

***Pruebas internacionales de logros de aprendizaje:***

***Impacto sobre los procesos de mejoramiento de la calidad de la educación  
y criterios para guiar las decisiones sobre la participación del Perú y otros  
países de América Latina***

**Informe Final**

Presentado al

Consortio de Investigación Económica y Social

por

*Patricia Arregui  
J. Guillermo Ferrer*

GRADE

Lima, febrero del 2002

## **TABLA DE CONTENIDOS**

Acrónimos y abreviaciones

Introducción

Fundamentación y objetivos del estudio

Estructura del informe

Sección I

Las reformas educativas y la evaluación del logro académico en la década de los 90

Sección II

¿Qué son las pruebas internacionales?

Sección III

Beneficios y problemas de la participación en pruebas internacionales

Sección IV

Marco teórico-conceptual

Sección V

Metodología

Sección VI

Resultados

Sección VII

Discusión

Sección VIII

Recomendaciones para futuras oportunidades de participación del Perú y otros países latinoamericanos en pruebas internacionales

Recomendaciones para futuras investigaciones

## **Acrónimos y abreviaciones**

AFT: American Federation of Teachers- Federación Estadounidense de Maestros.

ALL: Adult Literacy and Lifeskills (Prueba para Adultos sobre Alfabetización y Destrezas para la Vida).

BID: Banco Interamericano de Desarrollo.

ETS: Educational Testing Services (Servicios de Evaluación Educacional), Universidad de Princeton, EEUU.

ICFES: Instituto Colombiano de Fomento de la Educación Superior.

IDECE: Instituto para el Desarrollo de la Calidad Educativa, Ministerio de Educación de Argentina.

IEA: International Association for the Evaluation of Educational Achievement (Asociación Internacional para la Evaluación del Logro Académico)

IRT: Item Response Theory- Teoría de Respuesta al Item.

LLECE: Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación (UNESCO-OREALC).

NCES: National Council for Education Statistics- Consejo Nacional de Estadística de la Educación, EEUU.

NCTM: National Council of Teachers of Mathematics- Consejo Nacional de Profesores de Matemáticas, EEUU.

NSF: National Science Foundation- Fundación Nacional de Ciencias, EEUU.

OCDE: Organización para la Cooperación y el Desarrollo Económico.

OEI: Organización de Estados Iberoamericanos.

ONG: Organización no-gubernamental.

OREALC (UNESCO): Oficina Regional para América Latina y el Caribe.

PIRLS: Progress in International Reading Literacy Study (Estudio Internacional sobre los Progresos en Alfabetización y Habilidades de Lectura).

PISA: Programme for International Student Assessment (Programa Internacional de Evaluación de Estudiantes).

PNUD: Programa de las Naciones Unidas para el Desarrollo.

PROMEDLAC: Comité Regional Intergubernamental del Proyecto Principal de Educación en América Latina y el Caribe (UNESCO).

REDUC: Red Latinoamericana de Información y Documentación en Educación.

SIMS: Second International Mathematics Study (Segundo Estudio Internacional de Matemáticas).

SACMEQ: Southern Africa Consortium for Monitoring Educational Quality (Consortio de Africa del Sur para la Evaluación de la Calidad Educacional).

SIMCE: Sistema de Medición de la Calidad Educativa. Ministerio de Educación de Chile.

SMART: Science and Mathematics Achievement Required for Tomorrow.

TIMSS: Third International Mathematics and Science Study (Tercer Estudio Internacional en Matemáticas y Ciencias, 1995).

TIMSS-R: Repetición de TIMSS (1998).

UMRE: Unidad de Medición de Resultados Educativos, Administración Nacional de Educación Pública, Uruguay.

## **INTRODUCCIÓN**

### **Fundamentación y objetivos del estudio**

En los últimos años han surgido diversas recomendaciones, provenientes de organismos internacionales y regionales, sobre la conveniencia de la participación de países latinoamericanos en pruebas internacionales de logro académico. Esas recomendaciones suelen sustentarse con el argumento de que los datos comparativos sobre el rendimiento estudiantil son útiles para informar a los países sobre la calidad relativa de sus sistemas educativos y, consecuentemente, sobre su competitividad en el mercado global de bienes y servicios. Están también basadas en el argumento de que las pruebas internacionales garantizan más alta calidad técnica y eficiencia que los sistemas nacionales de medición y que pueden contribuir significativamente al desarrollo de capacidades locales en el campo de la evaluación.

La utilidad efectiva de las comparaciones internacionales de logro académico para informar políticas educativas no es, sin embargo, algo sobre lo cual exista total acuerdo en la literatura actual. Tampoco se ha recogido ni sistematizado suficiente evidencia empírica sobre la difusión que los países han dado a la información recogida en las pruebas internacionales ni sobre la manera en que han utilizado esa información para promover el mejoramiento de sus políticas, planes, programas y prácticas educativas.

Ante esta falta de evidencia, y ante la posibilidad de que el Perú y otros países de la Región comiencen a privilegiar la participación en pruebas internacionales sin la debida consideración previa sobre la necesidad de consolidar el desarrollo de sus propios sistemas y capacidades de evaluación y medición, se consideró oportuno iniciar una indagación sistemática sobre los riesgos y beneficios de esta participación y sobre las condiciones técnicas y políticas en que debiera realizarse. En la medida de lo posible, esa indagación se focalizará en la experiencia vivida por algunos países que ya han participado en pruebas internacionales de logro académico.

El propósito último que se persigue con este estudio es el de contribuir a que dicha participación -- de resultar recomendable o, de alguna manera, ineludible -- pueda servir efectivamente para el mejoramiento de la gestión y resultados de los procesos educativos en el país, así como en otros países de la Región que recién se incorporan a este tipo de programas, a la vez que para que se tomen medidas para minimizar impactos negativos que dichas actividades pudieran eventualmente acarrear.

## **Estructura de este informe**

Dada la complejidad del tema investigado, y considerando que una gran parte de los lectores potenciales de este trabajo se enfrentan a este tema por primera vez, se ha decidido organizar este informe de la siguiente manera:

La **Sección I** presenta, sobre la base de la revisión bibliográfica inicial, algunos aspectos del contexto histórico sobre reformas educativas y de desarrollo de sistemas de evaluación dentro del cual se inscribe la participación de muchos países dentro y fuera de la Región en pruebas internacionales de logro académico. La **Sección II** presenta una descripción inicial de algunas de las pruebas internacionales más importantes en la actualidad y resume sus principales características. En la **Sección III** se resumen algunos de los beneficios y problemas de la participación en pruebas internacionales más frecuentemente citados por la literatura especializada. El marco teórico-conceptual se presenta en la **Sección IV**, y en la **Sección V** se describe la metodología seguida en este estudio. La **Sección VI** presenta los resultados del trabajo de campo, clasificados por tipo de prueba y organizados de acuerdo al marco conceptual seleccionado. En la **Sección VII**, en forma de discusión, se plantean algunas generalizaciones sobre la experiencia de participación de países latinoamericanos en pruebas internacionales. En la **Sección VIII**, por último, se ofrece un conjunto de recomendaciones para orientar las decisiones de los países de la Región sobre futuras participaciones en este tipo de pruebas.

## **I. LAS REFORMAS EDUCATIVAS Y LA EVALUACIÓN DEL LOGRO ACADÉMICO EN LA DÉCADA DE LOS 90.**

Los esfuerzos en el mundo por llevar a cabo reformas educativas en las últimas dos décadas han sido motivados, principalmente, por un creciente interés político en asegurar no solamente cobertura y acceso universal a la educación, sino también el aprendizaje efectivo, por parte de *todos* los alumnos, de los contenidos curriculares propuestos por esas reformas.

Este renovado interés por la calidad de la educación, entendiéndose por calidad más y mejores aprendizajes, se inscribe en un contexto de fuertes cambios políticos y económicos de orden mundial, donde las reglas de libre mercado, la globalización y las nuevas tecnologías de información imponen una redefinición y ajustes de las metas y objetivos de la educación pública, así como del rol del Estado para la consecución de esas metas. Estos cambios se caracterizan por una serie de objetivos que Carter y O'Neill (1995), en Ball (1998), definen como constitutivos de lo que ellos llaman una "nueva ortodoxia":

1. Mejorar las economías nacionales por medio del establecimiento de vínculos más robustos entre escolarización, empleo, productividad y mercado.
2. Mejorar los resultados de aprendizajes relacionados a las competencias y destrezas que requiere el mundo laboral.
3. Lograr un mayor y más directo control de los sistemas educativos nacionales sobre los contenidos curriculares y sobre la evaluación.
4. Reducir los costos de gobierno en educación.
5. Ampliar la contribución de la comunidad a la educación, mediante su participación en la toma de decisiones escolares y mediante la presión de la libre elección de mercado<sup>1</sup>.

Paralelamente, muchos países, entre ellos los latinoamericanos, buscan por medio de las reformas curriculares y de gestión fomentar y consolidar valores democráticos, de participación ciudadana y de respeto por la diversidad. En este sentido, es importante notar que la búsqueda del control más directo sobre los contenidos curriculares al que se aludía anteriormente se ha fortalecido en los países que han tenido sistemas educativos tradicionalmente descentralizados, mientras que en aquéllos donde la prescripción curricular ha sido históricamente centralizada los esfuerzos se han concentrado más bien en la flexibilización del currículo en el nivel local y escolar. En cualquiera de los dos escenarios, sin embargo, se percibe la necesidad de realizar evaluaciones centralizadas que permitan medir el logro académico de *todos* los alumnos bajo un mismo parámetro y en relación con contenidos curriculares a los que *todos* los estudiantes deberían poder acceder.

Estas tendencias vienen de la mano de nuevas formas de administración y gestión que ponen en el centro de la discusión la necesidad de mejorar la eficiencia de los sistemas educativos y de fomentar la responsabilización social y profesional por los resultados de la educación. Estas formas de gestión, usualmente denominadas de “gerenciamiento”, suponen el uso de estilos y herramientas de gestión diferentes a las tradicionalmente utilizadas por el sector público, y de alguna manera introducen matices a los cinco objetivos anteriormente mencionados. Son ejemplos de esas nuevas formas o estilos de gestión la delegación de responsabilidad y poder de decisión a las instituciones proveedoras de servicios (escuelas), así como el control “a distancia” de la calidad en la provisión de esos servicios por parte del poder central. De esta manera, el Estado reemplaza la intervención directa y la prescripción por la determinación de objetivos, la comparación y la responsabilización (Lyotard, 1984, en Ball, 1998).

Dentro de este esquema de gerenciamiento, una de las herramientas principales de gestión educativa que ha cobrado protagonismo en muchos países del mundo, y especialmente en los de habla inglesa, es la evaluación comparativa y cuantitativa de los logros de aprendizaje de los alumnos, y la medición del impacto de diversas variables escolares y extraescolares sobre esos logros. Los países que disponían de sistemas nacionales de evaluación desde antes de los años 80 reforzaron su aplicación y los vincularon más directamente a sus nuevas metas curriculares. Los países que no disponían de tales sistemas de evaluación, como fue el caso de la mayor parte de los países latinoamericanos, desarrollaron esos sistemas junto con el diseño e implementación de las reformas educativas de los años 80 y 90. El impulso por la creación de estos sistemas de evaluación en los países en desarrollo estuvo influido por la presión de algunas agencias de cooperación internacional, que condicionaban los préstamos para educación al desarrollo de sistemas nacionales de evaluación que permitieran medir los resultados de esas millonarias inversiones en reformas educativas. El Banco Mundial, como parte de la negociación por los créditos para educación, jugó un rol particularmente importante en la difusión de un modelo de gestión consistente con las actuales tendencias de globalización y de libre mercado (Jones, 1998, en Ball, 1998; Torres, 1999).

Los cambios económicos de las últimas dos décadas en el mundo también han dado lugar a una creciente preocupación de los países por su capacidad para competir eficazmente en los mercados internacionales. Para poder competir y prosperar, según se ha argumentado en realidad ya desde el siglo XVIII con la publicación del tratado sobre “la naturaleza y las causas de la riqueza de las naciones” (Smith, 1776), es menester invertir en capital humano. Desde entonces, la educación ha sido vista como el principal medio para incrementar ese capital y fomentar así el desarrollo y la posibilidad de competir comercialmente en el escenario global.

A las preocupaciones por la prosperidad económica, se suman metas de conciencia y participación ciudadana, así como del desarrollo emocional y actitudinal de los alumnos. Esto es particularmente visible en América Latina, donde las reformas educativas de los últimos 20 años se han emprendido en contextos políticos de democratización y pacificación, y en algunos países de Asia como Japón y Corea donde se ha hecho evidente la necesidad de complementar el enfoque tradicionalmente academicista de la educación con aspectos del desarrollo moral, afectivo y social de los estudiantes.

La responsabilidad que se le adjudica a la educación por incrementar el capital humano de las naciones se manifiesta actualmente en las múltiples iniciativas de reforma educativa en el mundo. Casi todas estas reformas, si bien muestran grandes diferencias de país en país y de región en región, guardan como principal semejanza los esfuerzos por actualizar permanentemente los currículos y por medir, como se explicó anteriormente, la implementación y aprendizaje efectivo de los contenidos de esos currículos.



Junto con el desarrollo y perfeccionamiento de sistemas nacionales de medición de aprendizajes, ha crecido notablemente el interés por las comparaciones internacionales de logro académico. Según Bempechat et al (2000) “el último cuarto de siglo ha sido testigo de un incipiente interés por las comparaciones internacionales de logro estudiantil”. Agregan las autoras que “la fascinación con el logro académico en diferentes naciones ha sido motivada por los rápidos avances tecnológicos que han transformado la economía mundial [...] Los Estados-nación expresan su preocupación por su capacidad para competir en un mundo que se vuelve más y más pequeño”.

El caso de Estados Unidos es paradigmático en cuanto al impacto que las pruebas internacionales comparadas pueden tener sobre las políticas educativas de una nación. Los resultados en matemáticas y ciencias de los estudiantes norteamericanos en pruebas internacionales, por debajo del promedio mundial y muy lejos de las potencias asiáticas, alarmaron a los líderes políticos de los inicios de los 80, quienes no dudaron en calificar a los Estados Unidos como “una nación en riesgo” (National Commission on Excellence in Education, 1983). La declaración hacía referencia al peligro que corrían los EEUU de perder liderazgo mundial si su sistema educativo público no lograba mejores resultados de aprendizaje en ciencias y matemáticas, y desencadenó una serie de medidas de reforma curricular y de gestión a lo largo y ancho del país, con el objetivo explícito de ocupar el primer puesto en los *rankings* internacionales de logro académico hacia el año 2000.

Pero las comparaciones internacionales no tienen solamente un valor competitivo, o de “*ranking*”. Además del interés, muy legítimo por cierto, que pueda tener cada país en saber su posición relativa a otros países en cuanto a logro académico, existen razones conceptuales que justifican las comparaciones entre países como un instrumento metodológico que ofrece mayor información que la que puede ofrecer un análisis intra-nacional. De acuerdo a Joseph Farrell (1979) no puede realizarse ningún tipo de generalización científica sobre educación si no es a través de un estudio comparativo. El autor sostiene que “los datos comparativos son esenciales para establecer la credibilidad de las proposiciones” y que “hasta que no se conoce la relación entre universos particulares y generales, ningún estudio es suficiente para probar una proposición”.

La teoría sobre educación comparada es extensa y controvertida, y los argumentos sobre el valor y validez de las comparaciones internacionales aún son materia de acaloradas discusiones en los medios académicos. Sin embargo, persiste el interés de los investigadores sociales, así como de los formuladores de políticas públicas alrededor del mundo, en realizar comparaciones, cada vez más precisas y sofisticadas, sobre diferentes aspectos de política social. Éste ha sido el caso de las pruebas internacionales de logro académico, que se siguen aplicando en una creciente cantidad de países cada año y que se van

perfeccionando gradualmente en lo que se refiere a la especificación consensuada de sus objetivos, así como a las técnicas de recojo y análisis de la información. Por ejemplo, y como se explicará más adelante, entre las razones más comúnmente citadas para llevar a cabo comparaciones internacionales en educación, se encuentra la posibilidad de hallar mayor varianza en las variables estudiadas que la que puede hallarse en un estudio intra-nacional, lo que permite llegar a conclusiones más claras y más robustas sobre los efectos de determinadas variables o factores sobre el rendimiento estudiantil.

Según la literatura actual, son varios los objetivos que suelen perseguirse mediante la participación en pruebas internacionales. Theisen y colaboradores (1983) citan tres objetivos principales para esa participación:

1. Comparar el logro académico relativo entre países y por asignatura;
2. Identificar implicancias de política educativa para un país a través de los determinantes de logro hallados en otros países; y
3. Re-evaluar las prioridades de inversión dentro de un país para elevar la calidad de los aprendizajes.

Husén (1987), cuando discute el impacto de la IEA (International Association for the Evaluation of Educational Achievement) sobre las políticas educativas en los países participantes, distingue dos categorías de “clientes”: 1) formuladores de políticas (mayormente en el nivel central), y 2) ejecutores de esas políticas (técnicos y educadores en el nivel central y local). Según Husén, los primeros suelen estar más interesados en datos y hallazgos de investigación referidos a la estructura del sistema de educación, mientras que los últimos prestan más atención a la información recogida sobre procesos de enseñanza-aprendizaje, prácticas y metodología de aula, materiales educativos y evaluación de aprendizajes en el aula.

Para muchos investigadores y tomadores de decisiones sobre políticas educativas, tal como fuera expresado en una conferencia del Centro Nacional de Estadística Educativa de los EEUU- NCE (Husén, 1987), la participación en pruebas internacionales constituye una “intervención en sí misma”, independientemente de los resultados del estudio, y alerta a educadores e investigadores sobre problemas que podrían no haber considerado anteriormente.

## **¿QUÉ SON LAS PRUEBAS INTERNACIONALES?**

Las pruebas internacionales comparadas son instrumentos de medición estandarizados que se aplican en varios países simultáneamente, en grados o

grupos etéreos seleccionados, y que recogen información sobre logros de aprendizaje conceptual y procedimental en diferentes disciplinas, y sobre los factores contextuales que se presume influyen en el logro académico de los estudiantes. Las pruebas suelen incluir ítemes de opción múltiple, preguntas abiertas y/o demostraciones prácticas, como por ejemplo de laboratorio en el caso de las ciencias naturales. Entre los factores contextuales más importantes que suelen estudiarse aparecen la inversión per cápita en cada sistema nacional, las características familiares y socioculturales de los alumnos, la infraestructura escolar, la preparación académica de los docentes, la cobertura curricular efectuada en clase en las áreas disciplinares evaluadas, y las actitudes y valores de los alumnos con relación a esas áreas.

Las pruebas pueden ser curriculares o “acurriculares”, dependiendo respectivamente de su vinculación o no vinculación a los currículos de los países participantes. Cuando las pruebas son curriculares, como en el caso de TIMSS o LLECE, los instrumentos se diseñan sobre una matriz curricular común que representa los contenidos básicos que se ofrecen en los niveles educativos a ser evaluados en casi todos los países participantes. A estos contenidos comunes se suman contenidos considerados necesarios para una educación de calidad en el siglo XXI, según la opinión de profesionales y académicos internacionales especializados en las áreas disciplinares evaluadas. En el caso de las pruebas “acurriculares”, tales como PISA, los diseñadores del estudio establecen estándares de contenido académico disciplinar que se consideran necesarios para una efectiva inserción social y laboral en el mundo de hoy, independientemente de los contenidos de los currículos de los países participantes en la prueba comparada.

Algunas de estas pruebas se han aplicado una única vez, a modo de diagnóstico o como ejercicio de investigación. Actualmente, se observa una tendencia al diseño de pruebas de aplicación cíclica de manera tal que se pueda seguir el rendimiento de una cohorte de estudiantes cada cierta cantidad de años, o de un grado escolar a través de los años, y así poder establecer relaciones longitudinales con respecto al desempeño académico.

Estas son pruebas diseñadas y coordinadas por agencias internacionales de evaluación, y con cierto grado de participación por parte de las unidades de medición de la calidad educativa de los países. Algunas de estas agencias son privadas e independientes, de un perfil básicamente académico, tales como la IEA (Asociación Internacional para la Evaluación del Logro Académico), o ETS (Servicios de Evaluación Educacional). Otras agencias son intergubernamentales y en ellas están representadas las máximas unidades políticas o técnicas de administración o evaluación educativa de los países miembro. Ejemplo de estas últimas son la OCDE (Organización para la Cooperación y el Desarrollo Económico), el Laboratorio Latinoamericano para la Evaluación de la Educación (LLECE), dependiente de UNESCO-OREALC, o SACMEQ (Consortio de Africa del Sur para la Evaluación de la Calidad

Educacional). Las pruebas de la IEA o de la OCDE son de cobertura mundial, mientras que otras, como las del LLECE o SACMEQ, son de cobertura regional. En la actualidad existen no menos de 10 pruebas internacionales que miden el logro estudiantil en diferentes áreas de conocimiento, tal como se resume en forma ilustrativa en el *Anexo A*.

Originadas como ejercicio principalmente académico, las pruebas internacionales buscan el desarrollo de nuevas metodologías de investigación comparada y enfoques analíticos (Forster, 2000); actualmente, sin embargo, se tiende al desarrollo de estas pruebas como un instrumento que puede proveer importantes insumos de información para el diseño y mejoramiento de políticas educativas orientadas a elevar la calidad de los aprendizajes. Para ese fin, algunas de estas pruebas ofrecen lo que en este informe daremos a llamar “opciones nacionales”, es decir la posibilidad de que cada país utilice el operativo de aplicación de instrumentos a la muestra nacional para medir subpoblaciones específicas que por alguna razón ameriten una medición más enfocada, o de áreas de conocimiento particulares y diferentes de las que evalúa la prueba general. Así, un país podría hacer uso del operativo internacional para evaluar, por ejemplo, un grupo étnico diferente o agregar ítemes de contenidos curriculares nacionales no contemplados en la prueba internacional.

Los resultados de las pruebas internacionales pueden ser reportados y difundidos por las agencias mismas que dirigen los estudios, por las agencias nacionales que coordinan los estudios en los países participantes, por investigadores independientes o por la prensa (Forster, 2000; Macnab, 2000). Los datos que las pruebas arrojan suelen ser cuantiosos y por esa razón se eligen diferentes formas de reportarlos, de acuerdo a las audiencias a quienes están dirigidos. Algunos reportes presentan un *ranking* de acuerdo al logro relativo de los países participantes, con relación al total del examen o a grupos de ítemes seleccionados para el reporte; también suele mostrarse la distribución de los resultados con relación a la media aritmética de cada país (dispersión), el logro de diferentes subpoblaciones, o las relaciones entre logro académico y diferentes variables contextuales relevadas por el estudio (Forster, 2000).

Algunas de las formas en que suele usarse la información recogida son: la motivación para mejorar políticas a partir de la simple comparación con el logro de otros países; la promoción de políticas de mejoramiento a través de la difusión masiva de los resultados; el establecimiento de nuevos estándares académicos de acuerdo al desempeño de alumnos en otros países; la identificación de los factores escolares y extraescolares que muestran un efecto positivo en el logro de aprendizajes; y el estímulo para las investigaciones dentro del país mediante la profundización de los resultados del estudio internacional (Forster, 2000; Macnab, 2000).

## **BENEFICIOS Y PROBLEMAS DE LA PARTICIPACIÓN EN PRUEBAS INTERNACIONALES**

El *valor* de las comparaciones de logro académico entre países o entre regiones es materia de discusión permanente entre investigadores. Los argumentos que se esgrimen a favor y en contra de esas comparaciones suelen ser eminentemente de carácter técnico, aunque también entran en juego factores de orden político e ideológico. A continuación se resumen las razones más frecuentemente citadas por la literatura especializada con relación a las ventajas y desventajas (o problemas) de la participación en pruebas internacionales.

### **Beneficios**

- La construcción de pruebas requiere de los países participantes una revisión cuidadosa de sus currículos y de su comparación con los de otros países (Beaton et al, 1999), lo cual puede conducir a su actualización y mejoramiento. También motiva a educadores e investigadores a estudiar problemas del sector que no se habían considerado antes (Husén, 1987).
- Se desarrolla la capacidad local para la medición de aprendizajes (Greany y Kellaghan, 1996; Torney-Purta, 1990; Bloom, 1974), en la medida en que educadores e investigadores de la educación se ven llamados a interactuar con especialistas internacionales y así se mantienen actualizados (Beaton et al, 1999; Mislevy, 1995).
- Se conoce la efectividad de las políticas implementadas por otros países (Beaton et al, 1999)
- Los resultados comparados internacionalmente atraen la atención de los medios de comunicación, movilizan fuerzas políticas (Greany y Kellaghan, 1996) y suscitan debates sobre las implicancias prácticas de los resultados obtenidos (Husén, 1987; Husén y Tuijnman, 1994, en Kellaghan, 1996).
- Además del logro académico propiamente dicho, existe una gran variedad de variables educativas que pueden compararse, tales como el clima escolar, la disponibilidad relativa de materiales educativos o los efectos de la educación preescolar. Estas variables contribuyen a comprender cuáles son los efectos relativos de los diferentes factores sociales y escolares que afectan el rendimiento (Husén, 1987; Torney-Purta, 1990).
- Los gobiernos pueden incluir en estos estudios preguntas e ítemes que de otra manera deberían hacer por su propia cuenta, a través de sus sistemas nacionales (Husén, 1987).

- La dispersión que se halla en los estudios comparativos internacionales suele ser mucho más fuerte y evidente que la que se puede observar en estudios intra-nacionales, y por lo tanto se puede indagar más profundamente sobre los efectos de diferentes variables sobre los resultados de aprendizaje (Husén, 1987).
- La construcción colectiva y aplicación de pruebas puede promover prácticas profesionales democráticas debido al tipo de procedimientos y metodologías utilizadas (Torney-Purta, 1990).
- Algunas pruebas, como TIMSS, utilizan diseños jerárquicos para la recolección de datos, lo cual permitiría, asumiendo que exista suficiente varianza en los datos totales, analizar cuánto de la varianza se debe a diferencias entre países o a diferencias entre estudiantes (Bempechat et al, 2000). Estos análisis pueden contribuir a revelar cómo algunas políticas similares pueden tener efectos diferenciales en los países, dependiendo de otras variables complementarias. Un ejemplo de esto es el efecto positivo que la tarea en el hogar puede tener en el logro académico de los estudiantes chinos, mientras que en Perú el efecto puede ser mucho menor debido al alto índice de analfabetismo de los padres.
- Si bien en el proceso participan técnicos internacionales, los informes de resultados son producidos por investigadores nacionales que conocen a fondo los problemas de su país; por esta razón se espera que exista una interacción entre investigadores y educadores por lo menos igual a la que puede existir en el caso de pruebas intra-nacionales, de manera tal que la información recabada y difundida provoque respuestas constructivas a los problemas detectados por el estudio (Bloom, 1974).

### **Problemas y limitaciones**

- Los resultados de los estudios, en particular cuando se presentan como *ranking* de logro entre países, suelen desatar comentarios y críticas por parte de políticos, medios de comunicación y público en general, de manera tal que predomina el efecto llamado “carrera de caballos” o “juegos olímpicos”, donde sólo importa la posición relativa de logro en el escenario internacional y no la generalmente abundante información sobre factores sociales, escolares y personales asociados al rendimiento (Greaney y Kellaghan, 1996). La prensa, particularmente, tiende a desconocer los aspectos más complejos y auténticamente informativos de los resultados y sólo destaca las posiciones en el *ranking*, a menudo con títulos dramáticos y escandalosos (Husén, 1987; Kellaghan, 1996).

- Si bien las pruebas internacionales utilizan diferentes formatos, algunos de ellos innovadores, los ítemes más numerosos continúan siendo diseñados en forma de opción múltiple. Aun cuando estos ítemes, si están debidamente contruidos, pueden evaluar habilidades cognitivas complejas, persisten críticas y dudas sobre su validez para ese fin (Greaney y Kellaghan, 1996). De la misma manera, se ha criticado a los estudios internacionales por utilizar un paradigma básicamente cuantitativo, positivista, alejado de la realidad de los países en desarrollo y carentes de metodologías que fomenten la construcción cooperativa y democrática de las pruebas (Husén, 1987).
- La complejidad de los estudios, sumada a la escasa capacidad que suele existir entre el público para interpretar los numerosos datos que derivan de los estudios comparativos, determina que la información que finalmente se difunde resulte ambigua o contradictoria, que un mismo dato pueda ser utilizado para argumentos antagónicos (dependiendo de la postura política), o que sencillamente se malinterpreten los datos y se llegue a conclusiones erróneas (Husén, 1987). Los indicadores, interpretados aisladamente de su contexto cultural y estructural, pueden conducir a interpretaciones y acciones de política equivocadas (Burstein, Oakes y Guiton, 1992, en Schmidt y McKnight, 1995).
- La validez de los estudios se ve amenazada por las grandes diferencias culturales entre los alumnos de diferentes países, que hacen que las “condiciones” en que se toman las pruebas no sean iguales en cada país. Entre estas diferencias destacan los diferentes grados de familiaridad que los alumnos puedan tener con las pruebas estandarizadas, los estilos de respuesta (Ej. tendencia a omitir o a adivinar respuestas), y las motivaciones que puedan tener los alumnos para hacer un esfuerzo significativo por resolver las preguntas (Mislevy, 1995).
- A menudo se toman decisiones de política sobre la base de interpretaciones de causalidad entre las variables estudiadas, cuando en realidad esa causalidad sólo podría confirmarse mediante estudios experimentales controlados (Platt, 1975, en Mislevy, 1995).
- La información cualitativa recogida por los estudios comparativos no siempre resulta suficiente para comprender las variaciones que existen al interior de cada cultura y así comprender mejor cómo las creencias y percepciones de los alumnos afectan diferencialmente su rendimiento (Bempechat et al, 2000).
- Si bien se ha propuesto el uso de las pruebas internacionales para fines de responsabilización de los gobiernos y funcionarios, aún no existe un modelo claro para operacionalizar una propuesta de este tipo, es decir definiciones claras sobre quién será responsable, o quién tendrá el poder o la autoridad

para modificar y mejorar el rendimiento académico de los alumnos. En general, y como sucede con muchos sistemas de evaluación nacionales, se ha dado poco espacio a la discusión sobre el uso exacto que se dará a las pruebas, y puede pretenderse, erróneamente, que la misma prueba sirva múltiples propósitos cuando en realidad esos propósitos pueden ser incompatibles (Kellaghan, 1996).

- Hasta la fecha, ni las agencias evaluadoras ni académicos especializados han llevado a cabo un registro sistemático de los usos que se ha dado en diferentes países a los resultados de las evaluaciones. Kellaghan (1996) resalta que los informes que actualmente existen presentan información limitada y relativamente superficial, y por lo tanto no se puede identificar los mecanismos exactos por medio de los cuales la información [de la IEA] se “traduce” en acciones de política. Aunque este no sea un problema inherente a las pruebas mismas, históricamente demuestra que los emprendimientos de evaluación internacional no siempre logran informar eficazmente a las audiencias o clientes principales.

El balance entre las ventajas y limitaciones que presentan las pruebas internacionales pareciera indicar, en principio, que la participación en estos proyectos puede ser valiosa para los países, inclusive para los países en desarrollo (Greaney y Kellaghan, 1996), siempre y cuando se tomen ciertos recaudos con respecto a las características técnicas de las pruebas y a las condiciones políticas e institucionales para su aplicación, incluyendo la previsión de mecanismos para la correcta interpretación y difusión de los resultados. La literatura actual propone una variedad de “condiciones” bajo las cuales debiera concretarse la participación de un país en pruebas internacionales comparadas<sup>ii</sup>. Esas condiciones, que en esta sección se presentan como nota marginal, serán tomadas en cuenta para sustentar algunas de las recomendaciones que se efectúen hacia el final de este informe.

## **MARCO TEÓRICO-CONCEPTUAL**

El presente estudio fue estructurado sobre la base de dos marcos teóricos principales. El primero permite encuadrar la información recogida sobre evaluaciones internacionales comparadas dentro de un modelo *secuencial* o *cronológico* de política y cambio educativo. Este modelo secuencial, que se explica a continuación, se seleccionó porque representa la lógica con que muchos de los informantes en la Región justifican la participación en pruebas internacionales y porque permitió anticipar algunos de los potenciales problemas que obstaculizan la implementación efectiva de los cambios deseados.

La perspectiva *secuencial* o *cronológica*, permite observar los contextos y procesos de política educativa, y evaluativa en particular, como una sucesión de instancias lógicamente ordenadas. Los autores que aplican esta perspectiva



utilizan diferentes nomenclaturas y números de fases para su descripción<sup>iii</sup>, aunque en conjunto coinciden en la existencia de las siguientes etapas en el desarrollo de políticas educativas centradas en el logro de aprendizajes:

a) Reposo, o situación de *status quo*, en que tanto los proveedores como los beneficiarios se sienten conformes con la calidad del servicio educativo, concretamente con los aprendizajes logrados por los alumnos.

b) Descontento: En esta fase, algún actor, o grupo de actores del sistema percibe, o demuestra con algún tipo de evidencia, que el servicio educativo no está proveyendo suficientes oportunidades de aprendizaje y que por lo tanto el rendimiento de los estudiantes es inadecuado.

c) Iniciativa y adopción: Ese mismo grupo, u otros interesados, deciden proponer cambios en el sistema, impulsan el diseño de programas y estrategias comprehensivas que apuntan a elevar el nivel de los aprendizajes, buscan apoyo político y financiero para asegurar la viabilidad inicial de esos programas, y en algunas ocasiones buscan sensibilizar o movilizar a otros actores clave del sistema.

d) Implementación: Los programas se ponen en práctica, buscando articular nuevos elementos (currículo, textos, capacitación docente, etc.) con recursos ya existentes.

e) Evaluación de proceso y resultados: En esta fase se monitorea el proceso de implementación para poder realizar ajustes a los nuevos programas en curso (evaluación formativa), o se decide medir los resultados e impactos de esos programas (evaluación sumativa). La evaluación sumativa requiere, en el contexto del estudio que nos ocupa, la aplicación de instrumentos que permitan medir objetivamente los logros de aprendizaje de los alumnos.

f) Retroalimentación: La información recabada mediante las evaluaciones se utiliza para mejorar los insumos o procesos de implementación, y así elevar las posibilidades de que los programas en curso tengan mayor impacto en los aprendizajes.

En el contexto de esta investigación, las etapas de política mencionadas estarán referidas específicamente a las políticas de diseño, implementación, evaluación y mejoras del currículo prescrito. Simultáneamente, se ha concebido el proceso de participación en pruebas internacionales como un continuo donde se suceden fases similares de desarrollo, desde la percepción de problemas de aprendizaje y la decisión de utilizar una herramienta de evaluación comparada para verificarlo, hasta el uso de la información de esa evaluación para el mejoramiento de las condiciones para un mejor logro académico.

Para fines analíticos, y asumiendo teóricamente que las fases de política curricular y las de evaluación comparada internacional son más o menos sincrónicas, encuadramos el análisis del ciclo de participación en pruebas internacionales de cada país dentro del ciclo de política, o reforma curricular, en el cual se inscribe históricamente. De esta manera se intenta ver, por ejemplo, en qué medida los procesos de implementación de pruebas internacionales (elaboración y aplicación de instrumentos) contribuyen al desarrollo de capacidades profesionales para la implementación y ajuste de los programas curriculares que se pretende evaluar. Dado que en la realidad cada país ha participado en más de una prueba, en diferentes momentos, y acompañando tiempos distintos del desarrollo curricular, este tipo de análisis no deja de ser solamente un intento por contextualizar los ciclos de evaluación comparada, sin pretender establecer relaciones exactas y sincronizadas entre ambos procesos.

De esta manera, los resultados de este estudio se presentarán conforme al siguiente esquema conceptual sobre política curricular y evaluación comparada internacional:

<b>Ciclo de política curricular</b>	<b>Ciclo de evaluación internacional</b>
<i>Reposo</i> , o situación de <i>status quo</i> , en que tanto los proveedores como los beneficiarios se sienten conformes con la calidad del servicio educativo, concretamente con los aprendizajes logrados por los alumnos.	<i>Reposo</i> : No existe evidencia empírica disponible sobre el rendimiento académico de la población estudiantil nacional con respecto a otros países, ni sobre los factores que en esos países impactan positivamente en la calidad de los aprendizajes. En consecuencia, no hay movilización política ni de la opinión pública y profesional para promover nuevas evaluaciones internacionales que, a su vez, estimulen el diseño o mejoramiento de los programas curriculares.
<i>Descontento</i> : En esta fase, algún actor, o grupo de actores del sistema percibe, o demuestra con algún tipo de evidencia, que el servicio educativo no está proveyendo suficientes oportunidades de aprendizaje y que por lo tanto el rendimiento de los estudiantes es inadecuado.	<i>Descontento</i> : Los resultados obtenidos en la participación en alguna prueba internacional, o en algunos casos los resultados de las mediciones nacionales, estimulan la opinión y la investigación sobre la necesidad de elevar la calidad de los aprendizajes y sobre las formas de lograrlo.
<i>Iniciativa y adopción</i> : Ese mismo grupo, u otros interesados, deciden proponer cambios en el sistema, impulsan el diseño de programas y estrategias comprensivas que	<i>Iniciativa y adopción</i> : La unidad central de administración educativa, o alguna entidad académica con el aval oficial, resuelven que es necesario medirse con respecto al logro académico de

<p>apuntan a elevar el nivel de los aprendizajes, buscan apoyo político y financiero para asegurar la viabilidad inicial de esos programas, y en algunas ocasiones buscan sensibilizar o movilizar a otros actores clave del sistema.</p>	<p>otros países y, de acuerdo a las prioridades curriculares y de política educativa en cada país, se opta por uno o más tipos de proyecto de evaluación internacional. Se espera de esta manera realizar un seguimiento de los alcances y efectos de los programas curriculares nacionales.</p>
<p><i>Implementación:</i> Los programas se ponen en práctica, buscando articular nuevos elementos (currículo, textos, capacitación docente, etc.) con recursos ya existentes.</p>	<p><i>Implementación:</i> En cada país se ponen en marcha los procesos técnicos de evaluación internacional, desde las primeras reuniones de discusión entre expertos sobre el marco conceptual a ser utilizado, hasta la selección y validación de ítemes, selección de la muestra poblacional y aplicación final de los instrumentos. Si la relación entre los equipos técnicos de medición y los profesionales del desarrollo curricular en cada país es fluida y cooperativa, es probable que el proceso de diseño de las pruebas contribuya a las primeras revisiones críticas de los contenidos curriculares nacionales, ya sean antiguos o de reciente diseño e implementación.</p>
<p><i>Evaluación de proceso y resultados:</i> En esta fase se monitorea el proceso de implementación para poder realizar ajustes a los nuevos programas en curso (evaluación formativa), o se decide medir los resultados e impactos de esos programas (evaluación sumativa). La evaluación sumativa requiere, en el contexto de estudio que nos ocupa, la aplicación de instrumentos que permitan medir objetivamente los logros de aprendizaje de los alumnos.</p>	<p><i>Interpretación y disseminación de resultados:</i> Se lleva a cabo la lectura de los resultados de evaluación por parte de técnicos de medición así como de los profesionales del desarrollo curricular convocados durante el proceso de implementación. Se preparan informes de resultados para la prensa, la opinión pública en general y para diferentes audiencias profesionales del sector, se ofrecen talleres de difusión y desarrollo de capacidades de interpretación y análisis de los datos, y se liberan las bases de datos para uso de los medios académicos con fines investigativos. Los esfuerzos de interpretación se centran en las inferencias de tipo curriculares que puedan realizarse, es decir en relación a los contenidos</p>

	prescritos y enseñados de las áreas de conocimiento evaluadas.
	<u>Decisiones, impacto político y nuevas líneas de investigación</u> : Los resultados de las evaluaciones influyen directamente, o indirectamente mediante la presión de la opinión pública y profesional, sobre las decisiones de funcionarios y profesionales del sector para elevar la calidad de los aprendizajes.
<u>Retroalimentación</u> : La información recabada mediante las evaluaciones se utiliza para mejorar los insumos o procesos de implementación, y así elevar las posibilidades de que los programas en curso tengan mayor impacto en los aprendizajes.	<u>Retroalimentación</u> : Las autoridades del sector o las agencias evaluadoras mismas llevan a cabo un sondeo o investigación sistemática sobre el tipo de impactos que ha tenido la experiencia de evaluación comparada internacional. Los resultados de esa investigación se utilizan para decidir futuras participaciones en pruebas internacionales.

El marco secuencial así establecido permite no solamente ordenar el contenido de las entrevistas y de los datos recogidos sino que, en las primeras fases del trabajo de campo, también permite anticipar uno de los aspectos del proceso que representa el mayor obstáculo para el uso efectivo de la evaluación internacional en el mejoramiento de la calidad de los aprendizajes. Ese obstáculo, concretamente el de ausencia de diálogo entre investigadores o evaluadores y los usuarios potenciales de la información, requirió para su explicación de un segundo referente teórico que se halló en el trabajo de Fernando Reimers (1999).

Reimers sostiene que en las altas esferas de administración educativa en muchos países del mundo, la mayor parte de las decisiones que se toman para incrementar las oportunidades de aprendizaje se realizan intuitivamente, especulando con la articulación de algunos de los elementos del sistema (currículo, capacitación, supervisión, etc.), en vez de tomarse como resultado del análisis cuidadoso de la evidencia que muestran las investigaciones y los análisis de política. El autor argumenta que esto se debe a la forma en que el conocimiento derivado de la investigación es acumulado y procesado, y a que no siempre se reconoce la diferencia entre la investigación que pretende simplemente explicar la realidad, y la investigación de política que, además, busca describir cómo podría cambiarse esa realidad, que se asume es lo que los funcionarios necesitan saber. Esa función propositiva de la investigación debería darse en forma de diálogo participativo entre investigadores y funcionarios u

otros usuarios de la información, de manera tal que se logre un acuerdo entre las necesidades de información y las posibilidades reales de los investigadores de proveer esa información. Reimers cita a Husén (1994) para ejemplificar, precisamente, cómo la falta de diálogo entre investigadores y funcionarios ha impedido que muchos de los estudios internacionales de la IEA logren tener un impacto real en las políticas educativas de los países participantes, y cómo con el paso de los años se ha intentado reforzar ese vínculo entre proveedores y usuarios de información.

El modelo racional de toma de decisiones, que asume que los funcionarios seleccionarán opciones de política sobre la base de un análisis informado, está siendo descartado. Se reconoce que la mayor parte de los problemas no están, ni pueden ser, perfectamente definidos y que los funcionarios no siempre pueden establecer prioridades de acción racionalmente sino en función de variables de oportunidad política y de poder. Por eso, para lograr decisiones informadas, es preciso ejercer cierto grado de persuasión de los decisores de políticas mediante un proceso de negociación, y no de diseño unilateral, del proceso de reforma. Ese proceso debe ser interdisciplinario (que es, en realidad, la forma en que los funcionarios perciben los problemas de política), debe ser propositivo y no solamente descriptivo, y debe tener en cuenta que el análisis debe realizarse y comunicarse dentro de los límites de tiempo que impone el ciclo de política. Reimers, adicionalmente, propone el uso de un “mediador de conocimiento” entre el mundo de la investigación y el mundo de la política. La evidencia recogida para este estudio muestra que, en ciertas ocasiones, esa mediación se logra a través de la prensa y de otras instituciones académicas y no gubernamentales formalmente convocadas *durante* el proceso de evaluación internacional. Sostiene Reimers que “el problema del uso de la información no puede ser considerado recién después de concluida la investigación o el análisis [...] la investigación debe comenzar con la consideración de las necesidades del cliente si se intenta evitar el riesgo de que los resultados del análisis resulten irrelevantes para los funcionarios”.

El uso de los dos referentes teóricos mencionados, sobre etapas secuenciales de política y sobre diálogo informado entre investigadores y funcionarios, servirá en este informe para organizar el reporte de los datos recogidos en forma ordenada y contextualizada, y a la vez sustentar conceptualmente uno de los hallazgos clave en este estudio de investigación.

## **II. METODOLOGÍA**

Sobre la base del referente teórico sobre usos de la información para decisiones de política, y tomando en cuenta el modelo secuencial o cronológico de política antes descrito, el diseño del presente estudio quedó metodológicamente configurado de la siguiente manera:

1. Revisión bibliográfica inicial sobre pruebas internacionales existentes, países participantes y principales características de esas pruebas. Elaboración inicial de marco teórico-conceptual.
2. Diseño del cuestionario de entrevistas, principal herramienta de investigación en este estudio, con una concepción *a priori* de los momentos o fases que presentaría la participación de diferentes países en pruebas internacionales. Ese cuestionario (ver *Anexo B*) fue revisado y corregido en dos oportunidades por un consultor externo<sup>iv</sup> y se aplicó en una primera fase en varias entrevistas a altos funcionarios y ex-funcionarios de las oficinas nacionales de evaluación educativa en cuatro países de América Latina y a académicos de los Estados Unidos involucrados en el diseño y aplicación de pruebas internacionales (ver lista total de entrevistas en *Anexo C*).
3. Pilotaje y afinamiento de cuestionario de entrevistas y recolección de materiales bibliográficos y documentales adicionales.
4. Primer análisis de datos, con el objetivo de afinar el marco conceptual, revisar los instrumentos de campo e identificar nuevos posibles informantes.
5. Reelaboración de marco teórico-conceptual de política para encuadrar el análisis del resto de las entrevistas. Búsqueda de sustento teórico adicional para explicar el quiebre de comunicación entre evaluadores/ investigadores y decisores de política.
6. Selección de ocho países americanos como muestra o materia de estudio en relación a su experiencia de participación en pruebas de la IEA, LLECE y OCDE. Esta muestra no se considera representativa, pero es suficiente para identificar algunas tendencias y patrones comunes en los procesos de participación de América Latina en pruebas internacionales. La selección de los países estuvo condicionada por las limitaciones financieras de los investigadores, que prefirieron realizar entrevistas personales y telefónicas solamente, haciendo uso de tiempo y recursos financieros adicionales de otros proyectos de investigación.
7. Finalización del trabajo de campo y desgrabaciones de entrevistas a cargo de personal contratado y revisadas luego por los investigadores.
8. Análisis de datos: codificación de las entrevistas y reordenamiento de la información de acuerdo a los momentos de la experiencia de evaluación internacional y su relación con los momentos de reforma curricular nacional,

por tipo de prueba aplicada, por fases del proceso de participación, y por país.

9. Discusión sobre los hallazgos del estudio y presentación de algunas posibles generalizaciones aplicables a todo América Latina a partir de los mismos.
10. Redacción de recomendaciones para futuras participaciones del Perú y otros países latinoamericanos en pruebas internacionales.

## **RESULTADOS**

A continuación se presentan los resultados del trabajo de campo, concretamente los contenidos de las entrevistas y del material documental recogido. Vale aclarar que del total de pruebas internacionales en las que América Latina ha participado, o está actualmente participando, se decidió reportar los datos de las experiencias del Primer Estudio Internacional Comparativo del LLECE (UNESCO-OREALC), de TIMSS, TIMSS-R y Educación Cívica de la IEA, y de la prueba PISA administrada por la OCDE. El trabajo del LLECE se seleccionó por ser el único estudio comparado de cobertura estrictamente regional, y porque en él participó un número importante de países latinoamericanos. La pruebas de la IEA fueron seleccionadas porque han convocado la participación de algunos países de la Región, y porque esa experiencia puede ser contrastada con la de otros países, como Estados Unidos, donde el impacto de la misma evaluación ha sido mucho más significativo que en América Latina. La prueba PISA se seleccionó por ser una prueba “acurricular” que propone estándares internacionales, porque ha concitado el interés de varios países de la Región, y porque es la única prueba en la que Perú está participando en la actualidad. Una breve descripción de las pruebas mencionadas se presenta en el *Anexo D* de este informe.

Si bien los datos presentados en esta sección pueden contener sesgos derivados de la interpretación subjetiva de los investigadores, se ha intentado que la información reportada se ajuste estrictamente a las expresiones de los entrevistados. Para una mejor comprensión de los datos, la información se ha clasificado, en primer lugar, de acuerdo a las agencias que administran las pruebas (LLECE, IEA y OCDE). En segundo lugar, se organiza la información de acuerdo a las etapas en la secuencia de política y de evaluación explicadas en el marco teórico-conceptual. Finalmente, para cada una de esas etapas, se presenta la información de los países de la muestra sobre los cuales se posee información relevante.

**Agencia: LLECE (UNESCO-OREALC) – Prueba: Primer Estudio Internacional Comparativo sobre Lenguaje, Matemáticas y Factores Asociados.**

### **Etapa: Descontento**

En 1989 la UNESCO llevó a cabo un estudio comparativo exploratorio de siete países de la región que incluía pruebas de rendimiento y encuestas. No fue un estudio riguroso pero sentó precedente y los bajos resultados alertaron a actores políticos y académicos de la Región sobre las deficiencias de los sistemas educativos y pusieron en cuestión la eficacia de los últimos treinta años de políticas educativas. En una reunión en Campinas, Brasil, donde se presentaron los resultados de ese primer estudio exploratorio, los representantes regionales solicitaron a la UNESCO-OREALC que diseñara un mecanismo de evaluación que permitiera reunir información en forma sistemática a través de una organización estable. OREALC coincidía en advertir que existían falencias importantes en términos de calidad educativa, no ya medidas en términos de cobertura y universalización sino también en cuanto al logro de aprendizaje académico.

### **Etapa: Iniciativa y adopción**

Hacia el momento en que se crea el Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación (en adelante, “Laboratorio”), y que empieza a gestarse la idea de realizar una prueba comparativa latinoamericana, predominaba básicamente la idea de la reforma de la administración del Estado, no solamente de reforma de la educación. Esto se siente desde los inicios de la década de los 90 con la caída del muro de Berlín. Es allí cuando se inicia un proceso de crítica muy fuerte al “Estado benefactor” y comienzan en la Región procesos muy marcados de ajuste estructural, que en el área social se sintió muy especialmente en el sector educación. Se instala un discurso y algunas medidas concretas de descentralización, como la transferencia de gestión que realiza el Ministerio de Educación de Argentina a los ministerios provinciales, y donde la administración central se reserva el rol de evaluador a través de su sistema estadístico y de medición de aprendizajes. Esto implicaba comenzar a concebir una nueva forma de gestión que permitiera diseñar políticas educativas sin tener la gestión directa de las escuelas. Se observa una transición de un modelo de planificación normativa a un modelo de gerenciamiento donde se pretende controlar al sistema no ya desde los insumos provistos sino desde los resultados. Este modelo de gerenciamiento no necesariamente se proponía en forma explícita desde los diferentes planes de gobierno y sectoriales en la Región, pero su necesidad sí se sentía a través de las formas de gestión que planteaban las agencias crediticias internacionales, especialmente mediante el condicionamiento de los préstamos a la creación de sistemas nacionales de evaluación de la calidad educativa. En ese contexto político y de administración



pública, era palpable un fuerte mensaje de que la riqueza y la competitividad de las naciones dependían de la eficiencia de sus sistemas educativos. Si bien esa relación no se sustenta hasta el día de hoy con evidencia consistente (Forster, 2000), en su momento promovió el deseo de los representantes nacionales por develar el estado de la competitividad de sus países según indicara la capacidad competitiva de sus sistemas educativos en relación a otros países.

La reunión en Campinas antes mencionada fue importante porque fue representativa de cada país, de la sociedad civil (ONG de educación), del gobierno y del entorno académico. Fue destacable en ese momento la participación del aparato ejecutivo de diferentes países, representados no por los ministros de educación directamente, sino por la línea de funcionarios casi inmediatamente después de ellos. A pesar de que entre estos actores y representantes existía una clara percepción sobre la crisis de la educación en América Latina, no se contaba con un referente empírico que ofreciera evidencia contundente sobre las falencias de los sistemas, particularmente en relación al logro de aprendizajes. La petición que se realiza en ese momento a OREALC tiene entonces mucha fuerza, porque proviene tanto de la academia como de los representantes políticos de diferentes países. Tuvo especial impacto también el evento de PROMEDLAC de 1993 en Santiago de Chile, donde los ministros de educación discutieron la necesidad de que UNESCO impulse el desarrollo de la evaluación y la creación de estándares en la Región. Esta petición por parte de los gobiernos se transforma en un mandato para la OREALC, quien comienza a invitar a los países para participar en una experiencia de evaluación conjunta. El consenso final para la realización del Primer Estudio Internacional Comparativo se logró en México en 1996.

Paralelamente, la OEI comenzaba a generar su propio programa de evaluación en educación y, por su parte, la IEA ofrecía la posibilidad de que América Latina participara en TIMSS. Sin embargo, a pesar del interés inicial que manifestaron varios países por la prueba de la IEA, muy pocos llegaron a inscribirse formalmente en el proyecto definitivo, y sólo uno de ellos, Colombia, culminó el proceso y publicó los resultados internacionalmente. Esto, según la interpretación de OREALC, pudo deberse a que la participación en TIMSS era costosa y los países percibían que no estarían adecuadamente representados en la experiencia. La oferta de OREALC, por otra parte, implicaba una idea de identidad regional donde se generara una visión compartida sobre la evaluación en un contexto organizacional de confianza. Cuando se aprobó la idea de crear el Laboratorio prácticamente todos los países optaron por adherirse. El Laboratorio, que es considerado más que el estudio comparativo mismo, se concibió como una red de los sistemas nacionales de evaluación emergentes a principios de los 90, en momentos en que la responsabilidad de crear y sostener un sistema de evaluación era complejo desde el punto de vista técnico y político. En esos momentos, el Laboratorio ofrecía un espacio protegido de diálogo y participación para los funcionarios y técnicos responsables de esos sistemas, y brindó apoyo para la creación de varios sistemas nacionales de evaluación.

Según lo percibe la OREALC, los sistemas nacionales de evaluación fueron financiados por “los Bancos” pero se crearon en el contexto político y cultural del Laboratorio, que a su vez ofrecía ciertas garantías a los financistas de que el dinero sería bien invertido. El proyecto de evaluación comparada del Laboratorio fue financiado inicialmente con los fondos de cada país participante, y posteriormente con aportes del BID, que se hicieron efectivos luego que el estudio había concluido.

Se considera que hubo dos razones fundamentales que motivaron la decisión final de realizar el estudio comparado. La primera fue la necesidad de establecer una comparación propia, “de raíz latinoamericana”, que tuviera el mínimo posible de sesgos. La segunda, era obtener una herramienta de información que permitiera apoyar las decisiones de política educativa. En relación a los posibles sesgos en las pruebas, se consideraba que era fundamental contar con pruebas regionales que establecieran parámetros curriculares acordados por todos los países participantes y que permitieran realizar inferencias más pertinentes a la realidad educativa regional. Una tercera razón que motivó a los países a participar en la prueba del Laboratorio en vez de participar en pruebas de cobertura mundial fue que el Laboratorio ofreció una alternativa extremadamente accesible en términos económicos en comparación con la oferta de otros estudios internacionales.

El primer objetivo para el Laboratorio era entregar una comparación de logro, de rendimiento académico entre países. En segundo lugar, se pretendía ofrecer un análisis de factores asociados a ese rendimiento. Otros objetivos complementarios eran fortalecer los sistemas nacionales de medición y evaluación por la vía de la práctica a través del primer estudio comparado, establecer estándares educativos, y generar una línea de investigación a partir de los resultados.

En **Argentina**, la decisión de participar se tomó por dos razones. En primer lugar, por las oportunidades para la capacitación que ofrecía la experiencia. En segundo lugar, porque el Laboratorio constituía un espacio para la discusión de temas comunes a las reformas de varios países, como por ejemplo aquella sobre desarrollo curricular. Se consideraba que se debía tener capacidad instalada en el país, y tener datos sobre el rendimiento en el ámbito nacional, antes de embarcarse en una propuesta internacional comparada. Argentina realizaba operativos nacionales desde el año 1993 y conocía las dificultades que supone institucionalizar y legitimar políticamente la evaluación externa. Dadas esas dificultades, era necesario considerar los efectos de las evaluaciones en la opinión pública, y en ese sentido convenía participar primero en una prueba regional antes que arriesgarse a salir en los últimos lugares en una prueba de cobertura mundial como TIMSS. Los efectos de un “ranqueo” desfavorable a Argentina se consideraban una amenaza a toda la estrategia de implantación del sistema nacional de evaluación. La solución era graduar la estrategia, estableciendo primero estándares latinoamericanos y después recién estándares

internacionales. Allí se expresa la clara expectativa de Argentina de participar no solamente en un proceso de medición sino de establecer parámetros curriculares regionales sobre los cuales efectuar tal medición.

En el caso de **Colombia**, el Ministerio de Educación también mostró un interés temprano en la participación en el estudio comparado. Para los representantes técnicos del ICFES, una unidad semiautónoma con muchos años de trayectoria en medición educacional, la idea del estudio comparado regional era atractiva aunque ellos no decidieron la participación, simplemente respondieron a la petición oficial del Ministerio de Educación para que fueran co-representantes de Colombia, junto con un representante del Ministerio, ante el Laboratorio. Sin embargo, existía una motivación intrínseca por parte del ICFES en sumarse al proyecto comparado, en tanto se esperaba adquirir mayores capacidades técnicas y contribuir a un proceso de definición conjunta entre países de las expectativas de logro académico y de un marco teórico (epistemológico) y de política educativa que permitiera interpretar adecuadamente los resultados de la medición internacional. En cuanto a las expectativas técnicas en particular, si bien ellos ya contaban con la más larga experiencia en América Latina en el campo de la psicometría, sabían que estaban atrasados, como el resto de la región, en muchos campos de conocimiento, como por ejemplo en relación a modelos de IRT (Teoría de Respuesta al Ítem). Simultáneamente, Colombia esperaba amistosamente contribuir con conocimientos propios al Laboratorio, pues además de las mediciones nacionales había pasado por la experiencia de participación en TIMSS en el año 1995.

En **Cuba** la decisión de participar se tomó luego de una visita del entonces Director del Laboratorio. Cuba no disponía del dinero para financiar su propia participación en la prueba pero el Laboratorio ofreció la posibilidad de pagarlo y entonces decidió sumarse al proyecto. Para Cuba era la primera experiencia de comparación internacional, y se daba en un momento en que el país comenzaba a atravesar serias dificultades económicas y un gran desafío por demostrar al mundo que sus políticas sociales eran efectivas. La comparación, indudablemente, implicaba riesgos pero a la vez una gran oportunidad de legitimación política para el país si los resultados eran favorables, y también de contar con un referente empírico para evaluar sus políticas educativas. Desde el punto de vista de la administración pública, Cuba decidió participar en un momento de relativa descentralización en que se reestructuraron algunos aspectos del sector educación, y en el que el Ministerio de Educación central, en aras de mayor eficiencia y ahorro de recursos, optó por transferir algunas facultades de decisión al nivel de las escuelas, tales como la acreditación de alumnos que anteriormente estaba en manos de un órgano evaluador central.

**Ecuador** participó en una aplicación piloto de la prueba, pero no participó en la aplicación general por razones de discontinuidad política. Sin embargo, realizaron una aplicación general en forma independiente más adelante, aunque aún no han recibido los resultados. Desde la agencia nacional de medición, que

no era en ese momento enteramente dependiente del Ministerio de Educación, se consideraba que era importante participar en una prueba comparada, pues tener referentes exclusivamente internos no se consideraba suficiente para juzgar la calidad de la educación. Se sabía cómo rendían las diferentes subpoblaciones del sistema educativo respecto de un conjunto de destrezas que se habían definido para el Ecuador, pero se hubiera valorado la medición con respecto a otras naciones para rescatar información sobre las estrategias de política que mejor han funcionado en esas naciones. Adicionalmente, se confiaba en que la participación habría contribuido a desarrollar capacidades nacionales de medición. En relación a futuras oportunidades de participación, y teniendo en cuenta que es necesario priorizar la asignación de recursos para decidir sobre el tipo de pruebas comparadas en las cuales participar, se considera que es mejor comenzar participando en estudios regionales que en estudios de cobertura mundial.

En **Perú**, la primera propuesta de adhesión al Laboratorio y participación en la prueba comparada la realizó la representante de UNESCO en el país, junto con otro funcionario de la organización. Su función era mantener el interés del Ministerio de Educación en el proyecto y que se cumpliera el compromiso financiero con la OREALC. La decisión de participación fue tomada por la jefa de la unidad de medición, junto con la Oficina de Planificación Estratégica y con el aval de la viceministra. Si bien el Estado estaba atravesando un proceso de modernización de la gestión, y el sector educación en particular experimentaba cambios de organización central y planes de mejoramiento de los servicios, la decisión final de participar en el Laboratorio no parece responder a un deseo, al menos no explícito, de reforzar una forma determinada de gestión, como puede ser la de gerenciamiento, o de efectuar mejoras de tipo pedagógico o curricular a partir de los resultados. Sin embargo, se sostiene que “los Bancos” ejercían una presión puntual para que se realizaran mediciones de logro académico. Las dos razones más concretas que se exponen para la participación son, por un lado, acceder a oportunidades de capacitación técnica en diseño de instrumentos y análisis estadístico vía el intercambio con expertos internacionales y, por otro lado, rescatar información que permitiera diseñar algunas medidas focalizadas de política, como por ejemplo en la educación rural. Esa segunda intención, sin embargo, no parece del todo fuerte dado que el Ministerio de Educación ya estaba desarrollando un operativo de medición nacional que le permitiría tener ese tipo de información en un grado más detallado, y la Unidad de Medición no tenía un interés particular en vincular los datos comparados con los datos que arrojava el sistema nacional. El valor comparativo-competitivo de la experiencia no fue considerado. En realidad, era más bien temido. El proceso de decisión y los primeros pasos de la participación fueron algo accidentados, especialmente por razones políticas internas, pero finalmente, y a pesar de no cumplir con el compromiso financiero con OREALC, Perú aplicó las pruebas del Laboratorio. Cabe destacar que durante todo el proceso de evaluación comparada, el vínculo entre UNESCO-Perú y el Ministerio de Educación fue particularmente débil.

En **Uruguay**, la decisión final de *no* participar en la prueba del Laboratorio fue tomada por razones financieras, aunque a la base de esa decisión había otras razones de carácter técnico y político. En 1996, cuando recibieron la invitación para sumarse a la evaluación, asistieron a una reunión en Buenos Aires donde se discutió el operativo e instrumentos a utilizar. Luego de esa reunión, Uruguay decidió no participar por dos razones principales. En primer lugar, no estaban convencidos sobre la calidad técnica de los instrumentos que se utilizarían. En segundo lugar, consideraban que políticamente convenía primero consolidar su sistema nacional, con las características que ellos querían imprimirle, sin complicar el proceso con la implementación de un segundo operativo con pruebas sobre las cuales ellos no tendrían verdadero control. En ese tiempo, consideraban prioritario garantizar que sus pruebas nacionales fueran óptimas y que fueran avaladas y respetadas por los maestros. Temían que la participación en una prueba internacional pudiera dañar la credibilidad interna de su sistema de evaluación, en gran medida por las fuertes resistencias a la evaluación que Uruguay ha tenido que confrontar. Finalmente, se argumenta que la decisión de no participar en pruebas internacionales en general se debe a que hasta ahora no se lo han planteado muy seriamente, excepto por la posibilidad de sumarse al proyecto PISA, al cual se hará referencia más adelante.

### **Etapa: Implementación**

El estudio tuvo sede en UNESCO-Chile, y desde allí se coordinó todo el operativo. Uno de los aspectos más complejos del estudio fue el financiamiento. En la medida en que se iba consiguiendo dinero, tanto de los países como de donantes internacionales, se iba contratando consultores para las diferentes etapas del proyecto. Sólo cuando se recibieron aportes del BID, sumados a algunos fondos de la Fundación Ford, se pudo constituir un equipo relativamente estable de profesionales. El dinero que los países participantes contribuían al Laboratorio se devolvía a los países en forma de servicios de capacitación para la medición.

Considerando la experiencia de ver cómo los países de la Región habían optado por no participar o por retirarse de TIMSS, se consideraba de vital importancia garantizar que los ítemes de las pruebas fueran significativos, en tanto tuvieran una fuerte vinculación a los currículos de los países participantes. Se invirtieron dos años y medio en el diseño de las pruebas, con reuniones para la selección de ítemes, verificando currículos nacionales de matemáticas y lenguaje. Este fue uno de los aspectos más beneficiosos del proceso, pues puso a los países en la necesidad de efectuar una revisión importante de sus currículos y así descubrir debilidades y fortalezas relativas. Si bien fue un trabajo “largo y tedioso”, fue absolutamente relevante para que todos se sintieran representados. Los representantes ante el Laboratorio, los llamados “coordinadores nacionales”, eran en su mayoría los directores de los sistemas de evaluación en sus propios países, o de la oficina ministerial de la cual esos sistemas dependían. Según la

percepción de los responsables del Laboratorio, los coordinadores nacionales tenían una representatividad político-técnica importante y todas las decisiones eran tomadas por acuerdo, tratando de asegurar que todas las voces estuvieran representadas. El diseño de la matriz curricular se hizo con la participación, en algunos casos, de los técnicos en currículo de los diferentes ministerios.

En algunos países hubo tensiones importantes entre las oficinas del currículo y los sistemas nacionales de evaluación, y en algunos casos esa tensión determinó que los departamentos de currículo no estuvieran representados en absoluto durante el proceso. En este sentido, se admite que la prueba final no tenía una base curricular de igual significancia para todos los países, lo cual repercutiría en las inferencias de tipo curricular que se hicieran más adelante. Esto, a pesar de que los contenidos que se incluían en las pruebas debían estar representados en los currículos de al menos el 80% de los países participantes. No se esperaba necesariamente una relación técnica entre las pruebas nacionales y la del Laboratorio, pero como se evaluaban los mismos grados que en varios sistemas nacionales, los países esperaban confirmar de alguna manera lo que ellos mismos venían evaluando internamente. Se aplicó la misma prueba en todos los países, y no hubo “opciones nacionales”, con excepción de México donde, junto con la muestra internacional, se realizó una muestra nacional de cada uno de los estados.

Desde OREALC se percibe que los países no tenían definiciones muy concretas sobre el tipo de capacidades que deseaban desarrollar mediante su participación en la prueba comparada. En la práctica, hubo ciertas capacidades que tuvieron mayor desarrollo que otras, siendo de mayor importancia las relacionadas a la construcción de pruebas, desde el análisis curricular hasta la construcción de los ítemes mismos, pese a que los ítemes en su gran mayoría fueron contruidos por especialistas contratados *ad hoc*. Desde la coordinación del proyecto en Santiago se percibe que la comunicación con los países era frecuente y la discusión profunda. Otro tipo de conocimiento que se generó ampliamente estuvo relacionado a las técnicas de muestreo. Si bien el esquema del muestreo fue diseñado centralmente por la coordinación, las muestras nacionales fueron seleccionadas por cada país individualmente y verificadas por la coordinación en combinación con un tercer país que obraba de árbitro o testigo. Ese procedimiento tuvo un primer propósito que fue el de asegurar la validez de las muestras, y otro que aspiraba a fomentar la transferencia de “*expertise*” entre países. Finalmente, la experiencia del Laboratorio contribuyó a desarrollar capacidades operativas para aplicaciones en gran escala, particularmente en los países que no tenían experiencia con sistemas nacionales. Junto con la experiencia práctica, el Laboratorio ofreció una serie de talleres para desarrollar capacidades en la construcción de escalas, en la selección de las muestras, en el uso de modelos jerárquicos lineales, entre otros. Durante el desarrollo del estudio también participaron expertos de la IEA y de ETS.

Las oportunidades de capacitación, y de aplicación práctica de las capacidades que habría brindado el Primer Estudio Internacional Comparativo, no aparecen como un elemento tan claro para otros actores del proceso, tanto internacionales como nacionales. Voces ajenas al Laboratorio, pero vinculadas estrechamente a la evaluación educativa en la región, consideran que el proceso pudo resultar confuso para algunos países, especialmente para aquéllos donde los sistemas nacionales se encontraban en ciernes o todavía ausentes. La OEI, que también ofrece programas de capacitación a los sistemas nacionales de evaluación, sostiene que hay que ser extremadamente cuidadosos en los países con más bajo desarrollo técnico, para evitar crear confusión en cuanto al objetivo de las capacitaciones. Si bien hay capacidades que son requeridas para ambos tipos de mediciones, nacionales e internacionales, los técnicos nacionales deben tener claridad sobre el tipo de aplicación que se le puede dar a los conocimientos adquiridos, y deben tener la posibilidad de experimentar gradualmente, quizás primero en el contexto nacional, y luego internacionalmente. Más concretamente, se considera que es imperativo generar sólidas capacidades nacionales de manera tal que los técnicos que asistan en representación de sus países para las pruebas internacionales tengan efectiva autoridad profesional para influir sobre las decisiones técnicas y así eviten terminar simplemente traduciendo y aplicando mecánicamente las pruebas que otros han diseñado.

Con esta postura de la OEI coinciden, en general, las percepciones que sobre el tema tienen quienes representaron a los diferentes países participantes. En **Argentina**, por ejemplo, se percibió que los países se embarcaban en la experiencia comparativa en desigualdad de condiciones, y que por lo tanto los aprendizajes durante el proceso dependían de sus capacidades iniciales. Argentina se sentía comparativamente fuerte en cuanto a conocimientos técnicos, especialmente porque habían ya ejecutado cuatro o cinco operativos en el ámbito nacional. Entre los aspectos más positivos de la experiencia están el transitar experiencias compartidas con América Latina, poder discutir con otros países en situación de desigualdad técnica --para lo cual se requiere mucha tolerancia y buena voluntad-- y llegar a acuerdos para la toma de decisiones. La muestra seleccionada fue una submuestra de la muestra del operativo nacional y se aplicó al mismo tiempo que el operativo del año 1997. Las contribuciones a la matriz curricular del estudio se realizaron sobre el nuevo currículo nacional (Contenidos Básicos Comunes). Se tiene la percepción que el Laboratorio no supo aprovechar adecuadamente los conocimientos y la experiencia que se tenía ya en algunos de los países, por ejemplo, en la utilización de algunos modelos estadísticos. Aun así, el contenido de las discusiones sobre esos temas era interesante e instructivo. La representación de Argentina fue técnico-política, de alto rango ministerial, como fue el caso de la mayoría de los países. En la opinión de estos representantes, los estudios internacionales son estudios “de cúpula”. Cuanto más internacional es el estudio, más focalizado, más restringido es el equipo técnico-político que interviene.

En la experiencia de implementación de **Colombia**, se considera que uno de los beneficios mayores fue el fortalecimiento de la relación entre el Ministerio de Educación y el ICFES, cuyos representantes en las reuniones de Laboratorio mantuvieron comunicación constante. En el ICFES se organizaron las bases de datos, se digitaron los datos y se envió la información procesada al Ministerio. El ICFES considera que su fortaleza mayor siempre ha sido el manejo operativo de las aplicaciones nacionales, así como el procesamiento mismo de la información. La debilidad mayor, por su parte, ha estado siempre relacionada con las estrategias de difusión de la información sobre logro académico y factores asociados para lograr mayor impacto en el diseño de políticas de mejoramiento. En este sentido, el ICFES esperaba aprender más mediante la participación en la prueba comparada. Si bien tenían expectativas de obtener mayores conocimientos técnicos mediante esta experiencia, los representantes sostienen que tuvieron poca participación efectiva en el diseño de ítemes, en parte debido al hecho de que se sumaron al proyecto una vez que éste hubo comenzado. Para cuando los técnicos del ICFES se incorporaron a la evaluación, si bien tuvieron la posibilidad de revisar las pruebas y cuestionarios finales, sus sugerencias no fueron tenidas en cuenta pues las decisiones ya estaban tomadas. Sólo pudieron influir en algunos cambios mínimos y más bien formales. Entre las deficiencias más importantes del proceso, mencionan que nunca tuvieron acceso al marco conceptual con el cual se habían construido los instrumentos. Esto se considera riesgoso pues afecta la validez del análisis que pueda hacerse de los resultados. La falta de un marco conceptual fundamentado se hizo evidente, por ejemplo, cuando se puso a discusión uno de los ítemes de respuesta abierta de la prueba. Colombia rechazó dos veces todos los criterios de calificación propuestos por los especialistas del Laboratorio, y finalmente el ítem fue descartado. El ICFES ya venía experimentando con ítemes abiertos y conocía los criterios psicométricos necesarios y aceptables para ese tipo de medición, y por esa razón tenía más elementos para juzgar el escaso sustento conceptual del diseño.

Otras recomendaciones que se hicieron al Laboratorio no fueron tenidas en cuenta. Una de las razones por las cuales se considera que la comunicación y el intercambio falló es que no se dio un uso apropiado al recurso de correo electrónico. La comunicación fluía de los países hacia la sede del Laboratorio, pero no se facilitó la comunicación entre los países mismos. Hubiera sido importante que toda la información y opiniones se distribuyeran ampliamente entre los participantes, si es que se pretendía desarrollar una visión regional sobre la evaluación, sobre las disciplinas, para poder hacer una aproximación en conjunto a problemas comunes de reforma, y de desarrollo científico, cultural y social en América Latina. En síntesis, hubo deseo e intención de comunicación y colaboración multidireccional, pero falló la estrategia. La participación del ICFES en esta fase de implementación del estudio fue relativamente alta, pero lo aprendido no necesariamente redundó también en beneficios de aprendizaje para el Ministerio de Educación, quizás porque lo que se aprendió no es de aplicación permanente en ese ámbito.



Para **Cuba**, la participación en la prueba comparada de Laboratorio fue un proceso de amplio aprendizaje. En él participaron investigadores del Instituto de Ciencias Pedagógicas en representación nacional, pero también de otras universidades, así como maestros y asesores provinciales y municipales en diferentes áreas disciplinares. Valoran particularmente la participación en el diseño de las matrices curriculares, pues fue en esa fase del estudio cuando se revelaron falencias del currículo cubano, como por ejemplo con relación al escaso enfoque comunicativo en el área de lenguaje, o la demora en introducir contenidos de matemáticas en los grados inferiores. También fue un momento de confrontación técnica entre diferentes formas de evaluar, con lo cual se generó debate y se accedió a nuevos conocimientos sobre evaluación, como por ejemplo con relación al diseño y calificación de respuestas abiertas. Finalmente, la experiencia permitió acceder a nuevas tecnologías de medición, tales como el modelo de Rasch y modelos jerárquicos lineales. Ese conocimiento se difundió a otros profesionales del país que no estaban directamente involucrados en la evaluación del Laboratorio. También hubo intención de incluir en el diseño del estudio una opción nacional que, por razones de financiamiento, no pudo concretarse.

Para **Perú**, la representación y participación durante el proceso de implementación de la prueba fue básicamente de carácter técnico. Las primeras dificultades estuvieron relacionadas a factores políticos, concretamente de antagonismo entre un dirigente sindical que fungió como representante ante el Laboratorio inicialmente y la cúpula directiva del Ministerio. Luego cambió ese representante y el proceso de implementación se volvió eminentemente técnico. Durante las reuniones internacionales se discutió el diseño de la prueba, la muestra y la utilidad potencial de la información que se recogería. La percepción en cuanto a las posibilidades de influir sobre el diseño es que sólo se materializó con relación a los cuestionarios de factores asociados, no a las pruebas mismas. Participaron en la revisión de los ítemes ya formulados, pero los aportes no fueron sustanciales, sino sólo para reducir algunos sesgos percibidos. Sin embargo, se considera que fue de gran utilidad un curso de capacitación de 15 días de duración que el Laboratorio ofreció en Santiago de Chile.

Al igual que en Colombia, no se percibe que la experiencia haya constituido un espacio de identidad y comunicación regional. Durante el proceso de desarrollo e implementación las reuniones eran más bien fluidas y Perú no realizó mayores observaciones técnicas, aunque en algunas de las reuniones los coordinadores nacionales llegaron a ciertos consensos para garantizar la comparabilidad de los datos. La escasa intervención durante ese tiempo pudo haberse debido al tipo de dinámica utilizada para las discusiones, que ahorra tiempo de debate y le permitía al Laboratorio continuar con los tiempos establecidos desde el inicio. Las objeciones técnicas mayores que Perú hizo al Laboratorio no se incorporaron sino hasta el final del proyecto, cuando se dieron a conocer los resultados, tema que se retoma en la próxima sección. Perú no propuso aplicar

opciones nacionales, pero sí acordó establecer estándares sobre los cuales realizar las mediciones, aunque luego ese acuerdo se cambió por decisión compartida entre los países.

### **Etapas: Interpretación y diseminación de resultados**

La entrega de los resultados finales por parte de la UNESCO-OREALC se hizo directamente a los ministerios, a la vez que se realizaron conferencias de prensa en Santiago de Chile y se editó un compacto para televisión que fue difundido por la Red de TV Española a América Latina. Todos los países participantes de la prueba del Laboratorio habían asumido el compromiso de publicar los resultados, es decir que los resultados pudieran hacerse públicos internacionalmente, independientemente de lo que cada gobierno decidiera hacer con la información. La OREALC considera que un sistema de evaluación internacional opera en la confianza, y por eso fue muy difícil resolver las situaciones que se presentaron hacia el final de la experiencia, cuando dos de los países participantes, México y Perú, cuestionaron la validez de los resultados y no autorizaron su publicación.

Otro problema, relacionado al anterior, es que no se había anticipado desde el Laboratorio que la información generada por la evaluación sería de interés principalmente para la cúpula de gabinete ejecutivo en cada país, y que el flujo de información más importante se daría solamente entre ministros de educación y sus presidentes de gobierno. Era decisión de los gobiernos cuál sería la difusión que se daría a la información en cada país, y en muchos casos los resultados sólo circularon en reuniones de cúpula, acaso compartiendo un resumen con la prensa, pero sin la voluntad de generar debate público sobre el tema. Esta falta de difusión en los países se atribuye, entre otras cosas, a la gran diferencia de rendimiento entre Cuba y el resto de los países. Se interpreta que algunos de los países que habían realizado cuantiosas inversiones en reformas con apoyo internacional, principalmente de Estados Unidos, sentían gran incomodidad al ver que un país no alineado con las políticas de desarrollo de los 90 estuviera tan por encima del resultado del resto de la Región. Esto puede en realidad haber inhibido una mayor difusión de los datos en cada país, sumado al hecho de que existe una fuerte resistencia cultural a la comparación, y especialmente a la comparación con países más desarrollados. Sólo Chile y Colombia se han comparado públicamente con países altamente industrializados. En un estudio latinoamericano, por otro lado, había la esperanza de que todos logran un resultado similar, y el hecho de que Cuba quebrara esa expectativa causó incomodidad en las esferas políticas de los países vecinos.

Los países participantes sabían con anticipación cuál sería el mecanismo de anuncio de manera tal que pudieran preparar sus propias estrategias de comunicación internamente y ante la opinión pública. El Laboratorio publicó dos

informes principales. El primero presenta los resultados comparativos en forma de *ranking* y desagregados por estratos. Esta primera información se consideró de carácter más bien político, simplemente para saber la posición relativa de los países. El segundo informe, que salió tiempo después, ofrece un análisis de factores asociados al rendimiento académico en el *conjunto* de países participantes, y se concibió con la idea de ofrecer mejores insumos de información para la toma de decisiones de política educativa y para estudios académicos en la Región. Si bien el informe no presenta esta información desagregada por países, lo cual limita el uso que cada sistema nacional le pueda dar a los datos, se espera que cada país realice luego su propio análisis de factores asociados con la asistencia técnica del Laboratorio.

Los equipos de representación técnica y política de los países participantes no cambiaron demasiado entre las primeras fases de la experiencia y el momento en que se entregaron los resultados. Esto debió permitir que la información tuviera más impacto dentro de cada país, pero sin embargo ese impacto no se percibió, como será descrito en la próxima sección. La organización del Laboratorio asume parte de la responsabilidad por esa falta de impacto debido a que no se dedicó tiempo y recursos suficientes a la planificación y ejecución de una estrategia de difusión regional adecuada. Este aspecto de las evaluaciones se considera como la gran debilidad de las experiencias de medición, tanto internacionales como nacionales. Sin embargo, a pesar de esta autocrítica, cabe recordar que el Laboratorio, como organismo internacional e intergubernamental, tiene limitaciones formales en cuanto al grado de involucramiento e influencia que puede ejercer sobre las políticas educativas de los países miembro.

En **Argentina**, la principal receptora de los resultados fue la Ministra de Educación, quien los compartió inmediatamente con el gabinete ejecutivo de la Presidencia. La lectura que se dio a los datos fue mínima, en parte porque el primer informe sólo fue comparativo y no permitía realizar inferencias sobre las razones que explicaban esos resultados. Se interpretó que los datos confirmaban de alguna manera lo que decían los resultados de los operativos nacionales en cuanto a, por ejemplo, diferencias significativas de rendimiento entre estratos socioeconómicos, o entre escuelas públicas y privadas. Como en el resto del continente, sorprendieron los resultados de Cuba, los que se atribuyeron al alto valor social que se le otorga a la educación en ese país. Antes de la devolución de resultados no se previeron estrategias de difusión y a la prensa sólo se le entregó una copia del informe oficial del Laboratorio, sin interpretaciones por parte del Ministerio. Se considera que habría podido darse más uso a la información del segundo informe del Laboratorio, si éste hubiera salido durante la gestión ministerial anterior. Retrospectivamente, se piensa que la estrategia de difusión que faltó debió consistir en una socialización amplia de los resultados, con los medios de comunicación, y entre directores de escuela, padres de familia y asociaciones civiles. Se asume que la escasa difusión de los resultados se debe, también, a que OREALC no propuso una estrategia definida al respecto.

El Ministerio de Educación de **Colombia** sostuvo con el ICFES algunas discusiones sobre la visión política del proyecto, sobre el tipo de información que se esperaba rescatar, y sobre el tipo de impacto que se esperaba luego de la difusión de los resultados. Sin embargo, debido a la falta de estrategia y de marco conceptual regional mencionada anteriormente, las conclusiones a las que llega el Laboratorio no están planteadas en una forma que pueda afectar la toma de decisiones de política. Además, se percibe que los resultados no están “contextualizados” en el continente, en parte por la ausencia de estándares regionales, y porque el proyecto no ofreció un marco interpretativo de los resultados con relación al alcance y cobertura de las reformas y políticas educativas implementadas en la última década. Es decir, si bien se estudiaron variables asociadas al logro académico, esas asociaciones no están contextualizadas en un marco conceptual de política educativa regional.

En **Cuba** existía la percepción de que los resultados serían altos debido a los muchos años de una política sectorial orientada a la distribución masiva y equitativa de oportunidades educativas. Además, se sabía que Cuba tenía, regionalmente, el más alto índice cultural de padres de familia, así como programas de salud infantil de amplia cobertura. Se esperaba con ansiedad los resultados del segundo informe, y se recibió con beneplácito la información sobre los efectos positivos del clima escolar sobre los aprendizajes, lo cual ellos interpretan como resultado de varias décadas de políticas sociales en su país que fomentan la solidaridad. La difusión que se dio a los resultados del primer informe tuvo un marcado, y deliberado, cariz político, en tanto se difundió por la prensa radial y televisiva, y entre los participantes del Congreso de Jóvenes Comunistas, como un logro significativo de la política social cubana.

En **Perú**, al tiempo de publicarse los datos, Fujimori aún se aferraba al *slogan* de “hacer de la educación peruana la mejor de América Latina”, y los resultados de la prueba comparada podían ser interpretados como evidencia de que poco o nada se estaba avanzando hasta entonces en esa dirección. Desde la conducción del Laboratorio, y también por parte de los representantes nacionales por Perú, se presumía desde antes de la conclusión del estudio que el gobierno no tenía intenciones de publicar los resultados. Sin embargo, los representantes peruanos también expusieron razones técnicas por las cuales se requerían más detalles técnicos sobre el análisis de los datos, y se solicitaba un plazo adicional antes de publicar los resultados internacionales. Anteriormente, y durante la fase de diseño y aplicación de pruebas, Perú ya había realizado algunas observaciones técnicas sobre los instrumentos, que al parecer no fueron tenidas en cuenta. Finalmente, la petición de un plazo adicional antes de la publicación fue interpretado como que Perú no autorizaba la publicación de sus resultados y el Laboratorio liberó el primer informe sin los datos peruanos. En un contexto de mucha tensión política en el Perú, donde las críticas al autoritarismo del gobierno de Fujimori se venían intensificando, la no-publicación de los resultados fue rápidamente interpretada como una privación al derecho

constitucional de acceso a la información pública. A pesar de que existieron razones técnicas fundamentadas por las cuales se solicitó que los resultados no se publicaran, los coordinadores nacionales del Perú son conscientes de que probablemente nunca hubo intención de publicar por parte del gobierno. Este hecho coincide con la no-publicación de los resultados de las pruebas nacionales, con lo cual se incrementan las razones para pensar que existía la intención de ocultar cualquier información desfavorable al gobierno. Las primeras demandas de publicación se realizaron en el Congreso de la República, en la voz de un congresista de oposición, y fueron seguidas por la demanda de información de al menos dos periodistas de televisión que sabían de la participación de Perú en la prueba del Laboratorio. En esa oportunidad, el Ministerio nuevamente esgrimió razones técnicas y no ofreció los resultados públicamente mientras duró la presidencia de Fujimori. Esos resultados fueron liberados luego, durante el Gobierno de Transición, mediante una publicación de la Unidad de Medición de la Calidad Educativa en el Ministerio.

### **Etapa: Decisiones, impacto político, y nuevas líneas de investigación**

Los responsables por la ejecución del estudio en el Laboratorio, como se expresó anteriormente, coinciden con algunos de los coordinadores nacionales en que la falta de una estrategia fuerte de difusión determinó que los resultados de la prueba no tuvieran un gran impacto político, a través de la opinión pública, ni en las decisiones de política educativa de la Región. Las reacciones por parte de los gobiernos fueron variadas, dependiendo de las expectativas y de la coyuntura política de cada país, aunque se reconoce que en general, exceptuando Cuba, fueron de “tono bajo”. Sin embargo, se espera que con la difusión del segundo informe sobre factores asociados y con el paso del tiempo, los datos comiencen a ser utilizados por investigadores nacionales e internacionales para análisis secundarios y proyectos de investigación, probablemente comisionados por agentes no gubernamentales, así como por académicos y organizaciones sindicales.

Luego del estudio general, algunos países han realizado aplicaciones independientes de las pruebas, como por ejemplo República Dominicana, Colombia y Costa Rica. Éstas han sido experiencias de investigación, no de exploración oficial para informar decisiones de política. Una forma en que los resultados pueden actualmente estar difundiéndose es por intermedio de los representantes nacionales que aún hoy siguen ocupando puestos técnicos o jerárquicos en los Ministerios de educación, o que se encuentran en medios académicos que les permiten difundir la información a audiencias especializadas. Al momento de realizar las entrevistas que sustentan este trabajo de investigación, se esperaba que el segundo informe sobre factores asociados llegara a manos de quienes puedan influir y decidir sobre políticas educativas, como por ejemplo para promover mejores relaciones humanas entre

alumnos dada la evidencia que ofrece el estudio sobre la influencia del clima escolar sobre los aprendizajes.

En **Argentina** no se percibieron grandes reacciones en el terreno político. En el ámbito del Ministerio de Educación particularmente, hubo escasa repercusión pues los resultados llegaron poco tiempo antes de que renunciara todo el equipo de alta dirección por diferencias políticas con la Presidencia. Se considera que el segundo informe, probablemente, habría permitido pensar en investigaciones de mayor profundidad. Pocos funcionarios de la gestión anterior, inclusive del área de medición, se desempeñan actualmente en el Ministerio de Educación, con excepción de algunos técnicos de escasa representatividad política. Sin embargo, muchos de esos funcionarios circulan actualmente en medios académicos, desde donde tienen cierta influencia en la formación de nuevos profesionales y especialistas en educación. Al igual que en otros países de América Latina, algunos de los funcionarios que accedieron a los datos finales decidieron visitar escuelas cubanas para obtener evidencia más directa de los factores escolares y sociales que explican los altos rendimientos académicos de ese país.

La difusión de los resultados en **Colombia** también fue débil y por lo tanto también lo fue el impacto sobre la educación colombiana. El gobierno no tenía recursos para publicar suficientes documentos, y sólo se entregaron copias del informe de OREALC a gobernadores o secretarios de educación de departamentos. Sólo el ICFES tomó la iniciativa de profundizar el análisis de los datos, desarrollando un marco conceptual *a posteriori* que permitiera analizar los datos con referencia a criterios establecidos por consultores contratados, en vez de utilizar la tabla original de especificaciones que sólo permitía un análisis referido a normas. Luego se pidió a los especialistas del Laboratorio que criticaran ese nuevo marco conceptual, el cual fue considerado apropiado pero se observó que la calidad de los ítemes no estaba a la altura de ese marco. El resultado de ese estudio no fue divulgado ni discutido con otros países. Según los técnicos del ICFES, habría existido mayor preocupación por difundir los resultados si la prueba hubiera estado referida a criterios y si hubiera estado sustentada por un marco conceptual sólido que permitiera capacitar a los docentes en interpretaciones significativas desde el punto de vista epistemológico y pedagógico. Dadas las condiciones actuales, se considera que el docente colombiano recibe información mejor y más relevante por parte del ICFES en el ámbito nacional que lo que podía ofrecer la prueba internacional comparada.

En **Cuba**, además de la difusión que hizo el gobierno de los excelentes resultados comparativos, se decidió premiar a los docentes con un incremento del 30% en sus salarios. Con esta medida también se pretendía desalentar el doble empleo de los maestros, pues los resultados del estudio mostraban una relación negativa entre el multiempleo y altos rendimientos estudiantiles. Finalmente, se creó un plan de formación docente de emergencia para

compensar las pérdidas de los docentes que habían “emigrado” hacia el sector turismo en busca de mayor remuneración. En el terreno curricular, se decidió trabajar con más diversidad de textos y se incentivó la producción de mayor variedad de materiales educativos.

Para **Perú**, quizás el impacto más evidente de la participación en el estudio internacional del Laboratorio fue la publicación de los resultados, que finalmente tuvo lugar durante el último gobierno de transición y que se considera un símbolo importante de la democratización del sector. Además, se observa que diferentes medios y actores sectoriales citan frecuentemente los bajos resultados del Perú con respecto a la Región para sustentar diferentes argumentos sobre prioridades de política educativa.

<b>Agencia: IEA - Pruebas: TIMSS, TIMSS-R y Educación Cívica.</b>
---

### **Etapa: Descontento**

Tanto en **Argentina** como en **Chile**, la participación en las pruebas de Educación Cívica y de Ciencias y Matemáticas de la IEA, se decidió con plena conciencia de que los resultados con respecto a los países desarrollados iban a ser extremadamente bajos. Se sabe, por las pruebas nacionales, y en el caso de Chile por participaciones internacionales anteriores, que existen importantes falencias por superar, tanto con relación al currículo prescrito como al implementado, en las áreas evaluadas.

En **Estados Unidos**, la participación en TIMSS estuvo precedida por el aval del Presidente George Bush (padre) a la propuesta de los gobernadores de estados, reunidos en cumbre en el año 1989 para proponer los objetivos “Goals 2000” que orientarían muchos de los esfuerzos de reforma educativa en la década de los 90. Uno de los objetivos nacionales explícitos a principios de esa década fue el de llegar al primer puesto en los *rankings* internacionales de logro académico en matemáticas y ciencias. Ya desde entonces Bill Clinton era el gobernador de mayor renombre. Él mismo organizó la agenda de la Cumbre Nacional de Educación y luego como Presidente dio particular ímpetu a la consecución de las metas de mejoramiento y logro académico. Esta percepción de que Estados Unidos está en desventaja internacional debido al bajo rendimiento de sus alumnos tiene origen en al menos dos fuentes de información que impactaron en el escenario político: por un lado, los resultados del SIMS (Segundo Estudio Internacional de Matemáticas, IEA), a partir de los cuales NCTM, Consejo Nacional de Profesores de Matemáticas, comienza a elaborar nuevos estándares en esa área; en segundo lugar, tiene singular efecto sobre la academia un estudio publicado por Curtis McKnight (1987), que refiere principalmente a los bajos resultados obtenidos por los estudiantes estadounidenses en matemáticas con respecto a algunos países asiáticos, y a

las escasas oportunidades curriculares de aprendizaje que pueden explicar esos resultados.

En el nivel estadual y local, los resultados de TIMSS dieron origen a una serie de iniciativas organizacionales, entre ellas la de participar en TIMSS-R, para elevar los resultados de aprendizajes en ciencias y matemáticas. Así, en un intento por sumar esfuerzos y aunar criterios curriculares y de política, se crean consorcios educativos en diferentes estados del país<sup>v</sup>. Uno de estos Consorcios, en el estado de Ohio, se dio a llamar SMART (Science and Mathematics Achievement Required for Tomorrow) y a él se harán referencias sistemáticas a lo largo de este informe, como ilustración de casos de unidades subnacionales que participaron en TIMSS-R independientemente de Estados Unidos como nación. La creación del Consorcio SMART tuvo origen en los magros resultados obtenidos por Estados Unidos en TIMSS. Una vez que se creó el Consorcio, con el objetivo central de elevar el rendimiento en ciencias y matemáticas, aparece la invitación a participar en TIMSS-R y las autoridades del Consorcio aceptan sumarse al proyecto de evaluación.

Por su parte, la organización sindical AFT (Federación Estadounidense de Profesores), jugó un rol importante durante el proceso de evaluación TIMSS en Estados Unidos. Según la sub-directora de asuntos académicos de esa organización, estudios comparativos anteriores a TIMSS habían demostrado que hacía falta elevar las expectativas de aprendizaje en el país y la AFT, en respuesta a esa necesidad, había producido una serie de libros con ejemplos de exámenes tipo que se toman en otros países, como Francia y Alemania, en diferentes áreas de conocimiento. Los representantes de AFT sabían que existían brechas importantes en el rigor y la profundidad con que se tratan los contenidos académicos en la escuela básica, y también estaban informados de que el gobierno federal había planteado el desafío de llegar a ser primeros en el mundo en el rendimiento académico en ciencias y matemáticas.

### **Etapas: Iniciativa y adopción**

**Argentina** tuvo un intento truncado de participación en TIMSS, debido a que la iniciativa fue individual, por parte de una universidad regional, pero no obtuvo el aval ministerial para continuar con el proyecto. Se consideró entonces que en el medio de una transformación curricular no se podía asumir “una prueba internacional con estándares fijos”. Los estándares propuestos por TIMSS, según perciben los entrevistados, eran más altos que los que se planteaba la reforma internamente, además de que las provincias iban aplicando los nuevos contenidos curriculares a tiempos diferentes. Para el momento en que el Ministerio debió decidir la participación, los estándares de TIMSS estaban acordados y no valía la pena insertarse en un proceso donde “ya estaba todo decidido”. Según los técnicos del Ministerio en ese momento, había que “tener un sistema nacional muy fuerte, con alta credibilidad, para realizar este proceso



de aplicación internacional”. El sistema de evaluación nacional estaba preocupado entonces por el lugar relativo que ocupaban las provincias en el escenario nacional, no por la posición de Argentina en el mundo. Se trataba de generar políticas para la igualdad, acciones concretas de política y para eso, se argumenta, no era necesario participar en una prueba internacional que confirmara que se estaba lejos de lograr el rendimiento de los países más industrializados.

Ese escenario ha cambiado en Argentina, y los representantes actuales ante la IEA consideran que el sistema de medición nacional está lo suficientemente afianzado y que no se debe temer a la comparación con países altamente desarrollados. Por lo contrario, se considera que sólo el efecto de “*shock*” puede movilizar al gobierno y a la opinión pública a efectuar cambios efectivos. Por esta razón es que el IDECE se ha embarcado simultáneamente en los proyectos PIRLS, Educación Cívica, TIMSS-R y TIMSS 2003, todos administrados por la IEA. Complementariamente, se decidió la participación en PISA y ALL para poder cubrir los tres grupos etáreos principales de la población (niños, adolescentes y adultos). El interés por la participación en el estudio de Educación Cívica deriva de la percepción del equipo que, en un momento político tan complejo para la Argentina, es importante saber cuál es el rol que cumple la escuela en la transmisión de valores y actitudes cívicas. Es de interés del IDECE capacitar equipos técnicos en psicometría y modelos estadísticos mediante la participación en estas pruebas, pues no hay en Argentina universidades especializadas en esas áreas.

En **Chile**, la participación en las pruebas de Cívica y TIMSS-R se dio en medio de un proceso de elaboración curricular intenso y prolongado, donde interesaba especialmente saber cómo en los países más desarrollados se logra alinear los contenidos curriculares complejos con instrumentos de evaluación adecuados. La prueba de Cívica se aplicó a una población de estudiantes que accedieron al currículo de educación cívica anterior, con lo cual se pudo obtener una línea de base antes de empezar a implementar el nuevo currículo de educación cívica, en un momento de democratización nacional después de muchos años de dictadura. Se considera de mucha utilidad tener este referente curricular externo para ver cómo evolucionan los estudiantes al ser expuestos a nuevos contenidos curriculares en la materia. Además, la prueba permite medir variables contextuales y escolares que inciden sobre la formación cívica, como por ejemplo las prácticas pedagógicas de la escuela o el contexto familiar. Se esperaba que la participación contribuyera a reforzar las capacidades nacionales de medición. Es importante destacar que la participación de Chile en la prueba de Cívica se decidió intersectorialmente, entre los Ministerios de Educación, Trabajo y Economía. La intención detrás de esta decisión no es comparar resultados competitivamente con otros países sino comparar el rendimiento de Chile con respecto a estándares internacionales.

La participación de **Colombia** en TIMSS fue liderada, y oficialmente representada, por un investigador de la Universidad del Valle, en Cali, donde dirigía un centro de investigación y desarrollo de materiales educativos. La idea original para que participe América Latina en la prueba la propuso un investigador canadiense, Richard Wolfe, entonces miembro del Comité Ejecutivo de TIMSS. Él fue quien promovía la idea de que América Latina se comparase con estándares internacionales y organizó una primera reunión en Venezuela donde invitó a 10 ó 12 países a participar en la prueba. Las primeras reacciones a la invitación fueron de escepticismo porque América Latina se sabía lejos de los estándares internacionales, y por lo tanto no había objeto en la comparación cuando las condiciones eran tan desiguales. Colombia percibió rápidamente que el estudio era importante para conocer la realidad de la educación en ciencias y matemáticas, en el contexto de un estudio que se presentaba como muy bien organizado y en el cual se veían representados curricularmente. También consideraron que el estudio permitiría revelar aspectos del sistema educativo nacional que no se habían estudiado antes. Para alentar la participación de los países invitados, y con la intención de ahuyentar los fantasmas sobre la excesiva distancia que habría entre América Latina y el mundo industrializado, Richard Wolfe sugirió hacer un estudio piloto con las pruebas del SIMS con algunas pocas instituciones en cada país, a modo exploratorio. En Colombia se llevó a cabo el estudio piloto, arrojó como resultado un 50% de respuestas correctas, y con eso se probó que se podía participar sin correr el riesgo de “efecto suelo” en la medición contra países desarrollados. La Universidad del Valle no consiguió al principio el aval oficial para afiliarse a la IEA, sin lo cual no podía aprobarse la participación de Colombia, pero finalmente la IEA le permitió continuar, aun sin el consentimiento del Ministerio de Educación. Justo antes del pilotaje general, cuando había que decidir la participación definitiva en el proyecto, la entonces Directora de Organización Escolar en el Ministerio de Educación decidió que era necesario participar en la prueba de TIMSS y obtuvo los fondos necesarios con la colaboración de la OEI.

Desde el ICFES, que también tuvo una fuerte participación técnica en el proceso de desarrollo y aplicación de la prueba TIMSS en Colombia, se percibe que participar en experiencias de evaluación de la IEA es positivo en varios sentidos. En primer lugar, el cronograma de desarrollo, aplicación y devolución de datos está perfectamente organizado y los plazos son inamovibles. En segundo lugar, las pruebas tienen un director de proyecto que es distinto del director general de la IEA, con lo cual se aseguran que la dirección del operativo esté enfocada en los aspectos exclusivamente técnicos del proyecto y no en cuestiones de política institucional ni en conseguir los recursos para poder continuar con las tareas programadas. La participación en un proyecto de evaluación de la IEA es considerada por el ICFES como una oportunidad de aprendizaje, no sólo de los aspectos psicométricos y estadísticos de la medición, sino también de las estrategias organizacionales desplegadas para poder cumplir eficientemente con el plan de proyecto. A pesar de las ventajas de la participación en estas pruebas, y según profesionales del ICFES y de la Universidad del Valle, las autoridades

actuales del Ministerio de Educación consideran que ya se han realizado suficientes mediciones internacionales como para saber en qué condiciones está Colombia con relación a otros países. Esta decisión, sin embargo, refleja que no se están tomando en cuenta las posibilidades de información que ofrecen las pruebas comparadas más allá del *ranking* internacional de posiciones.

En **Estados Unidos**, la decisión de participar no fue realmente meditada. Correspondía al ciclo de evaluaciones de la IEA aplicar una nueva prueba y Estados Unidos participa sistemáticamente en todas las aplicaciones de esa asociación. A principios de la administración de George Bush (padre), como se explicó en la sección anterior, Estados Unidos comenzó a observar que Japón estaba tomando ventaja económicamente. Allí comenzó a cobrar mayor importancia el logro académico en matemáticas y ciencias, y Estados Unidos decidió que financiaría la coordinación internacional del estudio sólo si se incorporaba a la prueba el área de ciencias. Así sucedió, y el estudio dejó de ser TIMS para ser TIMSS, incluyendo tanto matemáticas como ciencias naturales. También se sumaron a la prueba algunos estados y distritos escolares del país que participaron como unidades organizacionales independientes, al margen de su participación en la muestra general de Estados Unidos, para poder compararse individualmente frente al mundo. La razón principal para esta participación independiente era medirse con estándares internacionales y con sistemas más competitivos, pues los referentes nacionales se consideraban demasiado bajos como para ofrecer una buena medida de comparación.

La participación de muchos otros países en TIMSS se dio por cambios importantes en el escenario político mundial. Por ejemplo, los países de Europa oriental que querían entrar en la escena internacional pero que no podían ingresar a espacios institucionales como la OTAN, se adscribieron a la IEA como una forma de comenzar a insertarse globalmente. Pero para muchos de esos países, al igual que otros en Asia, el área de ciencias no era importante, por lo que Estados Unidos tuvo que realizar algunos esfuerzos de persuasión para que países como Singapur se midieran también en esa área y le permitieran a América del Norte compararse con estos grandes competidores asiáticos del mercado internacional.

El Consorcio SMART, al cual se ha hecho referencia anteriormente, decidió su participación en TIMSS cuando el Departamento Estadual de Educación de Ohio ofreció financiar la participación del estado en TIMSS-R, y para esto invitó a los distritos del Consorcio a formar parte de la muestra. Los distritos participantes debían representar, en conjunto, la conformación del estatus socioeconómico y racial del estado de Ohio. No todas las escuelas del Consorcio participaron, pues la convocatoria se cerró cuando se completó la matriz para la muestra deseada. Esa muestra fue estratificada e incluyó unos 1000 casos (estudiantes) en total. En uno de los distritos participantes, la razón principal que los motivó a inscribirse en el proyecto fue que deseaban contar con una línea de base, en términos de información, que pudiera utilizarse para alinear el currículo del

distrito con los estándares internacionales que planteaba TIMSS, y para conocer mejor las prácticas docentes y los patrones de aprendizaje de sus alumnos para luego desarrollar estándares de contenido y desempeño en las áreas evaluadas. Para el Consorcio no fue necesario persuadir a las escuelas, pues los distritos se sumaron con entusiasmo al primer llamado. La decisión de participar la discutió cada distrito con sus directores y docentes, y la estrategia para esa discusión fue diferente en cada uno de esos distritos. No había temor a las comparaciones entre escuelas o distritos dentro del consorcio porque las características de las muestras no las permitía.

En el ámbito nacional, si bien la decisión de participar en TIMSS no fue discutida con las asociaciones sindicales docentes, AFT fue invitada a formar parte de varios comités asesores durante el proyecto. En la percepción de sus líderes, esa invitación se debió a que la organización es apreciada por su reputación profesional, por su seriedad al defender la necesidad de elevar las expectativas de logro en matemáticas y ciencias, y por la comunicación permanente que mantiene con los docentes norteamericanos.

Respecto a la posibilidad de participar en TIMSS, **Uruguay** opinó en su momento que no se justificaba el esfuerzo y la inversión de recursos en compararse con países tan distantes geográfica, cultural y económicamente. Si bien se tiene en cuenta que los estudios de la IEA tienen valor analítico más allá de las comparaciones competitivas, y de que la participación puede desarrollar capacidades técnicas valiosas, la participación y la difusión masiva de los resultados mismos, particularmente de los *rankings* de logro, habrían significado un *shock* demasiado grande debido a las fuertes resistencias ideológicas a la evaluación externa que recién comienzan a superarse en el ámbito de las evaluaciones nacionales. En otras palabras, si bien se sabe que han perdido oportunidades de aprendizaje, también se han ganado oportunidades para afianzar la credibilidad del sistema nacional.

### **Etapas: Implementación**

**Chile** ha rescatado varios aprendizajes del proceso de implementación de las pruebas de la IEA. En el caso del estudio de Educación Cívica, por ejemplo, algunas de las capacidades desarrolladas estuvieron relacionadas con las estrategias operativas en el trabajo de campo, con diseños de manuales, aplicación de instrumentos y análisis estadístico. Ese desarrollo de capacidades deviene tanto de la interacción con expertos internacionales como de los procesos técnicos que se experimentan durante el proceso de evaluación. Se considera muy valioso lo aprendido con relación a los mecanismos utilizados por la IEA para asegurar el control permanente de la calidad y validez de los datos, o los requisitos estadísticos y de análisis que impone el estudio, en términos de procedimientos y de tecnología como, por ejemplo, en la selección de los sistemas informáticos (software) más apropiados para realizar análisis

multivariados de grandes muestras. En cuanto a los contenidos de la prueba, Chile consideró que el examen no estaba abordando en profundidad algunos de los temas que al país le interesaban particularmente. Por ello se hizo uso de la posibilidad de incluir una opción nacional y se elaboró un conjunto de ítemes especiales, aunque respetando los dominios temáticos de la prueba general. Esos ítemes se aplicaron al mismo tiempo que la prueba internacional. También hubo la posibilidad de retirar del cuestionario las preguntas sobre afiliación política, algo inaceptable en una encuesta avalada por el Ministerio de Educación. Durante el estudio de Cívica participaron básicamente técnicos y algunos expertos en las disciplinas sociales que desde antes trabajaban con el SIMCE. A pesar de que Chile ingresó tarde al proyecto, en las reuniones internacionales hubo oportunidad de discutir los ítemes individualmente, antes y después de la aplicación piloto.

En **Colombia**, los primeros pasos del proyecto TIMSS se tomaron con financiamiento propio de la Universidad del Valle. En esa primera etapa se llevó a cabo la investigación del marco contextual de la educación en Colombia, tal como lo requería TIMSS, tarea que resultó compleja pues las estadísticas en Colombia estaban muy atrasadas. La segunda fase consistió en la construcción del marco de referencia curricular de TIMSS. El trabajo “grueso” de esta etapa lo realizó un comité internacional en el cual Colombia no participó. Sólo recibieron instrucciones de cómo sería la codificación, en cada país, de las guías curriculares, textos y entrevistas con expertos en currículo, de manera de establecer cuál sería el currículo prescrito de referencia para Colombia. Luego se realizaron los estudios piloto de las pruebas y la evaluación de los ítemes del banco de TIMSS, de manera tal de evaluar el comportamiento de cada ítem en cada país. Para la definición de la muestra definitiva se convocó al ICFES. En ese momento se consiguió el aval y la cooperación financiera del Ministerio de Educación y, ya con la muestra definida, se decidió la participación definitiva en la aplicación general de la prueba. Por razones financieras, Colombia no participó en el relevamiento ni análisis de procesos instruccionales (v.g. videograbaciones de sesiones de clase). A lo largo del proceso de medición participó una docena de profesionales, como profesores del Taller de materiales didácticos de la Universidad del Valle, profesores del Instituto de Educación y Pedagogía, y profesores de algunos colegios de educación básica. En la medida en que se complejizaban las tareas vieron la necesidad de establecer una oficina para el proyecto, lo cual consiguieron hacer con fondos de la OEI y del PNUD. El mismo equipo codificó y calificó todas las preguntas abiertas siguiendo los manuales provistos por TIMSS. Todos los procedimientos estaban detallados en manuales, con lo cual se evitaban confusiones procedimentales, y cuando había dudas se solucionaban mediante la comunicación permanente con el centro de procesamiento de TIMSS. El ICFES mostró al principio cierta renuencia a aplicar preguntas abiertas, pues no creían que fuera posible calificarlas estandarizadamente. Sin embargo, una de las enseñanzas que dejó TIMSS es que pueden desarrollarse criterios adecuadamente fundamentados para la calificación objetiva de ese tipo de preguntas.

En el caso de la prueba de Educación Cívica, es importante destacar que la IEA invita a los países, hacia el principio del proyecto, a discutir el marco conceptual que guiará la construcción de las pruebas, y que luego permite realizar análisis más sustantivos de los datos recogidos. Cada país contrata a sus propios consultores para validar ese marco en el contexto nacional. Luego se envían los comentarios a la IEA, donde son revisados y utilizados para construir los instrumentos. Finalmente, la IEA convoca a una nueva reunión de países para discutir los ítemes propuestos. Colombia tuvo la oportunidad de participar en todo ese proceso, que es considerado de alto valor por los técnicos responsables nacionales.

El caso de **Estados Unidos** es peculiar en toda la experiencia TIMSS porque fue ese país quien, además de financiar la coordinación internacional, ejerció el liderazgo técnico del proyecto. Si bien hubo un comité técnico de participación internacional, toda la conceptualización del estudio y el desarrollo de los “*blueprints*” estuvo en manos de la Universidad Estatal de Michigan. Aun así, se daba representación a otros países en algunos aspectos especiales del estudio, como por ejemplo a Inglaterra en la evaluación de demostraciones prácticas, o a la organización *Statistics Canada* en cuestiones de muestreo. Debido a que las implicancias políticas de este estudio serían altísimas, el gran desafío de Estados Unidos fue asegurarse de que las muestras de todos los países fueran representativas. Según asegura uno de nuestros entrevistados, el rigor de los estándares de muestreo utilizados por TIMSS marcó un hito para ese aspecto de los estudios internacionales de logro académico.

Un aspecto relevante de esta fase de TIMSS en Estados Unidos es que se involucró activamente a los gremios de profesores y asociaciones profesionales de las disciplinas bajo estudio. Representantes de estos grupos fueron parte del comité asesor, lo cual contribuyó significativamente a la sensibilización de las audiencias académicas y profesionales de la educación, y aunque el centro de investigación en la Universidad Estatal de Michigan tenía una posición propia con respecto al diseño de la prueba, nunca dejó de consultar al Comité Asesor antes de tomar decisiones. Desde la conducción técnica del estudio en Estados Unidos, se consideraba que involucrar a las asociaciones profesionales era clave para lograr que los resultados tuvieran efecto en la política educativa. Por esa razón, a pesar de que las diferencias de opinión entre grupos complejizaron la tarea en algunas oportunidades, se decidió respetar siempre las intervenciones del Comité Asesor. Éste además de contribuir a la sensibilización de una audiencia particularmente importante, protegió al equipo de conducción de TIMSS en Estados Unidos contra algunas decisiones políticas de cúpula que pudieron afectar seria y negativamente los objetivos del proyecto.

En el Consorcio SMART, la participación técnica se centró básicamente en el análisis estadístico de datos. Para esto se contrató a un experto que analizó y criticó la información, y luego realizó recomendaciones sobre la presentación de

los datos. Algunos docentes, directores y profesores universitarios participaron en el análisis de esos datos antes de ser difundidos. Allí se decidió qué tipo de variables se estudiarían más en profundidad (tarea en casa, tiempo efectivo en clase, actitudes de los docentes hacia la disciplina, currículo prescrito vs. implementado). En el nivel distrital, se reunieron los directores de los departamentos de ciencias y matemáticas para discutir la alineación curricular entre escuelas, y entre el distrito y los estándares estatales. Se considera que el tipo de aprendizaje más valioso obtenido a través de la participación en TIMSS fue el relacionado al análisis e interpretación de datos. Todos los datos fueron analizados dentro del ámbito del Consorcio, para lo cual previamente se envió a algunos miembros del personal a entrenarse en un seminario que TIMSS ofrecía para que los participantes se familiarizaran con la base general de datos.

La participación de AFT en la fase de implementación de TIMSS estuvo relacionada con la discusión de los reportes preliminares del estudio para sugerir el tipo de información que ellos consideraban importante que se reportara al público. Además, tuvieron una activa participación en el comité consultivo de James Stigler para el análisis de los videos, así como en el comité consultivo para el diseño del reporte principal.

### **Etapas: Interpretación y diseminación de resultados**

En **Argentina**, el equipo técnico es consciente de que los resultados de las varias pruebas de IEA no van a ser favorables en comparación con otros países, especialmente en TIMSS-R y TIMSS 2003. Se preparan para un efecto de *shock* en la opinión pública, pero, como se explicó antes, ése se considera un efecto deseable si se quiere comenzar a cambiar el estado de la situación. Recientemente, y quizás como parte de una estrategia de sensibilización a los próximos resultados de la prueba de Educación Cívica, el ministro de Educación encabezó una Jornada Académica sobre "Evaluación de la Educación Cívica", realizada por la Academia Nacional de Ciencias Morales y Políticas. Del encuentro participaron varios asesores del Ministerio de Educación, funcionarios del IDECE y miembros de la academia y de organizaciones no gubernamentales. Los participantes se refirieron a las principales áreas del estudio, y realizaron análisis y aportes para la interpretación de los resultados de la prueba, que aún no han sido publicados.

Con respecto a los resultados de esta misma prueba, los funcionarios técnicos en **Chile** interpretan que, de acuerdo a los niveles de inversión en educación en Chile, así como al nivel cultural de los padres, el rendimiento de los alumnos chilenos en la prueba de Educación Cívica debió ser más alto. En las áreas de ciencias y matemáticas, por su parte, los docentes manifiestan gran inseguridad en sus conocimientos, lo cual indica por dónde habrán de orientarse algunas políticas de mejoramiento. En Chile no hay una organización institucional en el SIMCE ligada a la difusión de los resultados de pruebas internacionales. Los resultados de TIMSS-R fueron publicados solamente en forma de *ranking* por un diario nacional, mientras que la Unidad de Curriculum y Evaluación preparó un

documento de difusión resaltando los aspectos más sustantivos de los datos recogidos. A pesar de que la unidad de medición del Ministerio de Educación chileno está políticamente afianzada, la difusión de resultados de pruebas nacionales ha demostrado a esta unidad que se debe guardar extremo cuidado al momento de redactar los resultados de los estudios. Consideran que poco a poco las diferentes audiencias y la prensa aprenderán a realizar una lectura más interpretativa (es decir, no exclusivamente competitiva) de los datos, lo que permitirá gradualmente enriquecer el debate nacional sobre los aprendizajes en vez de polarizarlo.

Para el caso del estudio de Cívica, que incluyó una opción nacional, los resultados que primero se difundan incluirán todos los datos comparativos internacionales de base. Los datos de la aplicación diferenciada (ítemes de opción nacional) serán difundidos y analizados posteriormente. Los datos internacionales comparados serán desagregados por dependencia (escuelas municipales, subvencionadas y particulares). También en este caso el desafío mayor para el Ministerio será la difusión, de manera tal que no se lea solamente el *ranking* internacional sino que los resultados de la prueba sean un estímulo para elevar los estándares en la disciplina, es decir que impacte en el ajuste de expectativas para docentes y alumnos de educación cívica, y en virtud de esta expectativa es que se está organizando una capacitación para 40.000 profesores. Dado que después de un año y medio las bases de datos se hacen totalmente públicas, se espera que la comunidad académica las utilice y así contribuya a una mayor difusión de la información.

Con los resultados de Cívica en **Colombia** se hicieron talleres y seminarios para las escuelas participantes. Se hace evidente que es necesario superar la concepción que divulgación equivale solamente a documentos impresos, y comprender que la verdadera divulgación ocurre cuando se enseña a los actores a *interpretar* resultados. Un aspecto positivo de este estudio fue que durante la fase inicial se requería que cada país realizara algunos estudios de caso, lo que le permitió a Colombia contextualizar sus resultados dentro de un marco conceptual internacional sobre civismo y democracia. La IEA recoge toda esa información inicial de los países participantes y publica un primer documento que los países pueden traducir y difundir ampliamente mucho antes que salgan los resultados comparados de la prueba.

En el caso de TIMSS, y dado que el Ministerio de Educación había contribuido sustancialmente a la financiación del estudio, hubo también interés oficial por conocer y divulgar los resultados de la prueba. En la ciudad de Cali, antes de la publicación de los resultados, se realizó un seminario de casi una semana en el cual participaron funcionarios del Ministerio de Educación y del ICFES. La intención era comenzar a preparar a la audiencia en cargos directivos para comprender mejor los resultados y sus implicancias para la política educativa, y para ahuyentar el escepticismo que se percibía sobre las posibilidades de comparar países tan desiguales como pueden ser Colombia y Japón. En



general, sin embargo, los responsables técnicos del estudio no percibieron que los funcionarios estuvieran demasiado temerosos de los efectos negativos de los resultados en la opinión pública. El Ministerio de Educación financió algunas publicaciones especiales para maestros de matemáticas.

En **Estados Unidos**, luego de la prueba TIMSS, se publicaron primero los resultados del análisis curricular, y varios meses después se dio a conocer el *ranking* internacional. Esto contribuyó a que, para cuando se supo que Estados Unidos ocupaba un lugar modesto en ese *ranking*, las audiencias más especializadas ya tuvieran un conocimiento importante sobre los aspectos sustantivos del estudio. Los resultados de ese análisis curricular se divulgaron mediante la prensa televisiva, radial, e impresa. El director responsable del equipo técnico de EEUU apareció en varias cadenas de noticias, y frente al Senado de la Nación, comentando sobre los resultados del análisis y sus implicancias, mientras que el co-director de ese equipo ofreció un debate en la *National Public Radio*. Seis meses después se publicó el *ranking* internacional, en el cual Estados Unidos se ubicó alrededor de la media internacional, tanto en ciencias como en matemáticas, y muy por debajo de los temidos competidores asiáticos. A pesar de los desfavorables resultados, sin embargo, las asociaciones profesionales que co-lideraron el proyecto, como la Fundación Nacional de Ciencias (NSF) y el Consejo Nacional de Estadísticas Educativas (NCES), mantuvieron firme su compromiso de publicar la base de datos y de difundirla por todo el país. NCES pagó, y sigue pagando en la actualidad, un taller anual que se suma a las actividades del Encuentro Anual de la Asociación Estadounidense de Investigaciones Educativas (AERA). Este es un evento de amplísima asistencia y difusión en Estados Unidos, y por lo tanto un espacio ideal para la discusión y uso de los datos de TIMSS. Adicionalmente, NCES sigue distribuyendo manuales con los resultados del estudio y otorgando becas de financiamiento para análisis secundarios de los datos.

El TIMSS marcó un cambio en la historia de la IEA, desde el momento en que sus pruebas dejan de ser sólo un riguroso ejercicio académico para convertirse en el proyecto de un conjunto de países cuyos gobiernos se interesan por las implicancias políticas y “de política” de los resultados. Este fenómeno se sintió también en Estados Unidos, y es por eso que a la NSF se la convenció, por primera vez en la historia de esa Fundación, para que su donación a la Universidad Estatal de Michigan permitiera contratar a una prestigiosa consultora de prensa de Nueva York que coordinaría las relaciones públicas y las estrategias de difusión de TIMSS. La consultora proveyó entrenamiento a los coordinadores nacionales y asignó personal especializado a la “colocación” de noticias sobre los análisis de TIMSS en diferentes medios de comunicación del país. Es importante destacar que esa colocación no significa *pagar* por los espacios de prensa, sino identificar los medios de comunicación potencialmente interesados en difundir noticias educativas. Es un servicio muy caro que no podría haberse pagado si la empresa hubiera cobrado precios rigurosos de mercado. Pero las implicancias políticas del estudio y su difusión eran muy altas

y se decidió utilizar el servicio muy temprano, mucho antes que se liberaran los resultados, de manera que el público estuviera sensibilizado sobre los aspectos curriculares del estudio antes de conocer el *ranking*. De hecho, se considera totalmente inusitado el interés público que se despertó por la problemática curricular en Estados Unidos. Cabe resaltar también que en ese país existen muchos periodistas especializados en el área de educación, que tienen largas y cercanas relaciones con las fuentes que les pueden proveer información relevante. Además, la ética profesional de los medios de comunicación hace posible que se pueda facilitar a los periodistas información que no ha sido públicamente liberada de manera que puedan preparar sus informes y columnas de opinión con anticipación.

En Estados Unidos, a diferencia de América Latina, fue relativamente sencillo motivar e interesar a la academia, a las comunidades educativas y a las asociaciones profesionales por los resultados de TIMSS. Con el grado de involucramiento que se dio a los representantes de esos medios en el Comité Asesor del estudio, y con la difusión estratégica de prensa, la comunidad pronto prestó interés y estuvo incluso dispuesta a pagar a los investigadores nacionales de TIMSS para recibir mejor y más detallada información de los resultados. A pesar de esto, siempre existe escepticismo por parte del profesorado, y por lo tanto también se realizaron importantes esfuerzos en muchas localidades por difundir y utilizar la información que permitiera mejorar su currículo y sus logros de aprendizaje en ciencias y matemáticas.

También en el nivel local se realizaron importantes esfuerzos para lograr una mejor interpretación y difusión de los resultados. Los representantes principales de SMART, por ejemplo, asistieron a cuatro encuentros que TIMSS ofreció a los participantes de unidades subnacionales. Allí se analizaron y criticaron los datos y la mejor manera de presentarlos. Una vez analizados los resultados se presentó una sinopsis de los datos para que los docentes los comprendieran más fácilmente. La información resultante de TIMSS no puede relacionarse fácilmente con los exámenes del estado de Ohio pues estos últimos presentan ítemes de cuestionable calidad técnica, además de que cubren demasiados tópicos pero pocos conceptos en profundidad. A pesar de esto, se observan algunos patrones similares en los resultados de ambos tests. En relación a los cuestionarios sobre actitudes o hábitos de estudio, el experto estadístico del Consorcio considera que hay que ser sumamente cuidadoso con las inferencias, pues éstas pueden conducir a tomar decisiones equivocadas. Por ejemplo, los docentes del Consorcio expresaron en esos cuestionarios un alto grado de confianza en su dominio disciplinar de ciencias y matemáticas. Sin embargo, el Consorcio sabe positivamente que gran parte de los problemas de aprendizaje de los alumnos se debe precisamente a los baches de conocimiento disciplinar que presenta el profesorado. Por otro lado, los datos del estudio que se consideran más útiles son los relacionados al currículo prescrito y al currículo implementado. Por falta de tiempo no se ha realizado aún un análisis de ítemes individualmente pero sí se han identificado y analizado diferentes grupos o

frangas temáticas en las dos áreas evaluadas. Para muchos docentes, los datos de TIMSS son menos significativos que los del examen estadual, dado que estos últimos son censales y los resultados se devuelven a todas las escuelas.

En cuanto a la recepción de datos por parte de AFT, los representantes admiten que había cierto temor a los resultados comparativos con otros países, pero a la vez la organización tenía mucho interés por saber el nivel de rendimiento de los estudiantes estadounidenses con respecto a los de otros países. Desde la sede de AFT en Washington, D.C., se organizó una conferencia para los líderes de la asociación en todo el país, en cooperación con el Departamento de Educación de EEUU. Se invitó a representantes de seis países con altos resultados en TIMSS para entender mejor sus sistemas y profundizar en la problemática nacional. El foco de atención en esa conferencia fue la formación docente en diferentes países. Uno de los ponentes fue el director mismo del estudio en EEUU, quien habló en profundidad de las implicancias curriculares de los resultados de TIMSS. Los investigadores James Stigler y Harold Stevenson explicaron los resultados del análisis de los videos de clases, en el cual habían participado profesionales representantes de AFT. AFT había sido invitada por TIMSS a sesiones especiales destinadas a sensibilizar audiencias particularmente interesadas. Entre las estrategias que utiliza AFT para difundir los resultados de TIMSS, cabe mencionar los cursos de perfeccionamiento para docentes donde incluyen datos de TIMSS y sus implicancias. En esas sesiones se pone especial énfasis en lo que los matemáticos consideran como los elementos clave de una buena clase de matemáticas, y se intenta rescatar toda la información posible mas allá del *ranking* internacional. En general los docentes han respondido muy bien a la evidencia que presenta este estudio comparado. Si bien siempre hay profesionales escépticos sobre el valor de la información, la organización sostiene que, independientemente de las características de la prueba, siempre es posible reconocer cuándo se está enseñando bien, y eso es útil para producir mejoras.

### **Etapas: Decisiones, impacto político y nuevas líneas de investigación**

Los responsables técnicos de **Argentina** que participaron en la gestión ministerial anterior, si bien decidieron no participar en TIMSS, consideran que el estudio podría haber tenido gran impacto en las decisiones de política educativa. Aclaran, sin embargo, que para que esto suceda, la IEA debería reformular sus estrategias de aplicación y análisis de manera tal que la devolución de resultados a los países sea más rápida. Se percibe que los ciclos de las pruebas de la IEA son demasiado extensos y por la tanto, dada la ausencia de políticas de Estado coherentes y continuas, los resultados se devuelven cuando los gobiernos han cambiado y con ellos las prioridades de política. En cuanto a las informaciones que se aguardan de las aplicaciones recientes de pruebas de la IEA, los coordinadores nacionales esperan que los resultados causen alarma y que movilicen a la opinión pública y a decisores de política para realizar cambios y mejoras.

En las oficinas de medición de **Chile**, a pesar de que ya se ha pasado por la experiencia de difundir resultados de la IEA anteriormente, el impacto político que pueden tener los resultados de nuevos estudios sigue preocupando. Esta preocupación está justificada por el rol que ha jugado la prensa en otras ocasiones al difundir casi exclusivamente la noticia sobre el bajo rendimiento de Chile comparado con el mundo, sin prestarle demasiada importancia a los otros datos de valor que ofrecen las pruebas de la IEA. Una revisión de los recortes de prensa chilena (*on line*, [www.reduc.cl](http://www.reduc.cl)) sobre la participación de este país en pruebas de la IEA revela, efectivamente, un tono de alarma y bastante crítico hacia el sistema educativo público, aunque en segundo plano también se presentan algunas informaciones más sustantivas, como por ejemplo sobre algunos de los factores socioeconómicos que explican los resultados. Del total de recortes revisados, es notoria la cantidad de artículos escritos por académicos de uno de los centros de investigación sobre políticas públicas de Chile, que utilizan los datos para hacer críticas y sugerencias concretas sobre la educación pública en el país. Desde el Ministerio de Educación, se está intentando que la información llegue a los profesores, para que haya impacto en el aula y en los contenidos. Con ese fin se están sosteniendo algunas reuniones con funcionarios profesionales del área de formación y perfeccionamiento docente.

En **Colombia**, mientras que la prensa difundió los resultados de TIMSS con títulos más bien escandalosos y poniendo énfasis en el *ranking* de posiciones, el Ministerio de Educación decidió hacer una serie de publicaciones, basadas en los resultados, que ofrecían algunos recursos pedagógicos para que los profesores lograran mejores resultados de aprendizaje con elementos didácticos sencillos. Pero estas mismas publicaciones también mostraban que Colombia ocupaba el anteúltimo lugar en el *ranking* mundial, y terminó teniendo impacto también en el entorno académico y gubernamental. Se presentaron algunas propuestas al Congreso para la creación de un programa de calidad de la educación y de un ente independiente del Ministerio de Educación para evaluar la calidad. La idea de que ese ente fuera independiente no prosperó y por último terminó reestructurándose una unidad dentro del ICFES para realizar las evaluaciones nacionales. Uno de los grandes impactos del TIMSS fue que alentó a los técnicos del Servicio Nacional de Pruebas del ICFES a rediseñar los exámenes, con un marco conceptual de referencia a criterios en vez de normas, y con análisis de respuesta al ítem. Esto permite actualmente contar con mejores indicadores sobre la calidad de la educación en Colombia, y permite a los alumnos y a las instituciones comprender mejor cuáles son las deficiencias y las fortalezas del aprendizaje. Es también parte de los planes del ICFES hacer accesibles las bases de datos a los investigadores, aunque hasta ahora esto no se ha concretado.

En la Universidad del Valle, que como recordamos tuvo a su cargo la coordinación de TIMSS en Colombia, el Programa de Educación de Matemáticas

ha tomado los informes del estudio para hacer actividades de formación y capacitación de profesores. El análisis allí se concentra en los tipos de fallas conceptuales y procedimentales que los alumnos mostraron en la prueba. El coordinador del estudio remitió al Ministerio de Educación una propuesta de actualización docente basada en los datos de TIMSS, pero lamentablemente cayó en el olvido con el último cambio de gobierno nacional.

Los efectos de la participación en TIMSS en **Estados Unidos** han sido muy variados y demasiado numerosos como para describirlos todos en este informe. En forma resumida, sin embargo, podría decirse que los efectos se dieron en dos niveles principales. En el nivel federal y estadual, los resultados de TIMSS confirmaron que hay serios problemas curriculares en matemáticas y ciencias. La ausencia de un currículo nacional determina que la suma de currículos locales (en su mayoría dictados por los libros de texto seleccionados) conforme un mosaico irregular de contenidos, con una gran cobertura de tópicos pero tratados con muy poca profundidad. La evidencia en este respecto puso en marcha una serie de medidas gubernamentales, no gubernamentales y profesionales para establecer estándares curriculares más altos, y para alinear los diferentes elementos de la gestión y la práctica educativa desde una perspectiva sistémica que apuntara a elevar la calidad de los aprendizajes, y lograr los estándares propuestos. En el nivel local o distrital, las reacciones fueron variadas, y también lo fue el tipo de medidas que se tomaron para mejorar los aprendizajes y lograr los estándares establecidos por cada estado. En general, los esfuerzos se centraron en la alineación curricular local con los estándares estaduales y en mejorar las condiciones pedagógicas para la entrega de mejores oportunidades de aprendizaje.

Un efecto singular de TIMSS en el nivel local fue, como se ha descrito en secciones anteriores, la creación de consorcios educativos, como SMART, que luego de participar en TIMSS-R ha decidido proveer extensas oportunidades de capacitación a docentes sobre currículo, métodos instruccionales y, especialmente, sobre cómo analizar datos de pruebas estandarizadas. En el nivel estadual, se están revisando los exámenes estandarizados que, en comparación con TIMSS, cubren demasiados tópicos pero pocos conceptos en profundidad. El Consorcio ha realizado varias propuestas de mejoramiento, pero la decisión final sobre la adopción de esas medidas es, por ley, facultad de los distritos en forma individual.

Finalmente, cabe mencionar los esfuerzos de AFT por incluir datos de TIMSS en sus capacitaciones, y por difundir un mensaje claro en torno a la necesidad de definir en qué consiste un buen currículo, y de lograr aunar criterios nacionales ante la variedad enorme de fuentes curriculares que existe en el país.

### **Etapa: Iniciativa y adopción**

**Argentina** ingresó tardíamente a PISA Plus, es decir que no se registró oficialmente a tiempo para integrarse a la totalidad de las actividades del proyecto, pero se ha registrado a tiempo para PISA 2003. Son dos las razones principales que motivaron la participación en esta prueba. En primer lugar, porque ofrece un enfoque diferente al concentrarse no solamente en habilidades académicas o disciplinares sino en las destrezas cognitivas y cognoscitivas que se consideran fundamentales para desempeñarse competentemente en el mundo laboral de hoy. En segundo lugar, el equipo técnico consideró que los cuestionarios que acompañan la prueba están adecuadamente diseñados y muestran claramente cuál es el objetivo de cada una de sus preguntas. Al igual que con las otras pruebas internacionales en que Argentina está participando, se espera que los resultados sean insumos efectivos de información para decisiones de política.

**Ecuador** no ha participado en PISA pero estuvo interesado en hacerlo. El interés por sumarse a la prueba se dio en un encuentro internacional en el cual se publicitaban las diferentes pruebas disponibles.

La participación de **Perú** en PISA Plus, y probablemente en próximas aplicaciones, se origina en la búsqueda, por parte de las oficinas de planeamiento y de medición del Ministerio de Educación, de un sistema adecuado de indicadores educacionales. La búsqueda los condujo a familiarizarse con la OCDE, particularmente con el proyecto WEI (Indicadores Mundiales de Educación), y con la prueba internacional PISA. Impresionaron positivamente los avances realizados por la OCDE en evaluación, y en la generación y análisis de información educacional, y eso motivó la participación de Perú en la prueba de logro académico. Existió apoyo de la alta dirección del Ministerio para participar, además de que la experiencia no es demasiado costosa. Resultó atractiva la idea de medirse con estándares internacionales, aunque de especial importancia fue la posibilidad de acceder al trabajo conjunto con expertos en el campo de la medición a gran escala, y así contribuir al desarrollo de capacidades institucionales de medición y de formas de uso de la información. Se considera que toda comparación es útil, pero que estudiar lo que sucede en países con alto rendimiento no significa necesariamente adoptar modelos externos para aplicarlos localmente. El valor de la comparación reside en las preguntas que genera sobre cuáles son las vías posibles de mejoramiento. La participación también se decidió por la calidad de los ítemes de la prueba, por ser exigentes e ingeniosos. Si bien algunos de esos ítemes pueden resultar un tanto descontextualizados para las poblaciones rurales del país, se considera que el gran reto para el Perú es lograr, precisamente, el dominio de conocimientos y capacidades cognitivas más complejas, y para eso se requiere el acceso a formas de conocimiento de relevancia global.

En **Uruguay**, sólo a partir de 1999 comenzó a considerarse la posibilidad de participar en un estudio internacional, pensando básicamente en los beneficios que dicha participación pudiera reportar a la unidad de medición (UMRE), en términos de aprendizaje y desarrollo de la capacidad técnica. Entre las opciones consideradas se preferiría participar en PISA, pues es la prueba que ofrece, en la opinión de los técnicos, el esquema conceptual más completo, una buena definición de competencias, y una propuesta clara sobre la utilización de la información de factores de contexto. Si bien mostraron interés por participar, e inclusive asistieron a los primeros seminarios del proyecto, el cambio de administración presidencial no permitió que se tomara la decisión a tiempo y por lo tanto no ingresaron a la aplicación de PISA Plus.

### **Etapa: Implementación**

En la experiencia de **Argentina** con PISA Plus, se observa que el proceso de implementación ha sido básicamente técnico. No hay actores externos involucrados, excepto por los consultores o técnicos que se contratan para tareas muy puntuales, como por ejemplo las traducciones que requiere el proyecto, y la validación posterior de esas traducciones por parte de pedagogos en las disciplinas involucradas. Como Argentina ha ingresado tarde a las aplicaciones de PISA, no se ha podido diseñar opciones nacionales. Sin embargo, se planea utilizar esas opciones en el futuro para poder incluir ítemes de las pruebas nacionales y así poder comparar la información recogida.

En **Perú**, la participación en PISA está resultando costosa en términos de carga de trabajo para un grupo reducido de personal, aunque se considera que vale la pena el esfuerzo. Esta experiencia les da la posibilidad de ensayar con aspectos técnicos y conceptuales diferentes a los de las pruebas nacionales. Debido a que se ingresó tarde en el proyecto, la Unidad está comprimiendo en pocos meses mucho del trabajo que debió hacerse con más tiempo y que habría permitido aprovechar totalmente la experiencia de aprendizaje. Aun así, hubo oportunidades de modificar algunos ítemes para adaptarlos al contexto peruano. Para PISA 2003 los técnicos están teniendo más posibilidades de intervención que en la aplicación 2001.

Para los técnicos del **Uruguay**, serían varias las posibilidades de aprendizaje que ofrece el programa PISA. En primer lugar, permitiría generar más conocimiento para diseño de ítemes fuera de las áreas actualmente evaluadas por las pruebas nacionales. En segundo lugar, permitiría desarrollar capacidades específicas en relación al diseño, calificación y procesamiento de la información de pruebas de respuestas abiertas. Tercero, se valoraría la oportunidad de un aprendizaje más profundo de la teoría de respuesta al ítem, que hasta ahora sólo conocen por referencias. Por último, consideran que si bien ya tienen experiencia en el análisis de factores socioeconómicos asociados al rendimiento,

les sería valioso aprender más sobre la medición de variables escolares, curriculares e instruccionales. Se insiste particularmente en la importancia de participar en pruebas que estén fundadas en un marco conceptual sólido, como en el caso de PISA, donde es posible comprender el por qué del diseño de los ítemes, y donde importa el logro con respecto a competencias previamente definidas como deseables, y no en relación a un agregado de contenidos curriculares de varios países como fue el caso de TIMSS o del Laboratorio, en los cuales la tabla de especificaciones resultó un intento por establecer denominadores comunes pero sin contar con un marco teórico de referencia.

### **Etapa: Interpretación y diseminación de resultados**

En **Perú**, se considera que después de la experiencia con la prueba del Laboratorio es de suma importancia lograr que se difunda información sustantiva, no solamente el *ranking* internacional. Los técnicos nacionales consideran que uno de los principales inconvenientes para la difusión de este tipo de información en el país es que existen pocos expertos en educación que sepan analizar e interpretar datos cuantitativos para luego realizar inferencias sustantivas relevantes. En este sentido, hay otros campos, como la economía o la sociología, donde es más común encontrar profesionales con esas capacidades que contribuyan a la diseminación académica de las informaciones investigativas. Con la intención de superar ese obstáculo, la coordinación del estudio en Perú está planeando un taller de varios días de duración donde se explique a actores clave del sector cómo interpretar los resultados de PISA. Ahora que se percibe mayor estabilidad política en el país, se espera que sea posible elaborar una estrategia efectiva de difusión.

**Uruguay**, si decidiera participar en próximas aplicaciones de PISA, haría uso de esa oportunidad para difundir ampliamente los resultados de la prueba, así como lo que pudiera revelarse del instrumento mismo, tal como se ha hecho con las pruebas nacionales. De esta manera, se podría explicar a cada equipo de profesores, en cada establecimiento, qué es lo que las diferentes actividades de la prueba miden, para que ellos puedan aplicar la prueba, o ítemes seleccionados de la misma, y así comparar el desempeño de sus alumnos con el de los estudiantes en otros países.

### **Etapa: Decisiones, impacto político y nuevas líneas de investigación**

En las oficinas técnicas de medición del **Perú**, actualmente, no existe mayor preocupación por el posible impacto político de los resultados de PISA, quizás por lo que indica en ese sentido la experiencia pasada con la prueba del Laboratorio y con las pruebas nacionales. Se sabe que las noticias de prensa serán escandalosas, pero en la medida en que el país siga participando y el público se acostumbre a que hay otras informaciones rescatables además del *ranking*, los medios de comunicación deberán buscar otro tipo de noticias para



vender a la población. Entretanto, el Ministerio de Educación habrá tenido importantes ganancias en términos de aprendizaje técnico e institucional. Otro aspecto positivo de la prueba es que el país será propietario de los instrumentos y podrán ser aplicados nuevamente si se quiere profundizar las investigaciones.

Los técnicos nacionales de medición en **Uruguay** no descartan que, a pesar del buen uso que se pretendería dar a las pruebas, la medición contra estándares internacionales como los de PISA podría implicar un riesgo político considerable. Por lo tanto, antes de ingresar en una prueba de éstas, y si se tiene la intención real de difundir los datos, es necesario contar con el aval de las autoridades políticas del sector, y de la Presidencia, para asumir el costo de salir en los últimos lugares del *ranking* internacional.

## **DISCUSIÓN**

Como puede observarse, existe ya un conjunto importante de experiencias de participación en pruebas internacionales en América Latina. Esas experiencias han sido variadas y muestran diferentes grados de inserción o impacto efectivos en el escenario de las políticas y la práctica educativa de cada país, dependiendo de los contextos políticos e institucionales en que se inscriben. En esta sección se resume y discute algunos elementos comunes a la experiencia de participación en pruebas internacionales de los países seleccionados. El tamaño de la muestra de países en este estudio, así como el número limitado de actores entrevistados, restringen las posibilidades de hacer generalizaciones amplias a la Región en cuanto a las características de las experiencias en evaluación internacional y sus beneficios. Sin embargo, la cantidad de información obtenida amerita un resumen crítico que reúna los aspectos más importantes de las experiencias relatadas y que facilite la comprensión de los procesos descritos en la sección anterior. En esta sección se presenta ese resumen, organizado de acuerdo a las fases de evaluación comparada internacional utilizadas como marco conceptual en este estudio.

En cuanto a la **fase inicial** del ciclo de política curricular, donde teóricamente se perciben los primeros “descontentos” con respecto a los resultados de aprendizaje obtenidos por los sistemas educativos nacionales, vale destacar que esa percepción en América Latina se basa, al menos parcialmente, en algunos referentes empíricos, tales como los resultados de pruebas nacionales de logro académico, pruebas exploratorias comparadas en la Región, o resultados de pruebas internacionales anteriores. Parte del “descontento” sobre la eficacia de los sistemas educativos, que conduce a la búsqueda de información comparada sobre rendimiento académico, deviene del convencimiento, no del todo fundamentado, de que los grandes problemas económicos de la Región tienen origen, en parte, en las deficiencias de aprendizaje de los estudiantes del sistema, que se gradúan sin estar preparados para las demandas del mundo laboral de hoy. Ese uso de referentes empíricos es decididamente más fuerte en

países como Estados Unidos, donde se ha hecho un uso mucho más extensivo de los datos recogidos en pruebas nacionales y pruebas internacionales anteriores, y donde el temor por perder competitividad en el mercado internacional es más marcado.

Las **iniciativas de participación**, así como la elección del proyecto de evaluación comparada más conveniente para cada país en América Latina, han estado en la práctica definidas por tres objetivos principales. En primer lugar, acceder a mayores oportunidades de capacitación técnica en medición de aprendizajes y diseño de instrumentos. Existe, sobre la base de experiencias anteriores, una expectativa y percepción compartidas entre los países de la región en cuanto a los beneficios de participar en pruebas lideradas por agencias de alta capacidad técnica y organizacional como la IEA o la OCDE. En este sentido, se valora especialmente lo que se puede aprender de la organización operativa de los proyectos, de la calidad técnica de los instrumentos de medición, de los procedimientos para garantizar la calidad y validez de la información, y de los mecanismos que esas instituciones implementan para brindar apoyo y orientación a los países participantes. En segundo lugar, se espera obtener información objetiva sobre el logro académico de los estudiantes en relación a estándares de contenido internacionales, que impacte la opinión pública y que facilite una toma de decisiones mejor informada para el mejoramiento del sistema educativo. En cuanto a la vinculación de los proyectos comparativos con los sistemas nacionales de medición, se espera desarrollar capacidades técnicas que fortalezcan a estos últimos, pero no se ha diseñado generalmente un “cruce” sistemático entre los datos obtenidos en ambos tipos de mediciones. En tercer lugar, existe la expectativa, al menos por parte de algunos actores, de que la participación en una prueba internacional estimule la definición de estándares regionales contra los cuales medir el rendimiento de los países latinoamericanos. Si esto último no es posible, se espera al menos participar en pruebas internacionales de cobertura mundial que ofrezcan un marco conceptual sólido que permita interpretaciones sustantivas de los resultados. Cabe destacar que, si bien muchas de las decisiones de participación en pruebas internacionales en América Latina han estado inspiradas en la idea de establecer, o compararse con, estándares académicos más claros y exigentes, y si bien eso demuestra algún interés común por los aspectos curriculares de las reformas, la mayor parte de las decisiones se ha tomado sin la participación directa de los responsables de las unidades de currículo de los organismos de gestión educativa nacionales, ni de otros actores que puedan ejercer influencia sobre el desarrollo curricular en cada país. En este sentido, entonces, se podría afirmar que las iniciativas y decisiones de participación en pruebas internacionales están sólo parcialmente alineadas con los objetivos nacionales de reforma y logro curricular.

Si bien existen algunas opiniones “supranacionales” en la región que entienden la evaluación internacional dentro de un contexto de política educativa basada en la noción de *gerenciamiento*, en parte motivada por el condicionamiento de

los organismos internacionales, la mayor parte de las decisiones de participación no parece haberse tomado en función de ese modelo de gestión. Es decir, los beneficios que se esperan de la información recogida no están relacionados necesariamente a las nociones de descentralización y responsabilización, sino más bien al deseo de proveer mayores herramientas de gestión para un adecuado desarrollo e implementación curricular. Esto, pese a lo señalado líneas arriba respecto a la débil vinculación de las iniciativas internacionales con las unidades responsables del currículo.

En cuanto a la fase de **implementación** de las pruebas internacionales, encontramos grandes diferencias entre la experiencia de participación en la prueba del Laboratorio, por un lado, y en las pruebas de la IEA por el otro. Esas diferencias, como se explicó detalladamente en la sección anterior, están relacionadas principalmente a la capacidad técnica, organizacional, financiera y operativa de las agencias que administran las pruebas. Es claro que la IEA aventaja al Laboratorio en todas esas características, a pesar de que la prueba regional reportó también algunos beneficios a los países participantes. Con respecto a los aspectos estrictamente curriculares, vale decir que la representatividad de los países latinoamericanos en ambas experiencias es por lo menos cuestionable. La participación en pruebas de la IEA se ha dado en condiciones de desigualdad con respecto a los países desarrollados, pues América Latina ha ingresado tarde a los proyectos de evaluación, muchas veces con escasas capacidades técnicas para la representación, o con un débil aval financiero e institucional por parte de sus gobiernos nacionales. Por estas razones, las posibilidades de que los países de la Región influyan en las decisiones con respecto a los marcos conceptuales y curriculares han sido extremadamente reducidas, con algunas pocas excepciones.

En el caso de la prueba del Laboratorio, si bien la matriz curricular fue discutida con los países, la falta de un mecanismo de comunicación multidireccional<sup>vi</sup> no permitió lograr un consenso con respecto a un conjunto de estándares deseables ni a un marco conceptual común que permitiera una interpretación más sustantiva y contextualizada de los resultados, y por lo tanto la pertinencia curricular del proyecto se vio fuertemente debilitada. Recordemos que en las pruebas de la IEA en América Latina se valora precisamente que los países se midan con referentes o estándares externos a la Región, mientras que en la prueba del Laboratorio se esperaba llegar a acuerdos regionales sobre medidas de logro deseables.

Una última diferencia que vale la pena resaltar es que, mientras que las técnicas de muestreo y los instrumentos de las pruebas de la IEA son apreciados por su rigor y validez, la dirección técnica del Laboratorio enfrentó varios desafíos ante serios cuestionamientos técnicos por parte de los coordinadores nacionales. Por otro lado, la participación en pruebas internacionales en ambos casos ha favorecido la reflexión sobre los contenidos curriculares de cada país, en la medida en que las comparaciones revelan diferencias importantes entre países

en cuanto a los enfoques epistemológicos y a la distribución de contenidos a lo largo del currículo de educación básica. Sin embargo, esa reflexión se ha dado principalmente entre técnicos y expertos de las unidades de medición, pero con escasa o ninguna participación por parte de los profesionales de las áreas de desarrollo curricular. En términos de política curricular, esto implica que la medición internacional en cada país tiene escasa vinculación con los procesos de implementación de aquellos programas curriculares que, según se expresa, las evaluaciones desean mejorar.

Sin embargo, el aspecto más débil de los proyectos de evaluación, según se observa en la experiencia de la mayor parte de los países latinoamericanos, ha sido el de **interpretación y posterior diseminación de los resultados**. Además de las insuficiencias en el planeamiento estratégico, la escasa difusión de resultados estuvo marcadamente influida por la falta de comunicación y coordinación entre los equipos técnicos de medición y los de desarrollo curricular en cada país. Estos dos factores, falta de una estrategia de difusión y falta de trabajo conjunto entre actores clave, determinaron que finalmente existieran menos posibilidades de impacto en el mejoramiento de la implementación de reformas y en la reformulación de objetivos curriculares.

Durante la prueba del Laboratorio, la debilidad en la interpretación y diseminación se debió, principalmente, a la falta de una estrategia preconcebida, así como de un marco conceptual sólido y de estándares que permitieran una adecuada interpretación sustantiva de los resultados, a la vez que el segundo informe oficial, que se esperaba tuviera más impacto de política, llegó a los ministerios demasiado tarde, cuando las administraciones nacionales habían cambiado. Cuba fue, de acuerdo a las entrevistas, una excepción en este sentido, dado que allí los buenos resultados obtenidos fueron ampliamente difundidos desde la Presidencia, y favorecieron algunas decisiones sobre condiciones de trabajo docente, debido quizás en parte a la larga continuidad de la administración gubernamental.

En cuanto a los efectos de los resultados de pruebas de la IEA, es evidente que las desfavorables posiciones de América Latina en los *rankings* internacionales han causado un efecto de “*shock*” en algunas audiencias, especialmente por la forma de reporte que ha utilizado la prensa. En la actualidad, los países que han decidido continuar participando en pruebas de la IEA son concientes de la necesidad de “educar” a la prensa y otras audiencias influyentes, de volverlas más especializadas, de manera que puedan difundir informaciones más relevantes que la temida tabla de posiciones. Los reportes que redactan las oficinas de evaluación o currículo de los ministerios, por su parte, se consideran particularmente útiles cuando pueden tener impacto en el ámbito local, escolar y pedagógico, y cuando contribuyen a desarrollar capacidades de análisis e interpretación de datos. El caso de Estados Unidos durante la prueba TIMSS fue singular, como se describió en la sección anterior, por el alto involucramiento de

diversos grupos de interés y la consulta sistemática a actores del sector, así como también por la fuerte estrategia de difusión utilizada.

Con respecto al **impacto** de la participación en pruebas internacionales, en términos de decisiones sobre política educativa, de repercusiones políticas, o de investigaciones realizadas sobre los datos, nuevamente se hace evidente la diferencia entre la experiencia de América Latina y lo ocurrido en Estados Unidos. En nuestra región, tanto la evaluación del Laboratorio como las pruebas de la IEA han tenido un impacto político y de política relativamente bajo. Las excepciones, según nuestros datos, son Cuba, que tomó algunas decisiones concretas de políticas sobre la base de los resultados del Laboratorio, y Chile, que ha debido enfrentar el impacto sobre la opinión pública de los resultados en las pruebas de la IEA. En el caso del Laboratorio se han realizado algunas investigaciones posteriores, en los países y por parte de la OREALC misma, pero hasta el momento no han sido mayormente difundidas. En cuanto a la participación en pruebas de la IEA, la experiencia ha repercutido casi exclusivamente en el desarrollo técnico de las unidades de medición, lo cual no es un logro menor, pero no se reportan importantes medidas de política curricular, o educativa en general, en función de los resultados. Con respecto a PISA, una ventaja anticipada del proyecto que puede repercutir positivamente en el medio educativo es que las pruebas mismas, es decir los instrumentos mismos de medición, serán propiedad de los países participantes, y podrán utilizarse en aplicaciones nuevas e independientes para medir otras muestras y así favorecer investigaciones más profundas en cada ámbito nacional.

En Estados Unidos, donde existen más de 100 años de trayectoria en medición y evaluación educacional, la experiencia ha sido muy diferente, particularmente en relación a las pruebas TIMSS y TIMSS-R de la IEA. La creación de consorcios educativos, el involucramiento de organizaciones sindicales docentes y asociaciones profesionales de las diferentes disciplinas académicas, así como las múltiples líneas de investigación surgidas a partir de los resultados, confirman que el impacto de la participación en pruebas internacionales depende no solamente del potencial informativo de esas pruebas sino también de las estrategias de sensibilización previas a la difusión final de los resultados. Vale recordar que para cuando la IEA liberó los resultados de TIMSS, diferentes audiencias y medios de información ya estaban expectantes y deseosos de conocer los resultados. Estas diferencias entre las experiencias de Estados Unidos y América Latina confirman una vez más que una débil estrategia de sensibilización y difusión, así como la falta de un “contrato” entre funcionarios y usuarios en la búsqueda de información sistemática, deriva en un escaso impacto de las evaluaciones en las decisiones sobre políticas educativas, y sobre políticas curriculares en particular.

Además, no se monitorean los resultados sobre las pocas decisiones que se toman en función de las informaciones de estudios comparados, y por lo tanto es difícil saber si se ha avanzado o no en la dirección deseada. De esta manera, la

contribución de los resultados de pruebas internacionales a nuevos ciclos de política curricular, o de mejoramiento de los existentes, es mínima y en muchos casos inexistente. En otras palabras, esto implica que en el futuro podrán tomarse nuevas iniciativas y decisiones centrales de política curricular sin considerar la evidencia sobre los resultados de políticas y programas anteriores, que podrían indicar los resultados las pruebas internacionales.

### **III. RECOMENDACIONES PARA FUTURAS OPORTUNIDADES DE PARTICIPACIÓN DEL PERÚ Y OTROS PAÍSES LATINOAMERICANOS EN PRUEBAS INTERNACIONALES**

La evidencia recogida en este trabajo, si bien no puede considerarse representativa de la Región en su totalidad, ha permitido arribar a algunas conclusiones iniciales sobre la conveniencia de la participación de países latinoamericanos en futuras pruebas internacionales, así como sobre las condiciones en que esa participación debiera darse. Mientras que la sección anterior presenta resumidamente algunas conclusiones sobre los aspectos más relevantes de las experiencias pasadas de participación, y sobre la relación de esas experiencias con los marcos de reforma curricular de la Región, las conclusiones en este apartado se presentan directamente en forma de *recomendaciones* puntuales para participaciones futuras, y en curso, en pruebas internacionales de logro académico.

1. Dado que muchos de los países de la región ya han participado en alguna prueba internacional con anterioridad, se recomienda utilizar las informaciones de esas pruebas para generar opinión y sensibilizar a audiencias de interés antes de iniciar nuevos programas de reforma o mejoramiento curricular, y antes de adscribirse a un nuevo proyecto de evaluación comparada internacional. Una medida de ese tipo tendría el valor adicional de permitir el pilotaje de estrategias posibles de difusión de los resultados. Una forma viable de sensibilización es invitar a los representantes de la prensa y de organizaciones sindicales y profesionales de la educación a participar en las discusiones iniciales del proyecto, en la discusión sobre resultados anteriores y sobre la importancia de iniciar nuevas evaluaciones, y sobre el vínculo concreto que se percibe entre las evaluaciones y los contenidos curriculares. Es importante que los coordinadores en el ámbito nacional calculen con anticipación, y en colaboración con otras oficinas ministeriales, el costo material y el tiempo de trabajo que insumirá la fase de discusión y sensibilización, de manera tal que esta fase sea un aspecto formal y viable del proyecto de evaluación. Durante las discusiones, y mediante la informaciones que se hagan circular sobre resultados anteriores, se sugiere alertar sobre los posibles bajos resultados comparativos del país, pero simultáneamente explicar en forma ejemplificada las maneras en que la próxima prueba comparada puede informar sobre aspectos específicos de desarrollo o implementación curricular. Paralelamente, esta fase puede

comenzar a “educar” a las audiencias convocadas en algunos de los tecnicismos propios de la psicometría y de la sociometría, de manera tal que comiencen a desarrollarse algunas capacidades de interpretación y análisis de datos más allá de los *rankings*. Si el proyecto de evaluación está liderado por una institución académica o no gubernamental, es especialmente importante reunir a las autoridades ministeriales de medición y currículo, no sólo para obtener el aval oficial, sino para establecer una especie de “contrato” sobre las necesidades de información percibidas por el ministerio, y el tipo de informaciones que puedan derivarse de la evaluación comparada.

2. Al momento de decidir sobre el tipo de prueba internacional a aplicarse, cada país debiera considerar los siguientes **aspectos técnicos y contextuales**, que afectarán la mayor o menor utilidad de las experiencias de evaluación comparada:
  - a. El **momento histórico** en cuanto a la implantación de las reformas curriculares, la duración del gobierno en funciones, y el grado de afianzamiento del sistema nacional de medición de la calidad educativa.
  - b. Las **necesidades reales de la información** que se va a recoger, y las posibilidades concretas de aprovecharla, de acuerdo a los momentos de desarrollo e implementación curricular del país y de acuerdo a las dificultades percibidas o comprobadas anteriormente en esos procesos.
  - c. Las **características de las pruebas** en cuanto al marco conceptual utilizado, los dominios disciplinarios o cognitivos evaluados, y los factores escolares y socioeconómicos relevados en relación al rendimiento. Si la prueba es de cobertura regional y se presenta como un proyecto de elaboración conjunta entre los países participantes, es importante definir un **marco conceptual** que considere el contexto de política educativa y curricular de la Región y que permita una interpretación contextualizada de los resultados del estudio. Un estudio de esta naturaleza debiera representar también una oportunidad para definir consensuadamente, entre los técnicos de los países, un conjunto de **estándares** mínimos de contenido, y para realizar una revisión de los currículos nacionales y así poder comparar, mediante la evaluación, los alcances relativos de los currículos prescritos, implementados y logrados.
  - d. Las **oportunidades de participación efectiva** que el proyecto ofrece en relación a la construcción o adecuación de instrumentos de medición, o a la inclusión de “opciones nacionales” de evaluación, tales como ítemes especiales o submuestras de particular interés

nacional o regional. Ante esta posibilidad, debiera considerarse también cuál es la capacidad técnica de los coordinadores nacionales para hacer efectiva y viable esa participación.

- e. La **utilidad y pertinencia de la forma de reporte** de resultados utilizada por la agencia evaluadora.
3. Durante la etapa de implementación del proyecto, es decir durante el diseño de muestras, instrumentos y operativo de aplicación, se recomienda motivar y asegurar el vínculo y la **comunicación permanente** entre las unidades de medición y los equipos de desarrollo curricular en los niveles central y regional en cada país. Si partes de estos procesos son eminentemente técnicas y requieren poca intervención de personal de currículo, se recomienda por lo menos mantenerlos informados sistemáticamente sobre los avances del proyecto y sobre las futuras fechas en que serán convocados para discutir y analizar los resultados del estudio. Simultáneamente, debería mantenerse un tipo de comunicación igualmente sistemática con los representantes de la prensa y de las asociaciones sindicales y profesionales convocadas inicialmente (recomendación #1). El mismo tipo de mecanismo debería operar entre los coordinadores y los actores gubernamentales y no gubernamentales si el operativo de evaluación estuviera en manos de una institución académica independiente. Estos mecanismos de comunicación permanente son una forma de mantener vivo el “contrato” entre los técnicos y los consumidores de información que más pueden influir en la toma de decisiones para cambios de política y práctica curricular. Como las posibilidades son altas de que los usuarios gubernamentales (ministeriales) de la información cambien súbitamente debido a cambios políticos, se considera tanto más importante reforzar los canales de comunicación permanente con la prensa y con representantes de instituciones sindicales, académicas, y profesionales de la sociedad civil.
  4. Hacia el momento en que se reciben los datos ya procesados de la agencia evaluadora, y de acuerdo a las recomendaciones anteriores, los posibles usuarios de la información deberían estar informados de las fechas en que serán convocados a **analizar e interpretar** los datos de la prueba. Aquí se sugiere realizar primero una sesión de análisis y discusión entre los técnicos de evaluación y los representantes de desarrollo curricular, de manera tal de planificar en conjunto una agenda de discusión organizada para el momento de intercambio con representantes de la prensa, del mundo académico y de organizaciones no gubernamentales. Las discusiones con estos actores debiera conducir a una serie de recomendaciones puntuales sobre el tipo de reportes que sería deseable difundir, concretamente en relación al formato, al tipo de información que se debe incluir, a las audiencias prioritarias y secundarias que debieran alcanzar, al número sugerido de publicaciones, y a los diferentes objetivos que los reportes debieran perseguir, como por



ejemplo, si deben brindar insumos de información para medidas de política o para decisiones de tipo curricular y pedagógico en el nivel local y escolar. Según indica la evidencia hasta el presente, la sola discusión de la información por medio de reportes no es suficiente para que tenga un impacto real en las políticas y prácticas educativas. Es necesario organizar eventos locales con amplia participación de directores de escuela y docentes, en los cuales se explique cómo interpretar los datos y se discuta sobre diferentes maneras de mejorar el desarrollo curricular en las escuelas y su implementación en las aulas sobre la base de los datos analizados. Este es también el momento de **revisar el contexto de política curricular**, ver en qué momento de implantación de los programas curriculares se encuentra el país y sus regiones, cuáles son las dificultades que evidencian la pruebas nacionales, los informes de supervisión regionales, las investigaciones académicas y la prueba internacional misma para decidir qué tipo de mejoras (si no cuáles exactamente) se pueden recomendar a las autoridades nacionales y subnacionales de educación. En este sentido, el análisis de los datos debiera conducir a preguntas clave sobre la forma en que otros países logran mejores rendimientos estudiantiles, y sobre las condiciones bajo las cuales esos logros son posibles. Vale recordar que una de las ventajas de las pruebas internacionales es que pueden revelar información sobre los efectos diferenciales de algunas variables sobre el logro de aprendizajes que no es posible detectar en un estudio intra-nacional. También es importante recordar que las asociaciones entre variables dependientes e independientes que muestran los estudios internacionales deberían tomarse con cautela y ser analizadas dentro del contexto particular de cada país.

5. Se recomienda hacer un seguimiento del tipo de decisiones de política y práctica educativa, y curricular en particular, tomadas en función de los resultados difundidos. Una vez identificadas, se sugiere comisionar o alentar a instituciones académicas el estudio de los resultados que sobre el aprendizaje de los estudiantes han tenido, o no, esas decisiones y cambios de prácticas. Las informaciones recogidas mediante estos estudios deberían canalizarse y difundirse adecuadamente para ofrecer evidencia más concreta para la toma de decisiones sobre nuevas políticas y programas curriculares, así como para la decisión sobre nuevas evaluaciones internacionales y el tipo de estrategias que son más efectivas para la difusión y uso de sus resultados.
6. Una vez que se ha participado en una experiencia internacional de evaluación, y si la información obtenida de esa experiencia está a la altura de las expectativas de los coordinadores nacionales y de las autoridades políticas del sector, se aconseja continuar participando en nuevas mediciones, especialmente en las siguientes aplicaciones de la misma prueba. Esto puede permitir no solamente la medición del progreso académico de las poblaciones evaluadas, sino también que el público y los usuarios de la información se acostumbren a contar con la información

periódicamente, y de esta manera maximizar el uso técnico y político que pueda darse a esa información.

7. Se recomienda la participación de unidades subnacionales de gestión educativa, como por ejemplo estados, departamentos o grupos de escuelas asociadas. Si la participación no puede ser financiada por esas unidades, como probablemente será el caso en América Latina, la coordinación central de la evaluación (ministerios u otras agencias a cargo) debería considerar la posibilidad de conseguir recursos para promover esa participación. También se debería considerar la posibilidad de que la muestra nacional, en vez de ser conformada por alumnos de todo el país, fuera construida con casos de un departamento o estado que presente la suficiente heterogeneidad demográfica como para asegurar variabilidad en los resultados y, por ende, la posibilidad de analizar estadísticamente su relación con las variables predictivas del logro. ■
8. Esto permitiría, asumiendo que existe voluntad e interés por maximizar el impacto de la información, diseñar más fácilmente estrategias de difusión y uso de los datos en un contexto más reducido, para luego replicar esas estrategias en el ámbito nacional.

### **RECOMENDACIONES PARA FUTURAS INVESTIGACIONES SOBRE LA PARTICIPACIÓN DE AMÉRICA LATINA EN PRUEBAS INTERNACIONALES DE RENDIMIENTO ACADÉMICO**

Las sugerencias de investigación que se presentan a continuación guardan relación, principalmente, con aspectos del diseño que por limitaciones en los recursos materiales no pudieron controlarse totalmente en este trabajo y que por lo tanto restringen las posibilidades de hacer generalizaciones más amplias a la realidad de la Región.

En futuras investigaciones que profundicen o amplíen los hallazgos de este trabajo se recomienda, en primer lugar, expandir la muestra de países que hayan participado en pruebas internacionales, incluyendo países fuera del continente americano.

En segundo lugar, convendría definir más estrictamente los criterios para la selección de esos países, como por ejemplo que todos hayan participado en al menos una prueba regional y otra de cobertura mundial.

Luego, se recomienda expandir la muestra de actores entrevistados, incorporando especialmente representantes de medios académicos y actores involucrados en el diseño y entrega del servicio educativo en el ámbito local y escolar. Si estos representantes no han tenido la oportunidad de participar en muestras de pruebas internacionales, o si no han sabido de la existencia de

tales pruebas, las entrevistas deberían enfocarse en el uso potencial que esos actores consideran que podrían dar a las informaciones producidas por dichas pruebas.

Finalmente, se sugiere identificar más casos de usos concretos que se haya dado a los datos para la toma de decisiones de política educativa y curricular, en el nivel central, local y escolar. Adicionalmente, debería recogerse evidencia sobre el efecto o impacto que el uso de esa información ha tenido en los aprendizajes de las poblaciones evaluadas.

A pesar de las limitaciones de tipo metodológico impuestas por las condiciones materiales para este trabajo de investigación, y de las consecuentes amenazas a la validez de las conclusiones presentadas con relación a la totalidad de la Región, esperamos con este trabajo haber contribuido significativamente a la comprensión de las condiciones técnicas y políticas bajo las cuales los países de América Latina deberían participar en pruebas internacionales. Es nuestro deseo que los resultados de este informe de investigación alcancen y se difundan en los espacios de decisión sobre evaluación educativa en la Región, así como en el ámbito académico, y así coadyuvar a una mejor utilización de los recursos disponibles para la medición de rendimiento académico y, lo que es más importante aun, para su impacto en la calidad de los aprendizajes.

## **Bibliografía**

Atkin, Myron and Paul Black (1997) *Policy perils of international comparisons: The TIMSS case*. Phi Delta Kappan, 79,1, 22-28.

Ball, Stephen (1998) *Big policies/small world: An introduction to international perspectives in education policy*. Comparative Education, 34, 2, 119-130.

Beaton, Albert et al (1999) *The benefits and limitations of international educational achievement studies*. International Institute for Educational Planning; International Academy of Education. UNESCO: Paris, France.

Bempechat, Janine, Norma Jimenez & Beth Delamater (Borrador, 2000) *Cultural-cognitive issues in academic achievement: New directions for cross-national research*. Board on International Comparative Studies in Education (BICSE), National Research Council.

Bloom, Benjamin (1974) *Implications of the IEA studies for curriculum and instruction*. School Review, 82, 3, 413-435.

Burstein, L. (1989) *Conceptual considerations in instructionally sensitive assessment*. Paper presented in the American Educational Research Annual Meeting, San Francisco.

Burstein, L., Oakes, J. & Guiton, G. (1992) *Education indicators*, en M. Atkin (Ed.) *Encyclopedia of educational research* (6<sup>th</sup> edition). New York: Macmillan.

Carter, D.S.G & O'Neill, M.H (1995) *International perspectives on educational reform and policy implementation*. Brighton, Falmer.

Farrell, Joseph (1979) *The necessity of comparisons in the study of education: The salience of science and the problem of comparability*. Presidential address, annual meeting of CIES, Mexico City, 1978, reprinted from Comparative Education Review, 23, 1.

Forster, Margaret (2000) *A policy maker's guide to international achievement studies*. ACER Press, Australia.

Greaney, Vincent & Thomas Kellaghan (1996) *Monitoring the learning outcomes of education systems* The World Bank: Washington, DC.

Husén, Torsten (1994) *Educational research and policy-making*, in International Eyclopaedia of Education, Ed. Torsten Husen and Postlethwaite, N. Oxford: Pergamon Press.

Husén, Torsten (1987) *Policy impact of IEA research. Comparative Education Review*, 31, 1, 29-46.

Husén, T. & Tuijnman, A (1994) *Monitoring standards in education: Why and how it came about*, in Tuijnman, A. & Postlethwaite, T. (Eds) *Monitoring the standards of education*, Oxford: Pergamon Press.

Jones, P.W (1998) *Globalisation and internationalism: democratic prospects for world education. Comparative Education*, 34, 133-155.

Kellaghan, Thomas (1996) *IEA studies and educational policy. Assessment in Education*, 3, 2, 133-160.

Kohn, Melvin (1987) *Cross-national research as an analytic strategy* (American Sociological Association, 1987 Presidential Address). *American Sociological Review*, 52 (December: 713-731).

Levin, Benjamin (2001) *Conceptualizing the process of education reform from an international perspective. Education Policy Analysis Archives*, 9, 13. <http://epaa.asu.edu/epaa/v9n13.html>

Lyotard, J.F (1984) *The postmodern condition: A report on knowledge*. Manchester: Manchester University Press.

Macnab, Donald (2000). *Forces of change in mathematics education: the case of TIMSS. Educational Policy Analysis Archives* 8,15.

McKnight, Curtis (1987) *The underachieving curriculum: Assessing U.S. school mathematics from an international perspective. A National Report on the Second International Mathematics Study*.

Mislevy, Robert (1995) *What can we learn from international assessments? Educational Evaluation and Policy Analysis*, Winter 1995, 17, 4, 419-437.

National Commission on Excellence in Education (1983). *A nation at risk*. Washington, DC: U.S. Government Printing Office.

Platt, W.J (1975) *Policy making and international studies in educational evaluation*, en A.C. Purves & D.U. Levine (Eds.) *Educational policy and international assessment*. Berkeley, CA: McCutchen.

Schmidt, William and Curtis McKnight (1995) *Surveying educational opportunity in mathematics and science: An international perspective. Educational Evaluation and Policy Analysis*, Fall 1995, 17, 3, 337-353.

Smith, Adam (1776) *An inquiry into the nature and causes of the wealth of nations*. Dublin, Whitestone, 1776. 3 v. 22 cm. The Library of Congress, EEUU.

Theisen, Gary; Paul Achola and Francis Musa Boakari (1983) *The underachievement of cross-national studies of achievement*. *Comparative Education Review*, 27, 1, 46-68.

Torney-Purta, Judith (1990) *International comparative research in education: Its role in educational improvement in the US*. *Educational Researcher*, (volume, #?) 32-35

Torres, Rosa María (1999) *Reformadores y docentes: El cambio educativo atrapado entre dos lógicas* (Documento de trabajo para la discusión). IIPE Unesco: Buenos Aires.

Wirt, Frederick. *Comparing educational policies: Theories, units of analysis, and research strategies*. Reprinted from *Comparative Education Review*, vol.24, no.2, (June 1980).

Wolf, R (1979) *Sampling*. Bulletin 4: Secondary Study of Mathematics. Urbana, IL: Second International Mathematics Study.

## Anexo A: Cuadro resumen sobre pruebas internacionales principales.

<i>Examen</i>	<i>Agencia</i>	<i>Años de aplicación</i>	<i>Cobertura regional</i>	<i>Población objetivo</i>	<i>Áreas curriculares y competencias</i>	<i>Tipo de prueba</i>	<i>Principales factores asociados</i>	<i>No. de países latinoamericanos</i>
Civic Education Study	IEA		Global		Ciudadanía y democracia			
IALS- International Adult Literacy Survey	OECD Statistics Canada NCES ETS	1994-98	Global	Adultos (16-65 años)	Lectura		Variables culturales, sociales y demográficas	1/22
ALL- Adult Literacy and Lifeskills	Statistics Canada NCES OECD		Global	Adultos (16-65 años)	Lectura "Numeracy" Razonamiento lógico Trabajo en equipo		Variables sociales, educacionales, y económicas. Acceso y uso de NTIC.	
Primer Estudio Internacional Comparativo...	LLECE UNESCO-OREALC	1997	América Latina	3, 4	Lenguaje Matemática	Opción múltiple		13
PIRLS- Progress in International Reading Literacy Study	IEA	(RLS-1991)  2001+	Global	4 (9 años)	Lectura	Opción múltiple Respuesta abierta	Hogar Escuela Prácticas pedagógicas Currículum Políticas	2/?
PISA- Programme for International Student Assessment	OECD	2000 + PISA-Plus 2001?	Global	15 años	Destrezas "para la vida": Matemáticas Lectura Ciencias	Opción múltiple Respuesta abierta	Variables demográficas, sociales, económicas y educacionales	4/33
TIMSS- Third International Mathematics and Science Study	IEA	1995 +	Global	3/4, 7/8, último educación básica	Ciencias Matemáticas	Opción múltiple Respuesta abierta Demostración	Currículum Prácticas pedagógicas Políticas Actitudes Hogar Escuela	1/45 ?
TIMSS-R	IEA	1999	Global	3/4, 7/8, último educación básica	Ciencias Matemáticas		Currículum Prácticas pedagógicas Actitudes Hogar Escuela	1/38 +

## **Anexo B: Cuestionario de entrevistas.**

### **1. Iniciativa de participación en pruebas internacionales:**

1.1. ¿Quién propone / decide participar?

1.1.1 ¿Organizaciones internacionales-agencias evaluadoras?

1.1.2 Funcionarios. ¿De qué sectores de la administración pública, específicamente?

1.1.3 Grupos de interés / académicos.

1.1.3.1 ¿Otras? ¿Quiénes suscriben o apoyan estos emprendimientos? ¿Quiénes discrepan? ¿Por qué?

1.2 ¿Por qué se decide participar?

1.2.1 ¿Comparar con otros países? ¿Cómo se previó que esa comparación beneficiaría las políticas educativas (u otras) nacionales?

1.2.2 ¿Reforzar capacidades en el proceso?

1.2.2.1 ¿Cuáles capacidades: de medición, diseño, logística, análisis, diseminación y uso?

1.2.3 ¿Cuál era el contexto político y de reforma educativa en el momento de la decisión? ¿Había cambiado ese contexto al momento de divulgar y usar los resultados de las pruebas? Si hubo cambio de contexto, ¿Qué consecuencias trajo para el compromiso y la toma de decisiones de los funcionarios y técnicos involucrados?

1.2.4 ¿Qué relación se anticipó, al momento de decidir y *ex-post*, entre la prueba internacional y los sistemas nacionales y sub-nacionales de medición existentes?

### **2. Desarrollo y aplicación de pruebas:**

2.1 ¿Quiénes participan / participaron en las diferentes fases del proyecto (en cada prueba)?

2.1.1 ¿Académicos y técnicos expertos (del Ministerio de Educación o contratados)?

2.1.2 ¿Actores de la comunidad educativa con representatividad regional o local?

2.1.3 ¿Sociedad civil?

2.1.4 ¿De quiénes hablamos cuando nos referimos a los países participantes? ¿Sus representantes? ¿Garantizan sus representantes (coordinadores nacionales) que el país se vea bien



representado en las decisiones durante el diseño, o que se tomen decisiones acertadas sobre “opciones nacionales”?

- 2.2 ¿Cómo participan? ¿Cuánta participación efectiva tienen-han tenido los países menos desarrollados en la determinación de los objetivos curriculares / competencias a ser incluidos y en la construcción de los instrumentos de medición?
- 2.3 ¿Qué efectos ha tenido ese grado de participación sobre el diseño de las pruebas? ¿En el análisis de los resultados? ¿En la difusión posterior?
- 2.4 ¿Qué aprenden los técnicos/ académicos expertos en el proceso?
- 2.5 ¿Se ha decidido en algún caso la aplicación de ítemes de “opción nacional”? ¿Qué tipo de ítemes? ¿Con qué fin? ¿Se relacionaban con el tipo de ítemes incluidos en las pruebas nacionales?

### **3. Devolución de datos (resultados):**

- 3.1. ¿Cómo se preparan/ prepararon los funcionarios del Ministerio de Educación para la difusión de resultados? ¿Con qué capacidad cuentan / contaban para enfrentar posibles reacciones? ¿Déficit? ¿Cuál? ¿Cómo lo resolvieron? (entes externos, cuadros propios, descartaron información por razones técnicas, etc.)
- 3.2. ¿Quiénes reciben los resultados? ¿Coordinador nacional? ¿Qué garantías hay de difusión a funcionarios, actores de la comunidad educativa, prensa, etc.?
- 3.3. ¿Quedaron satisfechos con la calidad de las mediciones y comparaciones? ¿Se obtuvo el “producto” que se esperaba?
- 3.4. ¿Qué grado de influencia en las políticas educativas tienen actualmente los miembros de las organizaciones que participaron o lideraron la administración de pruebas en sus países? ¿Dónde están? ¿Qué hacen? ¿Cómo usan (ellos u otros) los resultados en la actualidad para influir políticas educativas?
- 3.5. ¿Qué tipo de información reciben (datos crudos, información procesada o interpretada, etc.)?
- 3.6. ¿Cómo se difunden? ¿A quiénes?
- 3.7. ¿Para qué se usan los resultados? ¿Cómo? Intención vs. realidad.

- 3.7.1. ¿*Ranking* "Carrera de caballos"? ¿Qué uso se le da a ese *ranking*?
- 3.7.2. ¿Mejorar los procesos pedagógicos/ capacitación y formación docente?
  - 3.7.2.1. Mediante información
  - 3.7.2.2. Mediante mecanismos de responsabilización (*accountability*)
- 3.7.3. ¿Mejorar el currículo intencional y/o implementado?
- 3.7.4. ¿Mejorar gestión institucional escolar?
  - 3.7.4.1. Mediante información
  - 3.7.4.2. Mediante mecanismos de responsabilización (*accountability*)
- 3.7.5. ¿Mejorar otros insumos y/o procesos de política y gestión (financiación, número de alumnos por aula, materiales didácticos, etc.)?
- 3.7.6. ¿Generar debate, llamar la atención de los medios masivos?
- 3.7.7. ¿Otros?
- 3.7.8. ¿Cuánto margen de control/ preparación del público existe para la lectura e interpretación de los resultados?

**Anexo C: Entrevistas.** Por razones de confidencialidad, solamente se reportan los nombres de las agencias o unidades ministeriales a la que pertenecen los 25 entrevistados de este estudio.

**Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación.** UNESCO-OREALC.

**Organización de Estados Iberoamericanos.** Programa Evaluación de la Calidad de la Educación.

### **Argentina**

- IDECE (Instituto para el Desarrollo de la Calidad Educativa). Ministerio de Educación.

- Ex-Subsecretaría de Evaluación de la Calidad Educativa. Ministerio de Educación.

### **Chile**

- Programa Formación Inicial Docente. División Educación Superior.

- Unidad de Curriculum y Evaluación/ SIMCE. Ministerio de Educación.

### **Colombia**

- ICFES (Instituto Colombiano para el Fomento de la Educación Superior)

- Universidad del Valle. Ciudad de Cali.

### **Cuba**

- Instituto Central de Ciencias Pedagógicas.

### **Ecuador**

- APRENDO. EB/PRODEC. Sistema semiautónomo oficial de medición de la calidad de la educación.

### **Estados Unidos**

- Universidad de Wisconsin-Madison.

- Universidad Estatal de Michigan/ Universidad Estatal de Nueva York, Albany.

- Federación Estadounidense de Profesores.  
Departamento de Educación. (American Federation of Teachers. Educational  
Issues Department.)

- Consorcio SMART (Science and Mathematics Required for Tomorrow).

### **Perú**

- Radio Programas del Perú.

- Unidad de Medición de la Calidad. Ministerio de Educación.

- UNESCO-Perú.

### **Uruguay**

- Unidad de Medición de Resultados Educativos (UMRE) de la Administración  
Nacional de Educación Pública.

## **Anexo D: Resumen descriptivo de pruebas internacionales de LLECE, IEA y OCDE.**

El **Primer Estudio Internacional Comparativo sobre Lenguaje y Matemáticas y Factores Asociados** del Laboratorio Latinoamericano para la Evaluación de la Calidad Educativa (UNESCO-OREALC) fue aplicado en 1997 en 13 países latinoamericanos, en los grados 3ro y 4to de educación primaria, en las áreas de lenguaje y matemáticas. La prueba incluyó ítemes de opción múltiple y también de respuesta abierta, y recogió información sobre variables asociadas al logro académico, entre ellas de contexto socioeconómico, familiar, escolar y áulico. Los resultados confirmaron muchos de los datos que algunos sistemas nacionales ya habían comenzado a producir, como por ejemplo las diferencias de logro a favor de las escuelas urbanas y privadas, escasas diferencias de rendimiento entre géneros, y una fuerte asociación positiva entre bajo nivel socioeconómico y bajo rendimiento académico. En cuanto a los resultados comparados internacionales, fue particularmente llamativa la diferencia entre Cuba y el resto de los países de la Región, con más de dos desviaciones estándar a favor de Cuba. Los resultados fueron publicados por LLECE en dos informes. Un primer informe presentó básicamente los datos comparativos de logro entre países. El segundo informe presentó los factores asociados al rendimiento, pero con datos agregados al total de la Región y no individualmente por país.

La prueba **TIMSS** administrada por la IEA fue aplicada en 1995 en unos 45 países de todo el mundo. En algunos países sólo participaron algunas provincias y no todos los países evaluaron todas las poblaciones objetivo del estudio. Esas poblaciones fueron los grados 3ro ó 4to de primaria, 7mo u 8vo grados, y último año de la educación básica (previo a la educación superior). Las pruebas se centraron en las áreas de matemáticas y ciencias y la matriz curricular se definió mediante un relevamiento previo de los contenidos curriculares prescritos tanto por documentos oficiales como por libros de texto y materiales didácticos de uso masivo en un número significativo de países. La prueba incluyó ítemes de opción múltiple, de respuesta abierta y de demostraciones prácticas (e.g. de laboratorio). También se realizó un estudio cualitativo, mediante videograbaciones, sobre procesos pedagógicos de aula en seis países de la muestra. Las principales variables relacionadas al logro académico que se estudiaron fueron el currículo prescrito e implementado, prácticas pedagógicas, políticas educativas, actitudes, y contexto familiar y escolar. Uno de los datos más interesantes obtenidos del estudio se refiere a las grandes diferencias en las oportunidades de aprendizaje que tienen los alumnos de diferentes países, dependiendo de la claridad y la profundidad con que se explicitan los contenidos curriculares en documentos oficiales y libros de texto. El estudio también reveló que en muchos países los docentes de ciencias y matemáticas desconfían seriamente de sus conocimientos conceptuales en estas dos áreas. En cuanto a los resultados comparados internacionales, quedó confirmado que los países del sudeste asiático, como Japón, Singapur y Corea, logran resultados muy

superiores a la mayoría de los países industrializados del mundo. El único país latinoamericano que participó y permitió la publicación de los resultados, Colombia, obtuvo uno de los últimos lugares en el *ranking* internacional. En 1998 se aplicó **TIMSS-R**, es decir una repetición de la prueba en unos 38 países en todo el mundo, incluyendo tres países latinoamericanos. Se evaluaron las mismas áreas, para las mismas poblaciones objetivo, y se observó una alta participación de sistemas educativos subnacionales, especialmente en Estados Unidos, como por ejemplo estados independientes y consorcios educativos. TIMSS será aplicado nuevamente en el año 2003. La prueba de **Educación Cívica**, también administrada por la IEA, se aplica regularmente en países de todo el mundo. Los contenidos de la prueba se centran en temas de ciudadanía, gobierno y democracia. También se recogen datos sobre factores sociales, familiares y actitudinales asociados al rendimiento académico.

**PISA**, la prueba comparativa de la OCDE, se aplica regularmente en países de todo el mundo a estudiantes de 15 años de edad, cualquiera sea su grado de escolarización. Los países miembro de la OCDE ingresan a la prueba automáticamente, mientras que los demás países son invitados a sumarse al proyecto mediante el pago de una cuota de membresía. Las pruebas de PISA no se diseñan a partir de una matriz curricular consensuada entre países sino a partir de un conjunto de estándares de contenido y desempeño, elaborados por un grupo de especialistas en diferentes áreas de conocimiento, que se consideran imprescindibles para la adecuada inserción de los jóvenes en la sociedad y en el mundo laboral de hoy. Los resultados de PISA (y PISA Plus, una aplicación posterior para los países que no pudieron sumarse a tiempo al proyecto original) aún no han sido publicados por la OCDE. PISA volverá a aplicarse en el año 2003.

## **Notas Marginales**

Traducción de los autores, al igual que todas las citas de este trabajo que han sido extraídas de textos originales en inglés.

<sup>ii</sup> Kellaghan (1996) plantea un conjunto de ocho condiciones principales que deben darse para obtener el mayor beneficio posible de la participación en pruebas internacionales:

1. Debiera asegurarse que los datos obtenidos en los estudios internacionales representan fielmente el logro académico de los estudiantes en cada país participante. Para lograr esto, es menester *a)* que exista consenso sobre qué resultados educativos son deseables y merecen ser medidos, *b)* que esos resultados puedan ser expresados con precisión y detalle suficientes como para permitir su medición, *c)* que exista la tecnología necesaria para realizar las mediciones, y *d)* que los alumnos evaluados sean representativos del sistema educativo de su país, o del grupo especificado (Ej. cuarto y octavo grados; último año de secundaria). Los datos derivados de las pruebas internacionales debieran permitir comparaciones válidas entre países. En este sentido, es particularmente crítico construir pruebas que reflejen los contenidos principales de los currículos de todos los países participantes. Si los currículos no están apropiadamente representados, la interpretación de logro comparado puede perder validez.
2. Se debe guardar especial cuidado al seleccionar los contenidos de las pruebas cuando se intenta medir el capital humano de las naciones. Existen fuertes argumentos que cuestionan la posibilidad de establecer juicios sobre el capital humano de un país solamente sobre la base de una prueba que mide un número limitado de capacidades y que generalmente evalúa grupos específicos y seleccionados de la población estudiantil.
3. También se requiere cautela al momento de establecer relaciones, especialmente relaciones causales, entre el capital humano de una nación y su desempeño económico. Se asume que, así como el nivel de educación de una persona suele estar incuestionablemente asociado a sus ingresos, de la misma manera el rendimiento académico de los alumnos determina el capital humano de un país y en consecuencia sus posibilidades de crecimiento económico. Existe un conjunto de razones por las cuales esa relación es difícilmente sustentable.
4. Se debe establecer con claridad, y con anticipación, los propósitos de la prueba para poder decidir qué tipo de datos se van a recoger. Difícilmente una misma prueba puede servir múltiples propósitos simultáneos, como por ejemplo emitir juicios sobre la calidad del currículo, establecer un mecanismo de responsabilización, y justificar la necesidad de una reforma estructural. Algunos de esos propósitos pueden ser simplemente incompatibles con un diseño de prueba en particular.

5. La información derivada de las pruebas debiera servir los propósitos y necesidades de los funcionarios del sistema educativo. Por ejemplo, la elección del tipo y cantidad de indicadores que se van a seleccionar debiera depender del tipo de información que los funcionarios necesitan. También es importante que se establezcan mecanismos efectivos de comunicación entre investigadores y funcionarios para aumentar las posibilidades de que la información recogida sea realmente utilizada.
6. Se necesita información contextual (social, política, económica, estructural) sobre cada país que permita establecer un marco dentro del cual interpretar los resultados de las pruebas. La experiencia de estudios comparados anteriores debiera servir a los investigadores para determinar qué variables suelen tener bajo valor predictivo del rendimiento académico y cuáles otras variables contextuales podrían ser de mayor utilidad.
7. Las acciones que los gobiernos tomen sobre la base de la información recabada por los estudios internacionales deberían ser evaluadas para conocer su impacto sobre la calidad de la educación en esos países. Conocer el impacto de esas políticas sería de suma importancia para justificar la participación de los países en nuevas pruebas internacionales.
8. La información que se obtiene de los estudios debería ser lo suficientemente valiosa para los gobiernos, de manera tal que quede justificado el gasto realizado en su participación. Por ejemplo, sería importante considerar qué otras oportunidades de investigación se están sacrificando al dedicar esfuerzos y recursos a la participación en estudios internacionales comparados.

Otros autores advierten, también a modo de condición, sobre la necesidad de diseñar las pruebas internacionales de manera tal que los datos recogidos permitan el análisis intra-nacional (Wolf, 1979, en Mislevy, 1995), y se enfatiza la necesidad de contar con datos desagregados para poder llegar a conclusiones útiles para el mejoramiento del currículo y de los procesos pedagógicos (Burstein, 1989, en Schmidt y McKnight, 1995). Finalmente, se argumenta que es absolutamente necesario, al momento de diseñar las pruebas, que se tengan en cuenta los contextos culturales de los estudiantes, de manera de saber qué significados y percepciones aportan los diferentes actores al proceso educativo, qué significa exactamente ser una persona instruida en diferentes países, y así poder comprender mejor la relevancia de los resultados de acuerdo a cada contexto particular (Bempechat et al, 2000).

<sup>iii</sup> Bowe, Ball y Gold (1992), en Levin (2001), por ejemplo, proponen un modelo tripartito para explicar los procesos de reforma educacional. En lo que los autores llaman una “trayectoria de política”, existen tres fases distintas y relativamente autónomas de desarrollo de reformas: *influencia*, *producción de texto* y *práctica*. En la primera, la idea de una nueva medida de política es originalmente concebida e



importa principalmente quién define su naturaleza. La segunda fase consiste en la concreción de la medida de política como producto (texto). Aquí puede presentarse una primera desviación entre las ideas originales y lo que finalmente queda registrado como norma. Luego, en la etapa de práctica, la medida de política se concreta en el campo educativo mismo. Nuevamente, pueden existir desviaciones entre el texto de la norma y lo que los profesionales realizan en su campo de acción. Este esquema guarda semejanzas con el modelo lineal propuesto por Levin (2001), quien distingue cuatro etapas principales en el proceso de formulación de políticas: *origen, adopción, implementación y resultados*. La primera se refiere a los actores que conciben la idea de cambio, el rol que juegan en su formulación, y las condiciones contextuales e ideológicas que influyen en la iniciativa. En la fase de adopción se suscitan las primeras desviaciones entre la idea original y la norma que finalmente registrará, generalmente debido a cambios de idea por parte de los actores, o a insuficiencias burocráticas o administrativas. Durante la implementación surgen los problemas de concreción de la medida de política en el terreno práctico, tales como oposición profesional, falta de comunicación o escasos recursos materiales. Finalmente, la etapa de resultados hace referencia a la evidencia disponible sobre los efectos de las reformas. Esta fase suele ser compleja debido a los diferentes tipos de resultados que se esperan de las reformas y del sistema, y porque a menudo esos resultados son difíciles de evaluar. A la fase de evaluación de resultados debiera agregarse un último elemento, el de retroalimentación, que permite, en teoría, promover la iniciativa de cambio para formular nuevas políticas que solucionen los problemas identificados por la evaluación. El proceso, idealmente, sería cíclico y permanente. Según Wirt (1980), sin embargo, ese ciclo puede interrumpirse en cualquiera de las fases, dependiendo del balance de fuerzas políticas que entran en juego.

<sup>iv</sup> Dr. Gilbert A. Valverde, Ph.D., State University of New York at Albany, EEUU.

<sup>v</sup> Un consorcio educativo es una asociación voluntaria de distritos escolares, que a su vez están compuestos por varias escuelas bajo la misma administración, con el fin de unir criterios y recursos para el mejoramiento de la calidad educativa de las escuelas participantes.

<sup>vi</sup> Se argumenta que el flujo de comunicación electrónica se daba principalmente entre el Laboratorio y cada país individualmente, pero no entre los países mismos.