con ejemplos específicos cómo las evaluaciones estandarizadas pueden ser utilizadas como herramientas, para el aprendizaje y para ayudar a las escuelas y a los docentes a mejorar su trabajo, desde un enfoque formativo antes que punitivo.

Las evaluaciones estandarizadas, sus instrumentos y sus resultados pueden ser utilizados para mejorar el modo en que los maestros evalúan y para identificar y dar a conocer experiencias de escuelas con niveles de logro destacados en contextos difíciles.

Para ello es fundamental producir estrategias y materiales de difusión específicamente pensados para apoyar el trabajo docente y, en especial, articular la divulgación de resultados con programas de capacitación en servicio. En este sentido, la articulación entre las unidades de evaluación y las áreas, programas e instituciones que desarrollan programas de formación y capacitación de docentes es clave para propiciar la reflexión a partir de los resultados de las evaluaciones y el cambio de las prácticas de enseñanza.

Pueden ser utilizados también para informar, involucrar y comprometer a las familias con el aprendizaje de sus hijos, pero desde la perspectiva de construir una relación de apoyo y colaboración con la escuela en una tarea común -–educar a los niños y jóvenes—, en lugar de generar una situación de enfrentamiento o una relación "cliente-proveedor".

La experiencia internacional indica que este tipo de aproximaciones favorecen mejor el desarrollo de una cultura de la evaluación y actitudes positivas hacia las evaluaciones externas que cuando las mismas tienen un carácter únicamente de control o una connotación punitiva.

¿QUÉ SON LOS "FACTORES SOCIALES"?

Comprendiendo el lenguaje de los iniciados

Muchas evaluaciones estandarizadas nacionales e internacionales incluyen, además de pruebas de aprendizaje, instrumentos complementarios de relevamiento de información acerca de las características de los alumnos, los docentes y los centros educativos.

Esta información es utilizada para analizar qué aspectos tienen incidencia sobre los resultados constatados a través de las pruebas. Con esto se busca ir más allá del reporte de resultados de aprendizajes, para intentar explicar qué es lo que influye sobre los mismos.

En América Latina se suele utilizar la denominación "estudios de factores asociados" a los trabajos de análisis e investigación desarrollados a partir de la interrelación entre resultados de pruebas e información recogida a través de los cuestionarios complementarios.

Esta Ficha tiene como propósito orientar al lector acerca de qué puede aportar este tipo de estudios, explicar algunos conceptos estadísticos relevantes para comprender sus metodologías y alertarlo sobre los alcances y limitaciones de los mismos.

1. El significado de la expresión "factores asociados"

Los denominados estudios de "factores asociados", como se ha dado en llamarlos en América Latina, responden a la preocupación por ir más allá de obtener información acerca de los resultados del sistema educativo.

Lo que se pretende con este tipo de estudios es explicar qué "factores" inciden en los resultados y, en especial, aquellos aspectos que puedan ser objeto de toma de decisiones por parte de las autoridades educativas y de los educadores, es decir, sobre los que se pueda actuar desde el sistema educativo.

FICHA 13

¿qué son los "factores sociales"?

Sabemos, antes que nada, que los principales "factores" que inciden sobre los aprendizajes son aquéllos de carácter sociocultural: el nivel educativo de los padres de los alumnos, el equipamiento cultural del hogar, su situación económica.

Como se explica en la Ficha 10, las escuelas que trabajan con alumnos que provienen de familias con mayor trayectoria en el sistema educativo y que viven en condiciones materiales razonables, tienen el camino allanado para lograr que esos alumnos aprendan. En cambio, las escuelas que trabajan con alumnos cuyos padres han tenido limitadas oportunidades educativas o que viven en condiciones de pobreza tienen una tarea bastante más difícil y exigente.

Sabemos también que el problema no es tanto el origen social individual de cada alumno, sino la composición social del alumnado de una escuela, es decir, la concentración de alumnos de origen social desfavorecido en ciertas escuelas.

Ahora bien, estos "factores sociales" están fuera del control de las escuelas y del sistema educativo. En este terreno la política educativa tiene poco para hacer, salvo evitar medidas que incrementen la segregación social de la matrícula o incrementen la motivación de los padres para apoyar a sus hijos en la escuela.

En cambio, hay "factores" a los que se suele denominar "escolares", que sí dependen de los educadores.

Investigar acerca de estos últimos es útil para comprender mejor lo que el sistema educativo puede hacer para mejorar los resultados de los alumnos, bajo dos premisas:

a. no hay soluciones mágicas que puedan derivarse de un análisis estadístico;
b. la acción educativa tiene límites y no es omnipotente; en condiciones de pobreza extrema es difícil que el sistema educativo pueda actuar y lograr resultados si dichas condiciones no se modifican desde el sistema social y político.

"Factores escolares" son pues aquéllos que pueden ser objeto de política educativa: el liderazgo educativo, el clima del centro escolar, la existencia de expectativas altas en relación al desempeño de los alumnos, la dotación de libros y textos en medios desfavorecidos, la experiencia y estabilidad de los equipos docentes, etc.; todo aquello en que se puede intervenir a través de la toma de decisiones dentro del sistema educativo.

Se los denomina "asociados" porque lo que se puede demostrar es que existe "asociación estadística" entre ciertos "factores" y los resultados medidos por las pruebas estandarizadas.

Sin embargo, como veremos más adelante en este Ficha, "asociación" estadística no es lo mismo que causalidad. Para hablar de *causalidad* y de *explicación* de los resultados se requiere de una teoría sólida, que interprete las asociaciones encontradas, al interior de un marco conceptual explicativo. Esto es lo que hacen corrientes de investigación como la denominada de "eficacia escolar".

El término "factores asociados" se utiliza entonces en la región, para denominar a los estudios que se realizan en paralelo a la aplicación de pruebas estandarizadas, normalmente mediante la aplicación de cuestionarios complementarios que recogen información sobre las características sociales de los alumnos, las características de las escuelas y de la experiencia educativa de los alumnos en ellas, con el fin de encontrar qué variables de tipo escolar están asociadas con los resultados.

2. ¿Qué es lo que muestran estos estudios?

Este tipo de estudios está emparentado con las corrientes de investigación educativa denominadas como de **"eficacia escolar"** y de **"mejora de la escuela"**.

La primera se propone investigar y explicar qué es lo que caracteriza a una escuela que logra los resultados educativos que se ha propuesto.

La segunda está más orientada hacia los procesos para generar cambios en la práctica que permitan mejorar la realidad de las escuelas.

Sobre estos temas existe abundante investigación en países europeos, así como en los Estados Unidos.

En América Latina existe una Red Iberoamericana denominada "RINACE" y una revista electrónica denominada REICE (www.ice.deusto.es/rinace/reice/) a través de las cuales el lector interesado puede encontrar abundante información[1].

Con el propósito de ilustrar al lector acerca de lo que la investigación en este terreno ha

1) Véase en particular los trabajos de Javier Murillo, El movimiento teórico-práctico de mejora de la escuela. Algunas lecciones aprendidas para transformar los centros docentes (REICE - Revista Electrónica Iberoamericana sobre Calidad, Eficacia y Cambio en Educación, 2003, Vol. 1, No. 2) y Un Marco Comprensivo de Mejora de la Eficacia Escolar (Universidad Autónoma de Madrid, 2004).

encontrado, en el Recuadro 1 se consigna una revisión de la literatura recientemente publicada por el Instituto Nacional para la Evaluación de la Educación (INEE) de México.

Los Recuadros 2 y 3 recogen los principales hallazgos realizados en este terreno en el marco de los Estudios Internacionales PISA y TIMSS. Como se puede apreciar, en ambos casos se destaca la necesidad de investigaciones más amplias para dar cuenta de lo que realmente incide sobre los resultados escolares.

En el sitio web del Grupo de Trabajo sobre Estándares y Evaluación, Sección Biblioteca, pueden encontrarse trabajos de este tipo realizados por diversas Unidades de Evaluación e investigadores en América Latina (Argentina, Chile, Ecuador, El Salvador, Honduras, Perú, Uruguay, etc.).

Recuadro 1

Factores de "eficacia escolar"

- La existencia de un liderazgo fuerte y compartido: el desarrollo de un enfoque participativo en la toma de decisiones que involucre todos los niveles de gestión y enseñanza en las escuelas.
- El desarrollo de una visión y metas compartidas: la unidad de propósitos, congruencia en las prácticas escolares, la realización de trabajo colegiado y la continua colaboración entre todos los niveles de autoridad en las escuelas.
- El desarrollo de un ambiente positivo para el aprendizaje: la creación de un clima de orden orientado a la tarea y de un ambiente de trabajo atractivo.
- La focalización en los procesos de enseñanza y aprendizaje: la maximización del tiempo de enseñanza, un énfasis en los aspectos académicos, centrados en el logro académico de los estudiantes.
- El desarrollo de prácticas favorables para la enseñanza: una organización eficiente del trabajo escolar, claridad en los propósitos educativos, lecciones estructuradas y reconocimiento de las diferencias entre los alumnos a fin de seleccionar las estrategias pedagógicas más pertinentes.
- La promoción de expectativas altas sobre el desempeño de los estudiantes: la comunicación de estas expectativas, la provisión de los medios para que los alumnos puedan alcanzarlas, el desarrollo de nuevas prácticas de enseñanza y la organización de actividades que signifiquen un desafío intelectual para los estudiantes a fin de que se den cuenta de su potencial.
- El desarrollo de una cultura de refuerzo positivo: involucrar a los estudiantes en actividades extracurriculares de manera que puedan usar y sintetizar el conocimiento abstracto aprendido en clase, una disciplina clara y justa y proporcionar retroalimentación a los estudiantes.
- La provisión de un sistema para supervisar el progreso de los alumnos: realizar evaluaciones continuas del desempeño de los estudiantes.
- El otorgamiento de mayores responsabilidades a los estudiantes: una práctica clara e invariable de derechos y responsabilidades de los alumnos, eleva su autoestima e incrementa la confianza en sus propios juicios.
- El desarrollo de una asociación escuela-hogar adecuada: un mayor involucramiento de los padres en el aprendizaje de sus hijos.
- La definición de la escuela como una "organización de aprendizaje": incorporar visiones e ideas para producir cambios, tanto dentro como fuera de la escuela; el desarrollo de un cuerpo académico basado en la escuela.

Fuente: Muñoz, C. y otros, (2004); Factores externos e internos a las escuelas que influyen en el Logro Académico de los estudiantes de primaria en México, 1998-2002. Análisis comparativo entre entidades con diferente nivel de desarrollo. México D.F., INEE- Universidad Iberoamericana.

FICHA 13

Recuadro 2

PISA: ¿Qué pueden hacer las escuelas que sea relevante?

"El entorno familiar influye sobre el éxito educativo y el estatus socio-económico puede reforzar sus efectos. De modo igualmente importante, el proyecto PISA identifica diversas acciones que pueden llevar a cabo las escuelas y que están asociadas con el éxito de los estudiantes. Este primer informe, al identificar una constelación de factores que interactúan o influyen sobre el rendimiento, no pretende proporcionar vínculos causales entre lo que las escuelas hacen y cómo rinden sus estudiantes. No obstante, los resultados iniciales ofrecen algunas claves sobre las condiciones de las escuelas que están más estrechamente asociadas con el éxito. Los resultados presentados a continuación consideran el efecto individual de cada factor identificado, una vez eliminados los efectos de otros factores escolares o del entorno social y de cualquier asociación con ellos. Los resultados presentados a continuación tienden a ser similares para la lectura, las matemáticas y las ciencias".

La utilización de los recursos de las escuelas por parte de los estudiantes está más estrechamente asociada con el rendimiento de los alumnos que la infraestructura física de las escuelas.

"Se preguntó a los estudiantes sobre su utilización de la biblioteca de la escuela, las computadoras, las calculadoras, los laboratorios y las conexiones a Internet. En las escuelas en las que el uso es relativamente alto, también lo son las puntuaciones en lectura, incluso cuando se ha eliminado el efecto de otros factores... Las deficiencias en la calidad de la infraestructura física o material de las escuelas, según lo referido por los directores de las escuelas, tienden a tener un impacto mucho más leve que la utilización de los recursos por parte de los estudiantes..."

El profesorado cualificado es uno de los recursos más valiosos de las escuelas.

"En el proyecto PISA se pidió a los directores de las escuelas que indicaran el porcentaje de profesores con una titulación universitaria en su respectiva área de docencia. La existencia de un mayor número de profesores con titulación universitaria está asociada con mejores resultados de los alumnos, como promedio en los países miembros de la OCDE. Por ejemplo, en lectura, un incremento del 25 por ciento en la proporción de profesores con una titulación universitaria en este área de contenidos está asociado con una mejoría en la puntuación de 9 puntos, como promedio en los países miembros de la OCDE,manteniéndose los otros factores constantes..."

La relación del número de estudiantes con el número de profesores es importante cuando está proporción es relativamente alta.

"En las escuelas en las que el número de estudiantes por profesor excede de 25 alumnos, el rendimiento promedio de los estudiantes es notoriamente menor cuanto mayor es esta ratio. En el rango habitual, de 10 a 25 alumnos por profesor, se observa una asociación mucho más débil con el rendimiento en habilidad lectora. De hecho, las escuelas con menos de 10 alumnos por profesor obtienen puntuaciones ligeramente menores que el promedio de los países miembros de la OCDE, lo que puede deberse a que muchas de estas escuelas atienden a estudiantes con necesidades especiales..."

Algunos aspectos de la gestión y la práctica educativa de las escuelas tienden a estar asociados con un mejor rendimiento de los estudiantes.

"Tres de estos factores, tal como los perciben los directores de las escuelas, tienen un impacto positivo y estadísticamente significativo, como promedio en los países miembros de la OCDE... incluyendo:

- los factores relacionados con el profesorado que afectan al clima escolar, tales como las expectativas del profesorado con respecto al rendimiento de los estudiantes;
- la moral y compromiso del profesorado; y
- la autonomía escolar".

Algunos aspectos de las prácticas en la clase están asociados con un mejor rendimiento de los estudiantes.

"Tres de estos factores, tales como los perciben los estudiantes, mantienen una asociación positiva y significativa estadísticamente con el rendimiento de los estudiantes:

- las relaciones entre el profesor y los alumnos;
- el clima de disciplina en las clases, y
- el grado en que los profesores enfatizan la importancia del rendimiento académico y exigen a los alumnos un alto rendimiento.

Los dos primeros factores son más importantes que el tercero..."

Es más probable que hagan deberes los estudiantes con éxito que aquellos que no lo tienen.
"El otro factor escolar que presenta la asociación más fuerte con el éxito de los estudiantes es el constituido por los deberes escolares. Dentro de cada país, es más probable que los estudiantes que hacen más deberes obtengan una mejor puntuación en lectura, como promedio de todos los países. La cuarta parte de los estudiantes que hace más deberes obtiene, como promedio, una puntuación 44 puntos más alta que la cuarta parte de los alumnos que hace menos deberes. Esta asociación es más fuerte en los países cuyos estudiantes hacen, como promedio, más deberes..."

Es necesario llevar a cabo una investigación complementaria

"En conjunto, teniendo en cuenta las tres áreas evaluadas, la influencia combinada de este conjunto de variables escolares da razón del 31 por ciento de la variación en lectura entre escuelas dentro de los países, y del 21 por ciento de la variación entre países. Junto con las características del entorno familiar, el conjunto de los factores explican el 72 por ciento de la variación entre escuelas dentro de los países y el 43 por ciento de la variación entre países... Estos resultados proporcionan una primera aproximación a los resultados del proyecto PISA. Será necesario llevar a cabo tanto gran cantidad de investigación complementaria como análisis ulteriores para identificar cómo opera cada factor escolar, cómo interactúa con el entorno familiar y cómo influye sobre el rendimiento de los estudiantes y de las escuelas".

Fuente: INCE / OCDE, 2001; Conocimientos y Destrezas para la Vida: Primeros Resultados del Proyecto PISA 2000: Resumen de Resultados. Madrid, Ministerio de Educación, Cultura y Deporte.

Recuadro 3

TIMSS: Escuelas efectivas en Ciencias y Matemáticas-Resumen de resultados

"El contraste entre las escuelas con mayor y menor rendimiento en ciencia y matemática de cada país mostró que los indicadores socioeconómicos del entorno familiar y del apoyo de los padres para el logro académico distinguían de manera especialmente consistente a dos grupos de escuelas. En casi todos los países, los estudiantes de las escuelas con mayor rendimiento poseían más libros y materiales de apoyo, niveles más altos de posesiones en el hogar y de educación paterna y pasaban menos tiempo trabajando en el hogar. Otro factor distintivo en lo vinculado al hogar eran las aspiraciones de los estudiantes respecto a seguir estudios superiores. En la mayoría de países, los estudiantes de las escuelas de mayor rendimiento reportaron con mucho mayor frecuencia planes de asistir a la universidad después de la escuela secundaria

"Los factores relacionados más directamente con la escuela resultaron menos uniforme eficacia para distinguir entre escuelas con logros altos y bajos. Si bien factores como el tamaño y la ubicación de la escuela, el clima de la escuela, la actitud de los estudiantes hacia la ciencia y la matemática, y las actividades pedagógicas en las clases de ciencia y matemática sí discriminaron entre escuelas con logros altos y bajos en algunos países, pocas variables de la escuela funcionaron de manera consistente en todos los países. Esto indica que es los análisis de las características de las escuelas eficaces posiblemente resulten más fructíferos si se utiliza diferentes variables en diferentes países o grupos de países, en lugar de las variables comunes que operan del mismo modo en todos los países.

"Los resultados que se presentan en el segundo capítulo muestran que la medida en que los logros en ciencias y matemáticas pueden asociarse con factores escolares varía considerablemente de un país a otro, así como que el grado en el cual los entornos familiares de sus estudiantes difieren de escuela en escuela tampoco es igual en todos los países. Queda claro que la forma en que el entorno del hogar de los estudiantes se relaciona con sus logros y la forma en que el sistema escolar modera o magnifica esta relación están estrechamente vinculadas a factores organizacionales sociales y escolares exclusivos de cada país, y que cualquie esfuerzo de análisis comparativo internacional debe esto tener en cuenta.

"A pesar de que sólo un pequeño conjunto de variables relacionadas con el aula sobrevivieron al proceso de selección, éstas dieron cuenta de una gran parte de las diferencias entre escuelas en la mayoría de países. El indicador más predominante fue la realización diaria de tareas en diversos cursos (lenguaje, matemáticas y ciencia). Las escuelas donde se esperaba que los alumnos de octavo grado pasaran tiempo haciendo tareas de diversos cursos obtuvieron mayores logros en ciencia y matemáticas, incluso después de controlar por indicadores del entorno familiar de los estudiantes de la escuela. Las características de los docentes, el clima social de la escuela y características demográficas tales como la ubicación de la escuela y el tamaño de la clase fueron predictores menos consistentes del logro de un país a otro. Entre las variables que posiblemente puedan estar

influenciadas tanto por el hogar como por la escuela (la interfase hogar-escuela), el nivel promedio de las aspiraciones de los estudiantes de seguir estudios superiores resultó ser un predictor significativo del logro escolar en ciencia en la mayoría de países y en matemáticas en casi todos los países.

"Si bien los resultados muestran que las variables relacionadas con el aula están relacionadas con el logro promedio de la escuela incluso después de controlar por el entorno familiar de sus estudiantes, la fuerte relación que persiste entre el nivel promedio del entorno familiar y el logro estudiantil ajustado también sirve como recordatorio de que, en muchos países, el entorno del hogar/familiar, la educación escolar y el logro de los estudiantes están estrechamente ligados, y que el discriminar las influencias relativas de los diversos factores involucrados continúa siendo un gran desafío."

Fuente: Martin, M.O. y otros, 2000; Effective Schools in Science and Mathematics, IEA's Third International Mathematics and Science Study. International Association for the Evaluation of Educational Achievement (IEA) / International Study Center, Lynch School of Education, Boston College. Traducido por el editor.

3. Conceptos estadísticos básicos para comprender los estudios de "factores asociados"

Este apartado tiene como propósito explicar al lector algunos conceptos básicos de estadística necesarios para comprender mejor los estudios de "factores asociados".

3.1. Variables dependientes e independientes

El término "variable" se utiliza en estadística y en la investigación de las ciencias naturales y sociales para designar aspectos de la realidad que pueden cambiar y son susceptibles de ser medidos o clasificados en categorías –y, por tanto, pueden asumir distintos valores o categorías-.

Lo que hasta este momento hemos denominado como "factores" pueden ser considerados, desde el punto de vista estadístico, como variables. Por ejemplo, el puntaje de los alumnos en una prueba, el nivel educativo de los padres de los alumnos, la condición de pobreza de las familias, el clima del centro educativo, el nivel de formación de los profesores.

Se denomina como *"variable dependiente"* a aquélla cuyo comportamiento queremos

explicar, que, en el caso de los estudios de factores asociados, suelen ser los resultados en las pruebas estandarizadas.

Se denomina "variables independientes" a aquéllas que se supone influyen en el comportamiento de la variable dependiente, como podrían ser, en el caso de estos estudios, el nivel sociocultural de la familia o del grupo del estudiante, la estabilidad del equipo docente, el clima del centro educativo, etc.

3.2. Asociación entre variables y correlación

Lo que el análisis estadístico se propondrá mostrar es en qué medida las variables están "asociadas". Dos variables están "asociadas" si ocurre sistemáticamente que cuando una se modifica, la otra también lo hace.

La asociación entre dos variables se mide a través de lo que estadísticamente se denomina "correlación".La correlación es una medida del grado en que dos o más variables varían en forma conjunta, es decir, que cuando una varía en un cierto sentido, la otra también varía.

Por ejemplo, según hemos analizado anteriormente, cuando el nivel educativo de la familia del estudiante aumenta, sus resultados en las pruebas tienden a aumentar.

La correlación entre variables se mide a través de un índice al que se denomina con la letra 'R', que varía entre –1 y +1. Si el valor de 'R' es 0, significa que no hay ninguna relación entre una variable y otra. Si el valor de 'R' es 1, significa que la correlación es perfecta, es decir, por cada unidad en que una variable aumenta o disminuye, se produce una unidad de cambio igual en la otra, un caso de determinación perfecta que difícilmente se da en la realidad social

Si el signo de 'R' es negativo, ello significa que cuando una variable aumenta su valor la otra disminuye. Por ejemplo, se anticipa que cuando la rotación de profesores en un centro aumenta, los rendimientos de los alumnos disminuyen.

Si el signo es positivo, significa que cuando la variable independiente aumenta, también lo hace la variable dependiente. Por ejemplo, si el nivel educativo de la familia aumenta, también lo hace el resultado del estudiante en la prueba.

Usamos el término probabilidad porque la determinación no es perfecta. No existe una ley natural que determine los resultados en función del nivel educativo de la familia de un estudiante. De hecho, muchos casos se desvían de la asociación. Hay alumnos de familias con escasa trayectoria educativa que logran buenos resultados, y viceversa. Por eso no es posible hablar de "determinación" ni de "causalidad", sino de "asociación" y "probabilidad".

'R' nos dice entonces tres cosas distintas:

a. Si dos variables están vinculadas o asociadas. Para ello el valor de 'R' debe ser distinto de 0.
b. Qué tan fuerte es la relación (una correlación inferior a 0,30 es considerada como débil; entre 0,30 y 0,60 como moderada; superior a 0,60 como fuerte, y cuanto más cerca de 1, más fuerte).
c. El sentido de la asociación (si el signo de 'R' es positivo significa que ambas variables aumentan o disminuyen juntas; si el signo de 'R' es negativo significa que cuando una variable aumenta la otra disminuye y viceversa).

3.3. Varianza o variabilidad

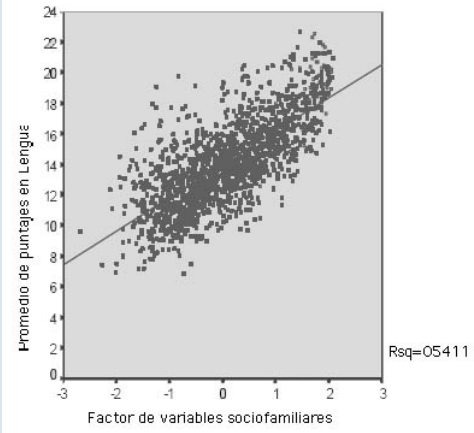
Muchos estudios utilizan el término *varianza* en lugar de correlación. La varianza es el grado de variación de una variable. Lo que se informa en muchos estudios es qué proporción de la varianza o variabilidad de la variable dependiente está vinculada a (o es explicada por) la variación de una o varias variables independientes.

Por ejemplo, en la evaluación nacional de aprendizajes realizada en Uruguay en 1996, el 54% de la variabilidad de los puntajes promedios de los grupos de 6º año en la prueba de Lenguaje estaba asociado a la variabilidad de la composición sociocultural del alumnado del grupo (véase la Figura 1).

La varianza es el cuadrado de 'R'. En el ejemplo anterior, la correlación entre el puntaje promedio de cada grupo y su composición sociocultural era de 0,7356 ('R'). Por lo tanto, la varianza explicada por este factor fue 0,7356 x 0,7356 = 0,5411. Redondeando las cifras, esto significa que el 54% de la variación de los puntajes promedios de las grupos está asociada a la variación del valor del factor sociocultural de dichos grupos.

FICHA 13

Figura 1 **Gráfico de dispersión**



Fuente: ANEP/ UMRE, 1999; Estudio de los Factores Institucionales y Pedagógicos que Inciden en los Aprendizajes en Escuelas Primarias de Contextos Sociales Desfavorecidos en el Uruguay. Montevideo, ANEP/UMRE.

3.4. Gráfico de dispersión

Un gráfico de dispersión es un modo de representar la relación entre dos variables. En el ejemplo de la Figura 1, el gráfico representa la relación o asociación entre los puntajes promedio en Lenguaje (eje Y) y la composición social del grupo (eje X). Cada punto en el gráfico representa un grupo de alumnos de 6° grado (los puntos o unidades de análisis también podrían haber sido alumnos individuales, escuelas, países, etc.).

Cada punto o sección está ubicado en el gráfico de acuerdo a su puntaje promedio y al valor del factor social. El gráfico permite observar que a medida que aumenta el valor del factor social, también aumentan el puntaje promedio en la prueba. Por eso la nube de puntos se eleva de izquierda a derecha.

La recta en el gráfico representa esta relación y se la denomina "recta de regresión".

Si todos los puntos estuviesen ubicados sobre la recta, significaría que la correlación entre las variables representadas en los ejes X e Y sería perfecta. La correlación 'R' sería igual a 1.

El hecho de que los puntos se desvíen hacia arriba y hacia abajo de la recta significa que en

algunas secciones el promedio en Lenguaje es mayor a lo esperable en función de su composición social, en tanto en otros es inferior. Esto simplemente demuestra que la asociación entre las variables no es perfecta, sino una tendencia que puede ser más o menos marcada. Pero no existe determinismo.

3.5. Análisis multi-variado

Hasta el momento hemos hablado de correlaciones o asociaciones entre dos variables, una dependiente y otra independiente. A este tipo de correlación, que se verifica entre dos variables, se la denomina ***bi-variada***.

Sin embargo, muchas veces la asociación observada entre dos variables puede tener carácter espúreo. Esto significa que en realidad las dos variables no están causalmente asociadas entre sí, sino que su relación se debe a que ambas están asociadas con una tercera variable que no está siendo considerada.

Un ejemplo de esto es la asociación entre resultados de pruebas y el carácter público o privado de la escuela. En muchos informes nacionales se reporta que las escuelas privadas tienen mejores resultados que las públicas. Pero lo que no se dice es que en realidad hay una tercera variable, la composición social del alumnado, a la que las dos primeras están asociadas (las escuelas privadas seleccionan alumnos de origen social más favorecido y este origen favorece mejores resultados en las pruebas) y que es la que realmente explica la relación.

También puede ocurrir lo contrario: que dos variables parezcan no relacionadas entre sí, porque en realidad están asociadas a una tercera con valores opuestos para cada una de ellas.

Un ejemplo de esto, investigado en Uruguay, es el siguiente. En principio no se encontró correlación entre un índice de actualización pedagógica del maestro del grupo y los resultados de los alumnos, cuando todo hacía suponer que debería haberla. Indagando en las relaciones entre las distintas variables, se encontró que existía una tercera, la zona geográfica de la escuela, que estaba interviniendo de la siguiente manera: en el interior del país los resultados tendían a ser mejores que en la capital debido a un mayor apoyo de la comunidad, menor marginalidad y mayor estabilidad de los equipos docentes. Pero la actualización pedagógica de los docentes del interior era menor, dado que las oportunidades de capacitación son menores.

En este marco, las variables se contrarrestaban entre sí. La actualización pedagógica era menor en el interior, donde los resultados eran mejores por otros motivos, y mayor en la capital, donde otros factores incidían en más bajos resultados pero donde el nivel de actualización sí gravitaba sobre esos resultados. Cuando se examinaba los datos agregados, el efecto sobre los aprendizajes quedaba oculto.

Para detectar estas situaciones es necesario realizar lo que se denomina un análisis «***multi-variado*** o de *"**correlación múltiple**"* que analiza en forma simultánea el efecto de un conjunto de variables independientes sobre la variable dependiente, teniendo en cuenta las asociaciones existentes entre las diversas variables independientes y el efecto conjunto de cada una de ellas.

Los análisis de asociación entre variables en educación necesariamente deben ser de tipo "multi-variado", dada la diversidad de elementos que interactúan en la producción de los resultados.

Por lo tanto, un cuidado que el lector debe tener al analizar los estudios de factores asociados es que no se basen exclusivamente en asociaciones bivariadas. Ante cada una de ellas debe preguntarse siempre si no existe la posibilidad de que una tercera variable sea la que en realidad explica la asociación.

3.6. Análisis multi-nivel

Un término que el lector probablemente encontrará en muchos estudios de factores asociados es el denominado "análisis multinivel" o de "niveles múltiples".

En educación los desempeños logrados por los alumnos dependen de variables que pertenecen a distintos niveles. Algunas corresponden al nivel ***individual:*** las características de la familia de cada alumno, su trayectoria educativa anterior, su motivación con el estudio, etc. Otras tienen carácter *grupal* y caracterizan por igual a todos los alumnos de una misma sección (la experiencia del maestro, el clima de disciplina en el grupo, la composición social del grupo, el tiempo destinado a tratar cada tema) o de una misma escuela (las características del director, el ambiente escolar, el equipamiento didáctico de la escuela, etc.).

El *análisis multinivel* es una técnica estadística compleja que permite distinguir el efecto de las variables independientes de los distintos niveles sobre la variable dependiente, es decir,

qué proporción de la variabilidad de los resultados de los alumnos depende de las características individuales de éstos, qué proporción se debe a variables propias de las secciones a las que los alumnos pertenecen, qué proporción depende de variables propias de la escuela.

Es a partir de este tipo de análisis que se llega a establecer con precisión que es más importante la composición social de la escuela a la que pertenece el alumno que su origen social individual, según fue indicado y ejemplificado en la Ficha 10.

3.7. Resultados ajustados

Otro concepto importante es el de resultados ajustados. Dado que la principal determinante de los resultados de una escuela es su composición social, su puntaje promedio no es, por sí solo, un buen indicador de qué tan bien la escuela está haciendo las cosas. Esto fue explicado en la Ficha 10.

Lo que importa es qué tanto la escuela logra **por encima o por debajo** de "lo esperable", dados los resultados promedio que obtienen escuelas de composición social similar de su alumnado. En términos de la Figura 1, lo que importa es la relación de cada punto con la denominada "recta de regresión".

Un punto ubicado justo en la recta representa una sección cuyo resultado corresponde con lo esperable. Un punto alejado hacia arriba representa una sección o grupo de alumnos cuyo puntaje promedio es superior al esperable. Es por tanto, una sección con muy buenos resultados. A la inversa, un punto alejado hacia abajo representa una sección cuyos resultados están por debajo de lo esperable.

Nótese que dos secciones pueden tener el mismo puntaje promedio (por lo tanto estar sobre una misma línea horizontal) pero tener méritos muy diferentes. Una puede estar hacia la izquierda del gráfico, es decir, en contextos sociales desfavorecidos, y quedar por sobre la recta de regresión. La otra puede estar hacia la derecha del gráfico, es decir, en contextos sociales favorecidos, y quedar por debajo de la recta de regresión. Ambas secciones tienen el mismo promedio, pero en el primer caso el resultado debe ser evaluado como positivo y en el segundo como negativo.

Es posible entonces considerar como medida de la calidad del trabajo de una sección o escuela no a su promedio, sino a la distancia a la cual se ubica de la recta de regresión.

Esta pasaría a ser la variable dependiente relevante: qué tanto se ubica el promedio de la sección o la escuela por encima o por debajo de lo esperable en función de su composición social. A esta nueva variable se la denomina *resultado ajustado* en función de la composición social del alumnado.

3.8. Valor agregado

Otro modo de resolver el problema de la fuerte influencia de la composición social en la determinación de los resultados escolares consiste en recurrir a las denominadas evaluaciones de *"valor agregado"*.

En estos casos lo que se hace es realizar dos mediciones de los desempeños de los alumnos, uno al inicio del año escolar y otra al final del mismo. Luego se comparan ambas y se establece como variable dependiente, indicativa de la calidad del trabajo de una escuela, a la diferencia entre la situación inicial y la situación final (véase la Ficha 3).

En este caso, tampoco importa el promedio absoluto que cada escuela obtenga, sino el grado de avance respecto a su situación inicial. Una escuela privada puede tener un promedio muy alto al final del año, pero si su promedio al inicio del año también era alto, su avance habrá sido escaso y no debería ser evaluada necesariamente como una buena escuela.

En cambio, otra escuela puede tener un promedio final bastante inferior al de esta escuela privada, pero haber partido de una situación inicial muy mala y haber mejorado mucho durante el año. Esta última escuela será evaluada como muy buena.

El enfoque de valor agregado tiene la ventaja de que permite neutralizar el efecto de la composición sociocultural y que permite establecer asociaciones o relaciones entre los que los alumnos aprendieron durante un año específico y lo que la escuela o el maestro hizo durante ese año con esos alumnos. Esto neutraliza no solo el efecto sociocultural, sino también el efecto de la historia escolar y acumulación de conocimientos previos de los alumnos. Por tanto, permite desarrollar estudios de "factores asociados" más precisos y pertinentes.

El principal problema de este enfoque y la razón por la cual se lo utiliza muy poco, es su elevado costo, dado que exige implementar dos operaciones de evaluación de los mismos alumnos en el mismo año.

Síntesis final

Alcances y limitaciones de los estudios de factores asociados

Los estudios de "factores asociados" deben ser apreciados en su justo valor, evitando esperar de ellos más de lo que pueden dar.

- El primer cuidado que debe tenerse es evitar creer que de un estudio de "factores asociados" se pueden derivar en forma automática recomendaciones de política educativa.
  Según se indicó más arriba, la existencia de asociación estadística no implica causalidad. Para establecer relaciones explicativas de causalidad hay que formular una teoría respaldada por múltiples observaciones e investigaciones.
  Del mismo modo, para establecer opciones de política educativa se necesita mucho más que analizar correlaciones entre variables escolares y resultados de pruebas. Estas pueden servir para acumular evidencia empírica, pero la formulación de opciones de política educativa y la toma de decisiones requiere de un esfuerzo mucho mayor de conocimiento y comprensión de la realidad educativa, así como de discusión y debate.

- Un segundo aspecto que debe ser tenido en cuenta por el lector es la limitación inherente a toda aproximación al conocimiento de la realidad.
  Los estudios estadísticos del tipo "factores asociados" están limitados a los aspectos de la realidad que pueden captarse a través de cuestionarios.
  Sin embargo, existen múltiples facetas relevantes de la acción educativa que escapan a lo que dichos cuestionarios pueden captar: la habilidad didáctica del docente, el enfoque desde el cual enseña su disciplina, la relación diferencial que establece con distintos tipos de alumnos, etc.
  Esta limitación implica que este tipo de estudios de corte cuantitativo debieran complementarse de manera sistemática con observaciones de carácter cualitativo en escuelas y clases especialmente seleccionadas, por ejemplo, por sus buenos resultados en contextos de pobreza.
  Esto se hace en algunos países de la región y lo ha hecho el Laboratorio Latinoamericano de la UNESCO. Es lo que hace la organización "Just For The Kids" mencionada en la Ficha 12.
  El estudio internacional TIMSS, además de la investigación referida en el Recuadro 3, llevó adelante un interesante estudio sobre enfoques de la enseñanza a través de la filmación en video de clases de Matemática en siete países[2].

- Una tercer elemento a tener presente al analizar un estudio de "factores asociados" es que los modelos estadísticos suelen explicar una parte limitada de la varianza de los resultados, quedando sin explicar la mayor parte de ella.
  Esto significa que en realidad, la mayor parte de los estudios solo logra explicar una parte limitada de los factores que influyen sobre los logros de los alumnos y que no sabemos qué es lo que explica el resto (probablemente, aspectos de la realidad educativa que no son captados por el tipo de instrumentos utilizados). Si esto no se explicita, el lector puede quedarse con la falsa impresión de que todo se explica a través de las variables incluidas en el estudio y que solo esas variables son importantes, cuando en realidad hay múltiples aspectos importantes que no están siendo contemplados.

- Finalmente, una debilidad que caracteriza a muchos estudios de "factores asociados" es su pretensión de elaborar un "modelo" universal que explique los resultados de escuelas de muy diverso tipo, lo cual implica un fuerte grado de sobre-generalización.
  La realidad educativa, en cambio, es muy diversa y lo que favorece el aprendizaje en ciertos contextos puede no hacerlo en otros.
  De allí la necesidad de tener miradas diferenciadas para diversos países, así como de regiones y tipos de escuela dentro de un mismo país.

Más que buscar un modelo universal que explicaría los resultados en todo tipo de escuela, lo que se requiere es acumular conocimiento acerca de cómo funcionan diversos sistemas educativos y diversos tipos de escuelas dentro de ellos, y qué factores tienen influencia sobre el aprendizaje de los alumnos en diferentes contextos

2) Véase : Hiebert, J. y otros, 2003; Teaching Mathematics in Seven Countries. Results from the TIMSS 1999 Video Study. Washington D.C., U.S. Department of Education /National Center for Educational Statistics (NCES).

FICHA 14

¿CÓMO ANALIZAR UN REPORTE DE EVALUACIÓN?

Las preguntas que el lector debe hacerse ante un informe de resultados

La Ficha Nº 14 intenta ser un resumen de cierre del conjunto del trabajo que sirva de guía al lector acerca de lo que contienen o deberían incluir los reportes de las evaluaciones –marcos conceptuales, datos, definiciones, ejemplos de ítemes, etc.–, relacionándolo con los temas analizados a lo largo de las Fichas. Para ello, esta Ficha propone al lector un conjunto de nueve preguntas fundamentales que debe formularse ante todo reporte de resultados de una evaluación:

1. ¿Cuál fue el propósito o finalidad de la evaluación?
2. ¿Qué fue evaluado? 3. ¿Cuál fue el universo estudiado?
4. ¿Cuál fue el enfoque de la evaluación?
5. ¿Qué tipo de datos se proporciona y qué significan esos números?
6. ¿Qué grado de precisión tiene la información?
7. ¿En qué grado y de qué modo se contextualiza la presentación de los resultados? 8. ¿Quiénes son los destinatarios de la información?
9. ¿Qué consecuencias e implicancias tienen los resultados?

GUIA PARA LA LECTURA DE REPORTES DE EVALUACIONES ESTANDARIZADAS

1. ¿Cuál fue el propósito o finalidad de la evaluación?

Todo reporte de evaluación debería incluir una explicitación de para qué se realiza la evaluación, qué tipo de consecuencias tendrán sus resultados, cómo van a ser utilizados los mismos y por quiénes.

2. ¿Qué fue evaluado?

Los reportes de evaluación deberían explicar claramente cómo fue definido el dominio evaluado. Esto implica explicar en forma comprensible cuál es la concepción de la disciplina o área evaluada desde la cual se concibió la prueba y cuáles fueron los contenidos y competencias que se seleccionó como fundamentales, así como la justificación de dicha selección.

Dicha justificación puede estar apoyada directamente en el currículo vigente o hacer referencia a algún otro tipo de proceso de discusión y consultas con especialistas, docentes y otros actores involucrados. La justificación debería incluir la referencia a por qué es relevante evaluar lo que fue evaluado.

Parte importante de la exposición acerca de qué fue evaluado es la ilustración con ejemplos del tipo de tareas que los alumnos debían resolver en la prueba, de modo que el lector pueda formarse una imagen concreta de lo que fue evaluado, más allá de las definiciones conceptuales.

Debería incluirse además, al menos como anexo, alguna información sobre los procedimientos de validación seguidos para asegurar la consistencia entre las actividades de la prueba y las definiciones conceptuales de las que se partió.

3. ¿Cuál fue el universo estudiado?

Los reportes de evaluación normalmente incluyen una descripción de la población que fue evaluada. Ello incluye información sobre los grados y disciplinas evaluadas, así como sobre los niveles en que la información podrá ser desagregada (nacional, regional, provincial, municipal, etc.). Debería ofrecerse, además, una justificación de estas decisiones, en términos de los propósitos y la estrategia general de la evaluación.

Cuando se trabaja con muestras debería incluirse, al menos en anexos, información relativa a cómo fueron seleccionadas las mismas y a los márgenes de error muestral.

4. ¿Cuál fue el enfoque de la evaluación?

El enfoque implica, en primer término, explicitar si la evaluación fue concebida con un carácter normativo o criterial (véase la Ficha 3).

Esto en parte podrá "verse" en la definición de propósitos así como en la definición de qué fue evaluado.

Los reportes deberían explicar al lector si la evaluación pretende establecer comparaciones con evaluaciones anteriores o futuras, en cuyo caso debería incluir una explicación de los recaudos técnicos tomados para hacer posibles dichas comparaciones.

En cualquier caso, los reportes deberían explicar cuál es el plan general de evaluaciones para el conjunto del sistema y para los años que siguen (qué grados se evaluarán, cada cuántos años, en qué áreas o disciplinas, etc.), así como la racionalidad de dicho plan.

En el caso de las evaluaciones de carácter criterial que incluya una definición de niveles de desempeño, debería explicarse de qué modo fueron construidos los mismos y cuál es el significado de cada uno de ellos.

Si la prueba incluye un estándar o expectativa respecto al nivel o puntaje que se espera todos los alumnos estén en condiciones de lograr, debería explicarse cómo se llegó a establecerlo y qué implica en términos de los que los alumnos deben conocer y ser capaces de hacer.

5. ¿Qué tipo de datos se proporciona y qué significan esos números?

E l lector debería encontrar en los reportes información que le permita saber qué tipo de datos numéricos encontrará, incluyendo si se trata de puntajes de la TCT o de la TRI (véase la Ficha 8).

Debe distinguir cuando se le entregan promedios de cuándo se le entregan distribuciones de frecuencias de los alumnos en categorías definidas.

6. ¿Qué grado de precisión tiene la información?

El lector debe poder encontrar información respecto al grado de precisión –o el error estimado– de los datos. Debe observar, además, si la información relativa a los márgenes de error e intervalos de confianza es tenida en cuenta a la hora de establecer conclusiones o juicios de valor.

7. ¿En qué grado y de qué modo se contextualiza la presentación de los resultados?

Los reportes de evaluación no deberían limitarse a entregar los resultados de las pruebas sino relacionarlos de algún modo con otros factores relevantes que ayudan a comprender y explicar esos resultados.
De primera importancia es que se incluya información relativa a los contextos sociales en que los resultados se producen. Ninguna comparación de resultados entre escuelas, provincias, países, etc., debería realizarse sin ofrecer alguna información sobre las características socioculturales de los estudiantes de dichas entidades.
En segundo término, el lector debe buscar si los reportes incluyen información sobre aspectos internos al sistema educativo que inciden sobre los resultados de los alumnos o que contribuyen a atenuar las inequiedades generadas desde el entorno social.

8. ¿Quiénes son los destinatarios de la información?

L os reportes deberían explicitar cuáles son los principales destinatarios o audiencias que se espera utilicen los resultados y de qué modo se espera que la información pueda serles útil.
Esto puede exigir distintos tipos de reportes, adaptados a audiencias distintas y a usos diferentes.

9. ¿Qué consecuencias e implicancias tienen los resultados?

Los reportes de evaluación deberían incluir algún tipo de "avance" hacia las consecuencias y el uso de resultados. Esto incluye aspectos tales como una reflexión sobre los desafíos que los resultados muestran para la política educativa y para los educadores, alguna indicación o recomendación del tipo de acciones que deberían tomarse, propuestas de perfeccionamiento de los enfoques y prácticas de enseñanza, etc.
Lo importante es que la evaluación no termine en la presentación de unos datos, sino que dé los primeros pasos hacia la generación de cambios. Obviamente no corresponderá a los evaluadores avanzar en la generación de dichos cambios, pero tampoco deberían dar por sentado que los mismos se producirán automáticamente. Es responsabilidad del evaluador propiciar la reflexión de otros actores y la generación de acciones a partir de los resultados de su trabajo.

PREAL

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

Desde su creación en 1995, el PREAL ha tenido como objetivo central contribuir a mejorar la calidad y equidad de la educación en la región, mediante la promoción de debates informados sobre temas de política educacional y reforma educativa, la identificación y difusión de buenas prácticas y la evaluación y el monitoreo del progreso educativo.

Este texto fue preparado en el marco de las actividades del Grupo de Trabajo sobre Estándares y Evaluaciones y tiene como propósito facilitar una mejor comprensión de los datos, usos, posibilidades y limitaciones de las evaluaciones estandarizadas por parte de diferentes audiencias. Su contenido puede utilizarse para la organización de talleres o seminarios, emplearse como guía inicial para cursos de evaluación dirigidos a personas interesadas en comprender mejor los resultados de evaluaciones o en participar en debates informados sobre el uso que se hace de los mismos.

Las actividades del PREAL, como las del GTEE, son posibles gracias al apoyo que brindan la United States Agency for International Development (USAID), el Banco Interamericano de Desarrollo (BID), el Banco Mundial, la International Association for the Evaluation of Educational Achievement (IEA), The Tinker Foundation, GE Foundation, entre otros.

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

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