Sistemas nacionales de evaluación del rendimiento escolar en América Latina

Santiago Cueto[1] — Grade
José Rodríguez[2] — Ministerio de Educación

El propósito del presente artículo es exponer y discutir los objetivos principales de los Sistemas Nacionales de Evaluación del Rendimiento, a través de dos preguntas cruciales: ¿qué información buscan recoger estos sistemas? y ¿qué uso se piensa dar a esa información? Además de ello intenta un breve balance y una sencilla exposición de las metas de la Unidad de Medición de Calidad Educativa para los próximos años.

En las dos últimas décadas prácticamente todos los países de Latinoamérica han llevado a cabo una o varias mediciones del rendimiento de sus estudiantes.[3]

La creación de unidades dedicadas a la evaluación del rendimiento –por lo general dentro de los Ministerios de Educación– a menudo ha sido parte de los programas de préstamos de los organismos internacionales que trabajan en la región.

Los Sistemas Nacionales de Evaluación generados a raíz de esto han sido bien acogidos al interior de los países, al punto que muchos de ellos han participado o piensan participar en estudios internacionales del rendimiento.[4]

Sin embargo, los fines de los Sistemas Nacionales de Evaluación no son uniformes en todos los países.

El propósito del presente artículo es presentar y discutir cuáles han sido los objetivos principales para estos sistemas y exponer brevemente los que busca la Unidad de Medición de Calidad Educativa del Ministerio de Educación del Perú para los próximos años. Una forma de hacerlo es tratando de responder a dos preguntas: ¿Qué información buscan recoger los Sistemas Nacionales de Evaluación? ¿Qué uso piensa darse a esa información?

¿Qué información buscan recoger los Sistemas Nacionales de Evaluación?

La mayoría de las evaluaciones del rendimiento en la región han utilizado el currículo nacional vigente como marco para el diseño de las evaluaciones.[5]

Sin embargo, el modelo de prueba empleado no ha sido uniforme. La mayoría de pruebas de la región ha utilizado un enfoque

1 Investigador Principal del Grupo de Análisis para el Desarrollo (GRADE). Dirige el equipo de asesores de la Unidad de Medición de la Calidad Educativa del Ministerio de Educación.

2 Jefe de la Unidad de Medición de la Calidad Educativa (UMC) del Ministerio de Educación.

3 ROJAS, C. y J.M. ESQUIVEL (1998). **Los sistemas de medición del logro académico en Latinoamérica.** Washington D.C.: The World Bank, LCSHD Paper # 25.

4 WOLFF, L. (1998). **Las evaluaciones educacionales en América Latina: Avance actual y futuros desafíos.** Documento Nº 11 del Programa de Promoción de la Reforma Educativa en América Latina y el Caribe. Santiago: PREAL.

5 En algunos países, como Brasil, se han creado pruebas al final de la secundaria que miden habilidades básicas que cualquier ciudadano debería tener independientemente del currículo (ver http://www.inep.gov.br/ para una descripción de las pruebas utilizadas en Brasil).

de *normas*. En este la evaluación se hace por grandes áreas, por lo general matemática y lenguaje. Lo que interesa es ordenar diferentes grupos de estudiantes –clasificados, entre otros criterios, según el tipo de gestión del centro educativo, el área de residencia o el género– de acuerdo con la proporción de aciertos en las respuestas de las pruebas.

Por la forma como se diseñan las pruebas bajo el modelo de normas, sus resultados no permiten medir los logros de aprendizaje de los estudiantes en términos absolutos. Es así, por ejemplo, que bajo este modelo no existe el equivalente a una nota aprobatoria: 80% de acierto no significa necesariamente que el estudiante domina el área, ni 20% de acierto implica que no la domine.

¿A qué se debe este resultado tan curioso? ¿Son útiles las evaluaciones por normas? La prueba por norma, como se mencionó antes, fue diseñada solamente para ordenar a los estudiantes y conocer las diferencias relativas, mas no para medir cuánto han aprendido los estudiantes.

La metodología de la construcción de pruebas por normas descansa en algunas propiedades estadísticas que le son impuestas al momento de seleccionar los ítems que se emplean en el diseño de las pruebas.

En particular, como se trata de ordenar a la población, lo que interesa es que la prueba discrimine entre los estudiantes. Por esa razón, preguntas que ofrezcan respuestas tales como *todos saben* o *ninguno sabe* no son empleadas, independientemente de la parte o porción del currículo que estas preguntas estén evaluando. Por ello, en este tipo de pruebas el acierto esperado es alrededor del 50%.

En este esquema de pruebas no tiene mayor sentido reportar el porcentaje de respuestas correctas.[6] En cambio se suele utilizar escalas abstractas –por ejemplo, promedio 500 y desviación estándar 50– para enfatizar que la variabilidad en el rendimiento de diferentes grupos debe ser vista de una manera relativa o comparativa y no asumiendo que 50% es equivalente a la nota 10 o "desaprobado".

En las nuevas propuestas curriculares, el objetivo del aprendizaje no es solamente el aprendizaje de conceptos, sino la integración de conceptos y la aplicación de éstos en situaciones de la vida cotidiana.

Un segundo modelo para la elaboración de pruebas es el de *criterios*, a través del cual se analiza el rendimiento de manera absoluta.

En este enfoque se suelen escoger ciertas capacidades (o contenidos) del currículo y generar un suficiente número de preguntas como para determinar si los estudiantes dominan las competencias evaluadas.[7]

Luego, sobre la base de la opinión de expertos en el área, se fija una línea de corte para cada competencia: por ejemplo, que de las 10 preguntas de una competencia los estudiantes puedan resolver 7 correctamente. La línea de corte indica el número mínimo de preguntas respondidas correctamente que un estudiante del grado evaluado debe saber como resultado de su educación escolar.

El reporte de resultados en este tipo de pruebas suele ser el porcentaje de estudiantes que se encuentra por encima de la línea de corte (de hecho, este enfoque es el que utilizan los docentes en el aula).

El ideal es, evidentemente, que todos los estudiantes estén por encima de esta línea. Cuando esto no ocurra se podrá determinar, entonces, en qué competencias del currículo hay deficiencias, puesto que la magnitud de tal deficiencia dependerá del porcentaje de estudiantes que están por debajo de la línea.

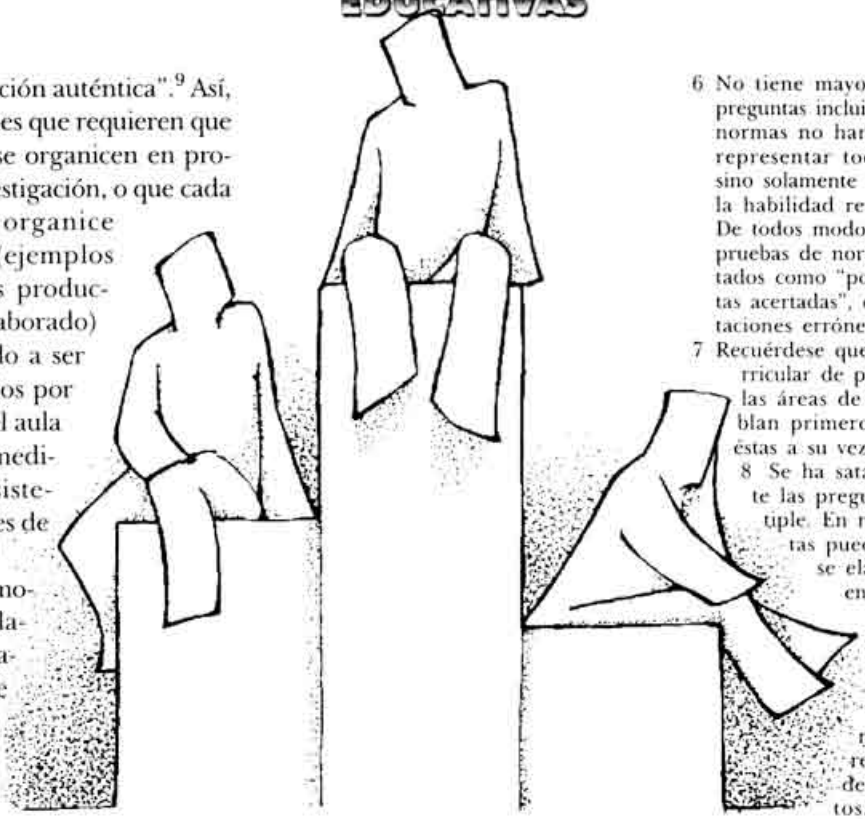
En las pruebas de criterios no se pretende obtener conclusiones acerca del "rendimiento en matemática" (o comunicación integral o cualquier otra área), sino en cada competencia que compone la evaluación, puesto que es difícil entender el conjunto y en cambio más fácil, y más valioso para la toma de decisiones pedagógicas, reportar por unidades pequeñas al interior de, por ejemplo, matemática.

Independientemente del modelo de evaluación empleado (normas o criterios) para desarrollar las pruebas, en los últimos años ha habido un gran desarrollo en cuanto al formato que deberían tener las preguntas. Esto va de la mano con cambios curriculares comunes a la región.

En las nuevas propuestas curriculares, el objetivo del aprendizaje no es solamente el aprendizaje de conceptos, sino la integración de conceptos y la aplicación de éstos en situaciones de la vida cotidiana. Las preguntas tradicionales de opción múltiple –en las que el alumno tiene que reconocer la respuesta correcta–[8] han perdido terreno frente a otros formatos en que el estudiante tiene que producir una respuesta de manera creativa.

A este enfoque se le ha denominado "evaluación de desempe-

ño" o "evaluación auténtica".[9] Así, las evaluaciones que requieren que los alumnos se organicen en proyectos de investigación, o que cada estudiante organice portafolios (ejemplos de diferentes productos que ha elaborado) han empezado a ser implementados por docentes en el aula y en menor medida por los sistemas nacionales de evaluación.[10]



De todos modos, queda claro que las evaluaciones de rendimiento deben estar orientadas a determinar qué pueden hacer los estudiantes con el conocimiento y no solamente si pueden repetir definiciones (aunque la definición de un concepto y su aplicación deberían estar ligados).

¿Qué uso se le da a la información recogida?

La segunda pregunta relevante para un Sistema Nacional de Evaluación es qué se piensa hacer con la información una vez obtenida. La respuesta a esta pregunta está definitivamente ligada a la respuesta a la primera (esto es, qué información recogen). Por ejemplo para pruebas de *normas* hay varias respuestas posibles:

Para generar competencia entre escuelas: En algunos países donde se han administrado pruebas censales[11] se han publicado los resultados promedio por centro, buscando que los padres de familia se enteren a través de estos resultados de la "calidad" del centro y, de ser los resultados bajos, presionen a los directivos escolares o incluso retiren a sus hijos del centro educativo.[12]

El supuesto es que el manejo público de la información de resultados desencadenaría un proceso de competencia que podría mejorar las condiciones bajo las cuales el servicio educativo es ofrecido.

La efectividad de este mecanismo depende –y en ese sentido, es condición necesaria también– de la existencia de condiciones para competir, es decir, que al menos existan los competidores.

Para dar incentivos a docentes: En algunos países se ha usado una combinación de diferentes factores –principalmente el promedio de los estudiantes en las pruebas de rendimiento y el porcentaje de mejora en el rendimiento desde la última prueba– para dar premios a los docentes de las mejores escuelas.[13]

El problema de este enfoque es que, para ser justo, debería medir todas las áreas que se desea influir con el sistema educativo, y esto ha sido hasta ahora imposi-

6 No tiene mayor sentido porque las preguntas incluidas en las pruebas de normas no han sido pensadas para representar toda el área evaluada, sino solamente para dar una idea de la habilidad relativa del estudiante. De todos modos algunos países usan pruebas de normas y reportan resultados como "porcentajes de respuestas acertadas", dando pie a interpretaciones erróneas de los resultados.

7 Recuérdese que en la estructura curricular de primaria, por ejemplo, las áreas de desarrollo se desdoblan primero en **competencias** y éstas a su vez en **capacidades**.

8 Se ha satanizado recientemente las preguntas de opción múltiple. En realidad estas preguntas pueden ser muy útiles si se elaboran de modo que en su resolución requieran un procesamiento complejo a nivel cognitivo (por ejemplo análisis, síntesis o solución de problemas) y no solamente repetición memorística de definiciones o eventos. Las preguntas de opción múltiple son relativamente económicas de administrar (en contraste con evaluaciones del desempeño, por ejemplo).

9 Véase, por ejemplo, *La evaluación en el marco de la Nueva Propuesta Curricular*, en **Agenda Educativa** N° 11: Lima: Foro Educativo y, BARON, M. & BOSCHEE, F. (1995). **Authentic Assessment: The Key to Unlocking Student Success.** Lancaster, PA: Technomic Publishing, Co.

10 Las evaluaciones de desempeño son más costosas que las de opción múltiple por el material y el personal involucrado. Además, a menudo requieren un trabajo de varios días, por lo que su uso en sistemas nacionales ha sido limitado

11 Una prueba censal se administra a todos los estudiantes matriculados en el sistema en los grados seleccionados mientras que una prueba muestral a un grupo representativo de estudiantes solamente. Una prueba censal es más costosa y por lo general solo se ha hecho en países con extensión relativamente pequeña y con muy buenas vías de comunicación interna (como en el caso de Uruguay, Chile y Costa Rica).

12 En Chile se publican los resultados de todos los centros educativos del país, indicando la región, el tipo de gestión y una aproximación al nivel de pobreza de las familias que asisten al centro, de modo que los padres puedan comparar centros que consideren similares en una o varias de las anteriores variables.

13 Para ello se requiere, por supuesto, que la información sea recogida a través de pruebas censales.

ble por razones técnicas y de presupuesto.

En el caso chileno, por ejemplo, las pruebas sobre cuya base se da el incentivo son lenguaje y matemática y por lo tanto el mensaje que se estaría dando a los docentes es, al menos, potencialmente peligroso: los incentivos salariales dependen de cuánto saben los estudiantes en estas dos áreas y por tanto la preocupación que los docentes deben tener por otras áreas (desarrollo físico, artístico, moral, etc.) debe ser relativamente menor.

Los organizadores de este sistema reportan sin embargo buenas experiencias y mejoras notables en el rendimiento escolar.[14]

Para identificar los centros educativos con peores rendimientos: Con las evaluaciones censales se puede ubicar a aquellas escuelas donde el rendimiento es relativamente peor, de modo que se aumente la inversión en ellas.[15]

Por otro lado, dado que la medición es en áreas (matemática, etc.), no se tiene un diagnóstico preciso de los problemas y para mejorar el centro educativo debe empezarse por elaborar uno.

Para contribuir a una discusión de los logros esperados en el Sistema Educativo: En este caso se espera que la información de los grupos relativamente rezagados (por ejemplo regiones) y avanzados contribuya a una discusión general sobre los aprendizajes que se espera dominen los estudiantes.

Este uso no se ajusta realmente al marco teórico en que se elaboraron las pruebas, porque las pruebas de normas permiten, como se dijo antes, decir quién está relativamente mejor o peor, pero no permite establecer si los estudiantes rinden a niveles aceptables.

En las pruebas de *criterios*, los usos que se suele o puede dar a la información son diferentes:

> *El interés en los últimos años ha sido estudiar* factores alterables, *es decir factores que en gran medida dependan del propio sistema educativo y no de características estructurales de la sociedad.*

Para certificar los logros adquiridos al final de ciertos niveles: En algunos países se han tomado pruebas para certificar que los estudiantes dominan los conocimientos del currículo y, por tanto, pueden recibir su diploma (por ejemplo, en Costa Rica).

El problema en este tipo de mediciones es el estrés que suele estar asociado a las pruebas[16] y la dificultad de basar un diploma en una sola medición.[17]

Para identificar logros y necesidades de aprendizaje dentro del sistema: En pruebas de criterios es posible identificar las áreas donde los estudiantes presentan rendimientos por debajo de lo esperado y planear políticas educativas sobre esa base. Por ejemplo: revisar el currículo, planificar capacitaciones docentes o reformular los materiales educativos.

Para informar sobre el desempeño de cada estudiante: En algunos sistemas, como el uruguayo, se le da a cada centro educativo los resultados particulares de sus estudiantes y se les compara con el puntaje promedio y con el de estudiantes de otros centros educativos similares.[18]

Para producir guías pedagógicas para docentes: En algunos casos se han generado guías para docentes sobre la base de los resultados de las pruebas. Estas guías analizan algunas de las dificultades de los estudiantes y sugieren a los docentes formas innovadoras de enseñar.

Estas guías se pueden hacer con pruebas de normas. Pero, como se dijo antes, las preguntas de las pruebas de normas no evalúan en detalle el aprendizaje de los estudiantes, sino que brindan información relativa del área global. Por ejemplo, es difícil determinar con pruebas de normas cuánto saben los estudiantes acerca de los números naturales, porque hay pocas preguntas sobre este tema y porque las preguntas son seleccionadas principalmente por criterios estadísticos y no pedagógicos, como se explicó antes.

Por lo tanto, el análisis en pruebas de normas se tiene que hacer a menudo a nivel de cada pregunta individual. Pero una pregunta es a menudo muy poco para hacer afirmaciones sobre una competencia.

En cambio las pruebas de criterios pueden dar una base más sólida para el desarrollo de estas guías, pues tienen un número mayor de preguntas por competencia evaluada e indican líneas de corte mínimas.

La elaboración de estas guías, como de las pruebas, debería ser un esfuerzo conjunto de las Unidades de Medición y las Unidades de Currículo y Capacitación Docente.

Tanto en pruebas de normas como de criterios es posible hacer análisis de factores asociados al rendimiento. Por ejemplo, analizando los factores del estudiante, de la familia, del centro o sistema educativo.

Para hacer estos análisis se suele administrar encuestas a estudiantes, docentes, directivos y en algunos países a padres o madres

de familia. La intención es cruzar los datos de las encuestas con los de las pruebas en busca de resultados que sugieran determinadas políticas educativas.

El interés en los últimos años ha sido estudiar *factores alterables*, es decir factores que en gran medida dependan del propio sistema educativo y no de características estructurales de la sociedad.[19]

Las pruebas de normas o criterios no son buenas o malas en sí mismas. El tipo modelo de pruebas a usar depende de los fines que se busquen para el Sistema Nacional de Evaluación. En todo caso, existen usos correctos e incorrectos de la información, como se ha intentado explicar en el presente apartado.

¿Qué estamos haciendo en el Perú?

La UMC utilizó pruebas de *normas*[20] en 1996 y 1998. En ambas se emplearon tanto preguntas de opción múltiple como de expresión escrita y respuesta abierta (en matemática).

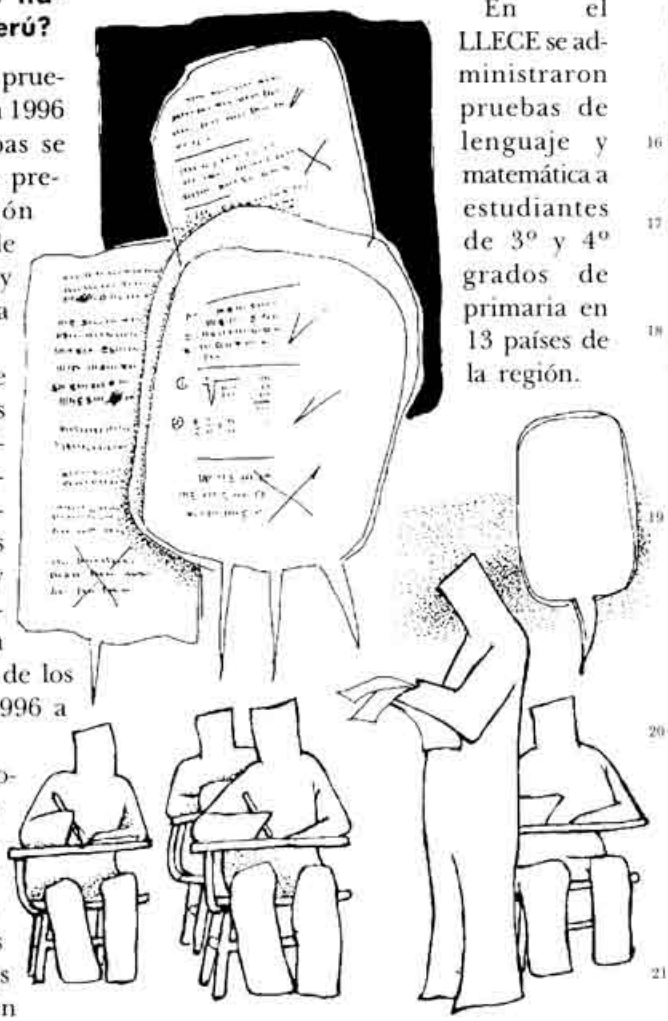
Sobre la base de estos resultados se están identificando departamentos relativamente rezagados en rendimiento y se está estudiando la evolución del rendimiento de los estudiantes de 1996 a 1998.

Estas evaluaciones servirán como línea de base para evaluar el rendimiento de los estudiantes en el futuro. Los resultados serán publicados en los próximos meses y se tiene planeado usarlos para informar y generar un debate entre los especialistas y el público.

En el análisis de los resultados, se deberá tener en cuenta los cambios curriculares. Así, las pruebas de 1996 se hicieron con un currículo que ha sido gradualmente sustituido en primaria y que está en proceso de sustitución en secundaria en los próximos años. Estos cambios dificultan la comparación del rendimiento, que debe centrarse en aquellos temas que permanecen en el currículo de una evaluación a la siguiente.

El Perú ha participado en un estudio de rendimiento escolar a nivel internacional: el Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE).[21]

En el LLECE se administraron pruebas de lenguaje y matemática a estudiantes de 3° y 4° grados de primaria en 13 países de la región.



Los resultados mostraron a Cuba muy por encima del resto de países. Luego venían, en apretada procesión, el resto de países de sur y centro América.

El Perú decidió finalmente no autorizar la publicación de sus resultados por discrepancias técnicas no resueltas oportunamente so-

[14] MIZALA, A. y P. ROMAGUERA (1999). *Sistemas de incentivos en educación y la experiencia del SNED en Chile.* Ponencia presentada en la Conferencia **Los maestros en América Latina: nuevas perspectivas sobre su desarrollo y desempeño,** organizada por el Banco Interamericano de Desarrollo en Costa Rica.

[15] Un ejemplo es Chile, que ha implantado el programa de las 900 y recientemente 1,200 escuelas. Ver RODRIGUEZ, C. (1996). Chile: *Sistema de medición de la calidad de la educación.* Presentada en el Seminario Internacional **Evaluación y estándares en la educación en América Latina**, organizado por PREAL.

[16] Muchos estudiantes no aprobarían por la presión externa o dificultades con el formato de la prueba.

[17] Para cualquier área, es preferible tener varias mediciones a lo largo del tiempo y no una sola para determinar la habilidad de una persona.

[18] Sin negar la importancia de la escuela en el aprendizaje de los estudiantes, es importante mencionar que los bajos resultados en las pruebas pueden estar vinculados a factores que trascienden el ámbito escolar. Por ejemplo, las características de la familia.

[19] El tipo de material educativo utilizado en el aula, el estilo de gestión del director, el tiempo que dedican los docentes a cubrir cada capacidad y la metodología para enseñar son, por ejemplo, factores más alterables desde el sistema educativo que el nivel de ingresos de la familia.

[20] En 1996 se evaluaron lenguaje y matemática en 4° de primaria. En 1998 se hizo lo propio con comunicación integral, lógico-matemática, ciencia y ambiente y personal social en 6° de primaria; y con lenguaje y matemática en 4° y 5° de secundaria. En base a estos resultados se elaboraron módulos didácticos para docentes.

[21] Ver página web: http://ns.unesco.cl/lab/.

bre la forma en que se preparó el informe final (publicado en 1998).

Para futuras evaluaciones dentro del Perú se tiene previsto emplear el enfoque *criterios*. El de *normas* ya generó la información que era útil generar y por tanto, luego de consultas dentro y fuera del Ministerio, incluyendo consultores internacionales, se ha decidido el paso a un enfoque de *criterios*.

Adicionalmente a las pruebas, se ha contemplado administrar otros instrumentos: escalas de actitudes sobre algunos rasgos afectivos propuestos en el currículo (utilidad, gusto por el área y autoeficacia) y encuestas a estudiantes, padres o madres de familia, docentes y directivos.

El sentido de estas encuestas es explorar algunos aspectos alterables en los que el Ministerio de Educación ha desarrollado programas en los últimos años: capacitación, gestión, reforma del currículo, uso de materiales educativos e impacto de programas especiales para Educación Bilingüe Intercultural y Escuelas de Frontera.

Los instrumentos están siendo consultados con especialistas dentro del Ministerio y con especialistas de la sociedad en general. Esperamos que los resultados sean útiles tanto para personas ligadas al quehacer educativo como para quienes están fuera de éste.

Finalmente, las pruebas de rendimiento que se administrarán buscan respuestas a preguntas que tienen que ver en parte con la calidad del sistema: cuánto saben los estudiantes y qué podría hacerse desde el propio sistema para mejorar sus aprendizajes.

Laboratorio Latinoamericano difunde estudio comparativo de lenguaje y matemáticas

El Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación, de UNESCO, dio a conocer en diciembre los resultados de su Primer Estudio Internacional Comparativo de Lenguaje, Matemática y Factores Asociados. Éste fue realizado con alumnos de Tercer y Cuarto Grado de enseñanza básica en 13 países, entregando finalmente información de 11 de ellos: Argentina, Bolivia, Brasil, Chile, Colombia, Cuba, Honduras, México, Paraguay, República Dominicana y Venezuela.

El estudio se orienta a responder cinco preguntas: ¿Qué aprenden los alumnos? ¿Cuál es el nivel al que ocurren los aprendizajes? ¿Qué competencias han desarrollado los alumnos en base de esos aprendizajes? ¿Cuándo han ocurrido los aprendizajes? y ¿Bajo qué condiciones se han producido los aprendizajes? Entre sus hallazgos, el estudio revela que:

Cuba se destaca entre los países de la Región por sus mejores resultados.

En el caso de Lenguaje, en Tercer y Cuarto Año, casi todos los países se sitúan en un rango de una desviación estándar a cada lado de la Media Regional. Argentina, Brasil, Chile y Cuba están sobre ella en Tercer y Cuarto Grado, Colombia, México y Paraguay se suman a los países ubicados sobre la Media Regional.

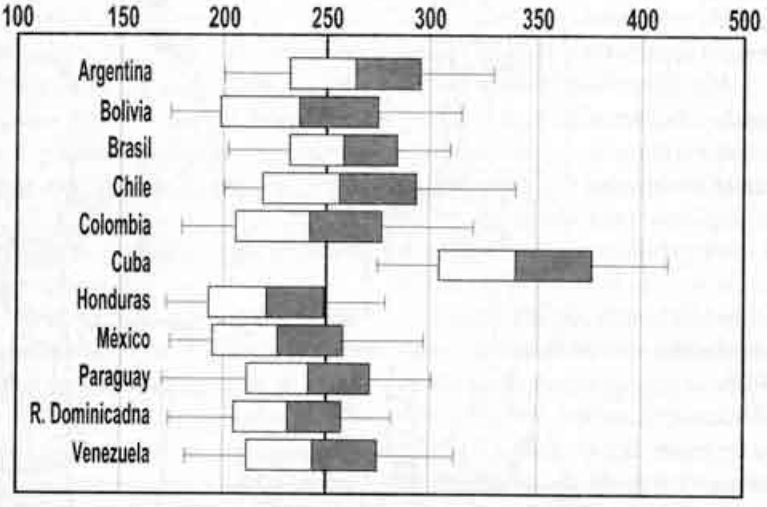
En cuanto a Matemática, en Tercer Grado, todas las medianas nacionales, a excepción de Argentina y Cuba, se sitúan bajo la Media Regional. En el caso de Matemática en Cuarto Grado, los puntajes son superiores a los que se obtienen en Tercero, ubicándose Brasil, Chile y Colombia por sobre la Media Regional, mientras que el puntaje de México coincide con ella.

La secuencia de logros de mayor a menor en casi todos los casos, es megaciudad-urbano-rural, excepto para Chile, el cual muestra que los resultados de sus escuelas urbanas superan a las de mega-ciudad. Salvo exepciones (como Colombia en Lenguaje, Tercer Grado) las escuelas rurales muestran los logros más bajos.

Por otra parte, en casi todos los casos las escuelas privadas tienen más altos logros que las públicas.

El informe de este estudio está disponible en Internet en **www.unesco.cl/lab**. Correo electrónico **laboratorio@unesco.cl**.

Resultados en lenguaje, 3° grado, Mediana cuartiles 25% y 75% y deciles 10% y 90%



Informe tomado de *PREAL Informa N° 2*, febrero de 1999.

El gráfico muestra que el rendimiento de los alumnos de la mitad más baja de Cuba es significativamente superior al rendimiento de la mitad más alta de los países que el siguen inmediatamente. En términos generales, se observan tres grupos: Cuba, cuya mediana de 343 está próxima a dos desviaciones estándar a la derecha de la Media Regional; Argentina (263) Brasil (256) y Chile (259), con medianas superiores a la Media Regional; y el resto de los países, con medianas inferiores a ella.