Evaluaciones educativas realizadas por ciudadanos en México: validación de la Medición Independiente de Aprendizajes

Felipe J. Hevia
Centro de Investigaciones y Estudios Superiores en Antropología Social (CIESAS), Unidad Golfo
Samana Vergara-Lope Tristán
Universidad Veracruzana, Facultad de Psicología

Resumen

En la última década se han desarrollado evaluaciones ciudadanas en diversos países de Asia y África. Estas evaluaciones son innovadoras, puesto que entregan insumos de información complementaria a las mediciones gubernamentales y fomentan la participación social en la educación. Las evaluaciones ciudadanas se basan en el desarrollo de instrumentos simples, pero robustos, que permitan medir aprendizajes básicos. El objetivo del artículo es validar el instrumento Medición Independiente de Aprendizajes (MIA), que consta de 10 reactivos y busca medir los aprendizajes básicos (operaciones matemáticas y lectura) de niños y adolescentes de entre 5 y 16 años de edad. El instrumento se puso a prueba para escalonar y descartar reactivos y obtener confiabilidad, además de validar tres versiones paralelas. Este instrumento obtuvo un índice de consistencia interna adecuado y un alto coeficiente de equivalencia en las pruebas paralelas. Se concluye que la MIA permite medir aprendizajes básicos en México de manera válida y confiable.

Palabras clave

Confiabilidad, evaluación del aprendizaje, instrumentos de medición, participación social, validez de pruebas.

Educational evaluations carried out by citizens in Mexico: Validation of the Independent Assessment of Learning instrument

Abstract

In the last decade citizen-led evaluations have been developed in different countries in Asia and Africa. These evaluations are innovative, given that they provide sources of information that complement governmental assessments and promote social participation in education. The citizen-led evaluations are based on the development of simple but robust instruments that allow measuring basic skills. The objective of this article is to validate the Independent Assessment of Learning instrument (MIA) that is comprised of 10 test items and aims to measure the basic skills (mathematical operations and reading) of children and adolescents between 5 and 16 years of age. The instrument was tested to determine to scale, eliminate items and determine reliability, as well as to validate three parallel versions. The instrument acquired an index of sufficient internal consistency and

Keywords

Evaluation instruments, learning evaluation, reliability, social participation, test validity.

Recibido: 12/10/2015
Aceptado: 17/02/2016

a high coefficient of equivalence in the parallel tests. The conclusion is that the MIA permits a valid and reliable measurement of basic skills in Mexico.

Introducción

La importancia de las evaluaciones de aprendizajes en el mundo

Existe un sólido consenso internacional sobre la necesidad de enfocar las políticas educativas en los aprendizajes. Esto implica, primeramente, lograr la universalización del acceso a la educación básica y, segundo, mejorar la calidad de los sistemas educativos (Unesco, 1990, 2000). Alcanzar la enseñanza primaria universal fue un objetivo de desarrollo del milenio; garantizar una educación de calidad inclusiva y equitativa y promover las oportunidades de aprendizaje permanente para todos es un objetivo de desarrollo sostenible aprobado en 2015 por Naciones Unidas. Con ello se busca asegurar que la asistencia a la escuela tenga impactos positivos y permanentes en los aprendizajes (Unesco, 2012).

Una de las consecuencias de poner los aprendizajes al centro de los sistemas educativos fue la necesidad de desarrollar métodos de evaluación y medición del logro educativo (Anderson, 2005). Estos métodos se nutrieron de la construcción de diversas corrientes de evaluación desarrolladas a lo largo del siglo XX que se centraron en la creación de exámenes. Posteriormente, se dedicaron a crear pruebas de alternativas múltiples (Escudero Escorza, 2003) y a desarrollar experiencias nacionales e internacionales de medición masiva, sobre todo a partir de la década de 1990 con la aplicación de los Trends in International Mathematics and Science Study (TIMSS) y del Programa para la Evaluación Internacional de los Estudiantes (PISA, por sus siglas en inglés), a cargo este último de la Organización para la Cooperación y el Desarrollo Económico (OCDE).

Muchas de estas experiencias fueron parte de reformas orientadas a incrementar la responsabilidad de los sistemas educativos en un proceso definido como *accountability* educativa. Según Kane y Staiger (2002), este proceso incluyó tres elementos: pruebas a los estudiantes, información pública sobre el desempeño escolar y recompensas o sanciones sobre la base de alguna medida de mejora o rendimiento en la escuela. Así, se llevaron a cabo enormes reformas en los sistemas educativos para tener información sobre los aprendizajes de los niños más allá de los indicadores tradicionales de abandono y repetición y, luego, compartir esa información con la sociedad para generar consecuencias en los sistemas educativos (Corvalán y McMeekin, 2006). La idea central

detrás del desarrollo de pruebas dirigidas a estudiantes es medir los aprendizajes sobre la base de estándares que definan con claridad lo que se espera de la educación y procurar que los currículos funcionen como hojas de ruta para lograr esos aprendizajes (Corvalán y McMeekin, 2006).

Los métodos preferidos –y recomendados– para identificar los aprendizajes fueron las pruebas estandarizadas de respuesta múltiple: "Para estandarizarse de esa manera, las pruebas se basan en respuestas a un conjunto cerrado de respuestas alternativas con un número fijo de 'distractores' y una sola respuesta 'correcta'" (McMeekin, 2006, p. 33). Estas pruebas estandarizadas también se utilizan para la selección de aspirantes a la educación superior y se han transformado en pruebas hegemónicas para la medición de aprendizajes y la evaluación educativa (Aboites, 2012), incluida la iniciativa presentada en la reforma de la Secretaria de Educación Pública (SEP) del 2010, que menciona que dentro de los Consejos Escolares de Participación Social se deben revisar los resultados de las pruebas de logro (SEP, 2010).

Las evaluaciones basadas en pruebas estandarizadas han representado importantes avances en el conocimiento de los aprendizajes (Hanushek y Raymond, 2004), pero las críticas a estas mediciones, tanto a su metodología (Kane y Staiger, 2002; Kreiner y Christensen, 2014) como a su uso general para mejorar la educación, son cada vez más comunes (Aboites, 2012; Ravitch, 2011; Sánchez, 2014).

Evaluaciones de aprendizaje en México

México no ha estado ausente de estos procesos mundiales para medir aprendizajes. En las décadas de 1970 y 1980 la Secretaría de Educación Pública se esforzó por sistematizar la información censal. En la década de 1990 se crearon diversos programas de medición de aprendizajes con múltiples propósitos, incluida la evaluación de docentes, producto del programa Carrera Magisterial implementado en 1993 (Santibáñez, Martínez, Datar, McEwan, Messan-Setodji, y Basurto-Dávila, 2007). También en esa década México participó en mediciones internacionales que facilitaron la creación de un Sistema Nacional de Evaluación, que se consolidó en los años 2000 (Zorilla, 2003). En 2002, se creó el Instituto Nacional para la Evaluación de la Educación (INEE), que ha pasado por una serie de reformas administrativas, la última en 2013, cuando se convirtió en un organismo constitucional autónomo (INEE, 2014).

Fue en el seno del INEE donde se desarrollaron los Exámenes de Calidad y el Logro Educativo (Excale), que se aplican desde 2003 a una muestra representativa de escuelas primarias y secundarias sobre una base muy seria de procedimientos técnicos para

lograr confiabilidad y validez en la construcción de sus indicadores (Backhoff, Sánchez, Peón, Monroy, y Tanamachi, 2006). Posteriormente, en 2006, la Secretaría de Educación Pública desarrolló la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE) (Campos y Urbin, 2011). A diferencia de los Excale, la ENLACE tenía un diseño censal, se aplicaba en la totalidad de las escuelas en México, y en los últimos años se posicionó como un parámetro aceptado para medir la calidad de la educación. Sin embargo, desde sus orígenes recibió fuertes críticas (Avilés, 2012; Bautista, 2012) y hubo constantes denuncias de corrupción en torno a su aplicación, al punto de que para el año escolar 2014 la SEP la suspendió (Backhoff y Contreras Roldán, 2014). Según las palabras del entonces secretario de Educación: "Con una prueba así nadie puede decir que es una prueba que refleja con transparencia el rendimiento escolar. Quien así lo diga parece que no ha leído los resultados de Enlace en 2013" (Martínez, 2014, s/p).

En efecto, una serie de errores de diseño y, en especial, el vínculo explícito que se propuso entre los resultados de la prueba y los programas de estímulos docentes desencadenaron la suspensión de la prueba de medición (Backhoff y Contreras Roldán, 2014). La prueba ENLACE, además, influyó para que la discusión sobre la calidad educativa se centrara exclusivamente en el docente, situación que se agravó con las discusiones sobre la denominada Reforma Educativa que implementó el Gobierno de Enrique Peña desde el primer día de su mandato (Hevia, 2014).

Participación social y evaluaciones educativas

Otra de las limitaciones de los sistemas de evaluación estandarizada que se vienen aplicando en las escuelas tiene que ver con la dificultad para comunicar sus resultados de manera accesible a la población general, de modo que puedan servir para fomentar la participación y el control social en la educación (Martínez, Bracho y Martínez, 2010). Por ello, se realizaron esfuerzos por acercar la información de estas evaluaciones a los padres de familia, en especial las reformas a los Consejos Escolares de Participación Social, para que pudieran revisar los resultados de las pruebas de logro (SEP, 2010). Sin embargo, los resultados muestran que la gran mayoría de consejos escolares no conoce ni usa los resultados de estas evaluaciones para incrementar el control social en sus planteles (SEP, 2015).

Así, a pesar de la gran cantidad de pruebas estandarizadas aplicadas en México, la participación social siguió limitada a la simple cooperación económica de los padres para la gestión educativa y restringida por la legislación y las prácticas institucionales (Latapí, 2004; Zurita, 2013), para las que la inclusión de los

ciudadanos en la evaluación del logro educativo no parece estar en el horizonte de las evaluaciones tradicionales en las escuelas.

La innovación educativa desde la ciudadanía: evaluaciones ciudadanas o citizen-led assessments

Frente a esta tendencia hegemónica enfocada a la medición educativa por medio de pruebas estandarizadas –ligadas a una serie de incentivos positivos y negativos para los maestros y escuelas, y que son ajenas a la ciudadanía y no fomentan la participación social– hace 10 años surgió en la India una alternativa desde la sociedad civil. Ante la ausencia de sistemas nacionales que permitieran medir el aprendizaje en los grados iniciales de la educación básica, en 2005, la organización Pratham desarrolló un modelo de valoración denominado Annual Status of Education Report (ASER), que mide las capacidades lectoras y matemáticas básicas de niños de entre 5 y 15 años de edad. Esto lo hace por medio de instrumentos de fácil comprensión aplicados por ciudadanos en los hogares (Banerji, Bhattacharjea y Wadhwa, 2013). Este modelo se expandió, en 2008, a Pakistán, mediante la iniciativa ASER-Pakistán; en 2009, a Kenia, Tanzania, Uganda y Mali, por medio de la iniciativa Beekunko; y en 2011, a Senegal, gracias a la iniciativa Jagandoo. De esta manera, se llegaron a medir los aprendizajes de más de un millón de niños en el sur de Asia y África subsahariana (Levine, 2015; Save the Children, 2013). Con un riguroso método de selección, este modelo generó muestras representativas de niños a escala nacional y subnacional. Las herramientas se diseñaron para que los padres, maestros, comunidades y cualquier persona pudieran aplicarlas y comprender, también, los resultados. Los indicadores se centraron en la habilidad para leer textos simples y resolver operaciones matemáticas básicas (Banerji, 2014). Estas experiencias llamaron la atención internacional (Rosenberg, 2014; Zaidi, 2014) y la información que se generó fue utilizada por organismos y expertos en investigación educativa (Dreze y Sen, 2013; Unesco, 2014).

Si bien el interés del ASER-Pratham se centró en la posibilidad de generar información independiente sobre el logro educativo en la India rural, este modelo también funcionó como una práctica innovadora para incrementar la participación y el control social en el ámbito educativo (Levine, 2015). En la opinión de los autores del presente artículo, este modelo tiene cinco elementos principales que podrían aplicarse en México, no sólo para generar una evaluación del logro académico independiente, sino también para incrementar la participación social en la educación.

El primero de los cinco elementos es la simpleza de los objetivos: estas evaluaciones se centran en los aprendizajes básicos (lectura y operaciones matemáticas elementales). Pretenden saber

si los niños –vayan o no a la escuela– los poseen, pues son elementos que ellos mismos y sus padres pueden reconocer y comprender en el momento en que se les aplican los instrumentos de medición. La sencillez también radica en que los resultados se pueden comunicar localmente de manera más simple y ser más útiles que los resultados tradicionales de las evaluaciones existentes, que suelen ser más difíciles de transmitir. Los instrumentos también se aplican en múltiples lenguas, lo cual permite medir estos aprendizajes en idiomas nacionales y regionales. Así, por ejemplo, en India se aplican en 19 lenguas, incluidas el inglés, el hindi y el talego; en Uganda, Tanzania y Kenia se aplican en inglés y suajili; mientras que en Senegal se aplican en árabe y en francés (PAL Network, 2015).

En segundo elemento es la sencillez de los instrumentos: son sencillos y rápidos de aplicar. Los instrumentos se desarrollan en cada país y se aplican de manera individual a cada niño (uno por uno). La prueba de lectura permite saber si los niños pueden leer letras, palabras, enunciados o historias simples; y, la prueba matemática, si pueden identificar números del 10 al 99 y realizar operaciones comunes, como sumar, restar y dividir. La simpleza de los instrumentos permite que cualquier ciudadano que sepa leer pueda aplicarlos, previa capacitación para ello. También permite retroalimentar directamente a los padres y a los propios niños que responden el instrumento, así como comunicar los resultados con mayor facilidad. Por ejemplo, la iniciativa Uwezo entrega información de todos los distritos en Kenia, Uganda y Tanzania en conferencias de prensa nacional, pero también entrega los resultados en reuniones con agentes interesados en cada uno de los distritos rurales y urbanos.

El tercer elemento son los resultados, pues posibilita que los padres, comunidades y escuelas puedan pensar en acciones de intervención, pues después de difundir los resultados, tanto en la localidad como en el país, surge la pregunta: "¿y ahora qué hacemos?" En este sentido, saber cuántos niños de una localidad o de una colonia pueden leer palabras, pero no enunciados, permite ubicar los niveles de aprendizaje y, con ello, diseñar intervenciones específicas para cada nivel. En India, por dar un ejemplo, se desarrollaron intervenciones específicas en los estados de Maharashtra y Bihar, donde se implementaron campamentos de fortalecimiento de lectura y matemáticas en miles de escuelas, con base en los resultados del ASER, con lo cual se pudo ubicar a los niños en niveles de aprendizaje y medir sus avances (Banerji, 2014).

El cuarto elemento es que la aplicación y difusión de los resultados la realizan voluntarios locales. En los países donde se ha aplicado este modelo se involucran miles de ciudadanos interesados en hacer algo por la educación en sus localidades. El reclutamiento de voluntarios implica un trabajo muy importante de construcción de alianzas con organizaciones civiles y de

identificación de voluntarios. Éstos son capacitados y se encargan de levantar la información en los hogares. Gracias a la simpleza de los instrumentos, quienes los aplican pueden tener un panorama bastante claro de los déficit de lectura y operaciones matemáticas que presentan los niños de sus propias comunidades, lo cual puede incrementar su motivación para que participen y ayuden a cambiar las cosas. Gracias a los procesos de reclutamiento y a la entrega de información, en la práctica se generan redes de organizaciones –algunas de ellas previamente existentes e interesadas en temáticas educativas– que generan información empírica de primera mano sobre la situación educativa en sus territorios. Así, por ejemplo, ASER-Pakistán trabaja con 144 organizaciones civiles y más de 3 000 voluntarios locales; mientras que, en India, ASER tiene 577 organizaciones socias (PAL Network, 2015).

Debido a que el modelo se aplica en los hogares y no en las escuelas, el quinto y último elemento permite ampliar la mirada de la educación más allá de los contextos escolares y focalizar esfuerzos en el involucramiento parental. En el caso mexicano, esto es central para complementar la discusión generada por la reforma educativa, muy enfocada en las políticas docentes y la evaluación del desempeño de los maestros.

Por tanto, en este contexto resulta pertinente y relevante adaptar y replicar la metodología ASER para el caso mexicano, esto es, que ciudadanos, organizaciones y academia midan, de manera independiente y en los hogares, los aprendizajes básicos de lectura y operaciones aritméticas de niños y jóvenes de entre 5 y 16 años de edad. Este proyecto comenzó a desarrollarse en México en 2014 bajo el nombre "Medición Independiente de Aprendizajes-MIA" (MIA, 2014). El primer paso fue validar los instrumentos de medición para la población mexicana.

En este sentido, el objetivo del presente artículo es mostrar los procedimientos que se realizaron para validar el instrumento Medición Independiente de Aprendizajes (MIA, Anexo 1), con el fin de que fuera simple, confiable y permitiera medir la capacidad lectora y de realizar operaciones matemáticas básicas de la población mexicana. Además, se presentan los procesos seguidos para obtener tres formas paralelas y equivalentes del cuestionario, de modo que se pudiera utilizar una distinta para cada niño, en caso de existir más de un menor por hogar.

Método

La adaptación de la metodología ASER a México implicó determinar qué temas tenían que ser incluidos en el instrumento para evaluar aprendizajes básicos de niños y jóvenes de entre 5 y 16 años de edad, con base en la parte del currículo mexicano relacionado con estos aprendizajes. También implicó crear los reactivos

correspondientes a cada indicador, pero siempre con la lógica de crear un instrumento sencillo, rápido y fácil de aplicar que pudiera administrase en hogares y por voluntarios de muy diversas características.

En este procedimiento participaron grupos de expertos que intervinieron en las diferentes fases de la investigación. Contribuyeron en decidir las dimensiones del instrumento MIA, así como en crear y realizar los cambios necesarios a los ítems de las tres versiones del mismo. La validez del contenido por medio del juicio de expertos se obtuvo a partir de dos fases.

En la primera, se formaron tres grupos de trabajo, dos de ellos con seis expertos cada uno y, el restante, con cuatro. Los grupos de expertos estuvieron conformados por maestros frente a grupo de 1º y 2º de primaria de escuelas urbanas y rurales en los municipios veracruzanos de Xalapa y La Antigua; maestros supervisores escolares; y un maestro académico de una escuela normal. En los grupos de trabajo se les presentaron, primero, los instrumentos que se aplican en las evaluaciones ciudadanas de Asia (ASER India y ASER Pakistán) y África del este (Uwezo). Luego, se les pidió determinar qué temas debía incluir un instrumento que, considerando el currículo mexicano de segundo y cuarto años de primaria y los instrumentos antes presentados, midiera lectura y operaciones matemáticas básicas. Estos resultados se analizaron a partir de la metodología cualitativa.

En la segunda fase se creó un grupo de expertos bajo la dirección de los investigadores Laura Oliva, María del Pilar González y Luis Rey, del Instituto de Psicología y Educación de la Universidad Veracruzana. En total, participaron seis académicos y especialistas en matemáticas y español de educación normal. Con base en los resultados obtenidos en la fase anterior, este grupo desarrolló los reactivos que integrarían las diferentes versiones del instrumento, añadió al instrumento el ítem de comprensión de inferencias, modificó la complejidad de los reactivos puestos a prueba, creó nuevos reactivos para la segunda versión, y elaboró el banco de reactivos para formar, aleatoriamente, las tres versiones paralelas finales. Así, estos expertos consideraron que la prueba actuaba como un instrumento de medición de aprendizajes básicos de lectura hasta 2º de primaria y de matemáticas, de 2º y 4º de primaria.

Es importante mencionar que una de las maneras de obtener la validez de un contenido es a partir del juicio de expertos que seleccionan indicadores para las dimensiones y proponen los ítems que deben conformar el constructo de interés (Abad, Olea, Ponsoda y García, 2011; Bollen, 1989). Por esta razón, se afirma que el instrumento MIA cuenta con validez de contenido, pues, como lo afirma Bollen (1989), ésta se basa más en argumentos conceptuales que en pruebas empíricas.

Para obtener la confiabilidad del instrumento se determinaron tres fases, que a continuación se detallan.

Primera fase

El objetivo de esta primera fase fue poner a prueba el instrumento, detectar si los reactivos se escalonaban adecuadamente en orden de complejidad y revisar las instrucciones y la organización del instrumento.

La primera versión del instrumento contenía 16 reactivos totales, seis de lectura y 10 de matemáticas. Las pruebas de lectura eran letra, sílaba, palabra, enunciado, historia y comprensión; y, las de matemáticas, cantidades, números del 0 al 9, números del 10 al 999, suma, resta, multiplicación, división, problema de suma, problema de resta y problema de multiplicación. Cada una de las pruebas, tanto de lectura como de matemáticas –excepto las de historia, comprensión y los problemas de matemáticas–, incluía seis posibilidades. Los niños debían acertar en dos de tres oportunidades que ellos mismos eligieran entre las seis opciones. La manera de calificar era correcto = 1, incorrecto = 0. Se aplicó el instrumento completo y siguiendo el orden de dificultad de los reactivos.

Se realizó un muestreo de las escuelas por conveniencia, que incorporara población urbana, rural, indígena y mestiza; y un muestreo por cuotas para la selección de los sujetos (Hernández, Fernández y Baptista, 2001; Kerlinger y Lee, 2002).

La muestra en la primera prueba piloto estuvo conformada por 273 personas de entre 5 y 16 años de edad, con una media de edad de 10.3 (DS = 2.7), de las cuales 53.1% eran mujeres y el resto, hombres. El 65.6% de la muestra cursaba la educación primaria y, el 34.4%, la secundaria. Aproximadamente 30 sujetos por cada grado escolar fueron incluidos, desde 1º de primaria hasta 3º de secundaria. El 33% de la muestra pertenecía a escuelas federales y, el 77%, a estatales. El 17.9% asistía al turno vespertino y, el 82.1%, al matutino.

Esta prueba piloto se llevó a cabo en cuatro escuelas de educación básica: una primaria general y una secundaria general, ambas públicas, ubicadas en el municipio de Xalapa, Veracruz; y una primaria general y una secundaria agropecuaria, ubicadas en el municipio de Tatahuicapan, Veracruz.

Los reactivos se pusieron a prueba en esta fase mediante los porcentajes de acierto/falla. Con base en lo anterior, se logró su validez interna y se valoró si escalonaban en orden de complejidad. También se obtuvo una consistencia interna total y por dimensiones, a partir del índice de Kuder-Richardson (alfa de Cronbach). Tras crear y poner a prueba la primera versión del instrumento, el grupo de expertos y los autores decidieron los cambios a realizar en los ítems (véanse los resultados). Por último, se creó el banco de reactivos del cual se eligieron al azar los pertenecientes a cada versión.

Segunda fase

El objetivo de esta fase fue crear y poner a prueba las tres versiones paralelas del instrumento. La segunda versión estuvo conformada por un total de 10 reactivos, cinco de cada dimensión: para lectura, sílaba, palabra, enunciado, historia y comprensión; para matemáticas, números del 0 al 99, suma, resta, división y problema. Para formar las tres versiones del instrumento MIA se realizó una igualación horizontal. Con el fin de obtener un banco de reactivos para cada una de las pruebas, se crearon unos nuevos con las mismas características que las de los diez ítems obtenidos en las fases anteriores. Los ítems de cada una de las tres versiones fueron seleccionados, aleatoriamente, de ese banco. Las tres versiones completas se aplicaron, en forma de entrevista, a cada uno de los sujetos y se alternó el orden de aplicación: se comenzó con la prueba de sílabas de la dimensión de lectura y de acuerdo con el orden de dificultad de los reactivos. Las calificaciones fueron, correcto = 1, e incorrecto = 0.

Para seleccionar a los sujetos se realizó un muestreo de las escuelas por conveniencia y otro muestreo por cuotas (Hernández, Fernández y Baptista, 2001; Kerlinger y Lee, 2002).

La muestra estuvo conformada por 210 sujetos de entre 5 y 16 años de edad, con una media de 10.84 (DE = 2.93). El 57.1% eran mujeres y el 42.9%, hombres. El 59% fueron alumnos de primaria y el 41%, de secundaria. La distribución por grado se muestra en el cuadro 1.

El 61.4% de los participantes asistían a escuelas federales y el 38.6%, a escuelas estatales. El 100% cursaba el turno matutino.

Cuadro 1. Distribución de la muestra por grado escolar, segunda aplicación.

	Frecuencia	Porcentaje
1º primaria	24	11.4
2º primaria	21	10.0
3º primaria	19	9.0
4º primaria	18	8.6
5º primaria	20	9.5
6º primaria	22	10.5
1º secundaria	28	13.3
2º secundaria	30	14.3
3º secundaria	28	13.3
Total	210	100.0

Fuente: elaboración propia.

La aplicación se llevó a cabo en cuatro escuelas públicas de educación básica: una primaria y una secundaria técnicas ubicadas en el municipio de Xalapa, Veracruz y una primaria y una secundaria técnicas ubicadas en el municipio de Pajapan, Veracruz.

Por medio de los porcentajes de acierto/falla se corroboró que los reactivos estuvieran adecuadamente escalonados. La consistencia interna se obtuvo a partir del coeficiente de Kuder-Richardson (alfa de Cronbach), con los métodos de división por mitades (pares y nones) y discriminación entre altos y bajos. Para valorar las formas paralelas se realizó una igualación horizontal; se obtuvieron medias y desviaciones estándar; se aplicó una prueba ANOVA a los resultados de las tres versiones; y, finalmente, se obtuvieron coeficientes de equivalencia o formas paralelas por medio de correlaciones de Pearson.

Tercera fase

Una vez obtenida la validación de las tres versiones finales de la MIA en la fase anterior, se realizó la tercera aplicación en hogares del municipio de Xalapa. Esto se hizo para observar cómo se comportaba el instrumento con las instrucciones de aplicación definitivas y en la situación real de aplicación en los hogares. Se asignó una sola versión por sujeto, misma que inició a partir de la prueba del enunciado, en el caso de la lectura, y de la resta, en el de las matemáticas.

Se utilizaron las tres versiones de la MIA, cada una con un total de 10 reactivos, cinco de cada dimensión: para la lectura, sílaba, palabra, enunciado, historia y comprensión; para matemáticas, números del 0 al 99, suma, resta, división y problema. Las versiones se aplicaron en orden de aparición –la primera versión al primer niño; la segunda, al segundo niño; la tercera, al tercer niño– y a manera de entrevista.

El modo en que se aplicó el instrumento varió en comparación con las dos fases anteriores, ya que en esta ocasión se usó la forma definitiva que propone la metodología ASER, que es la siguiente. La dimensión de lectura inicia con la prueba de enunciado: si las respuestas son correctas, se avanza a la prueba de historia y, posteriormente, a la de comprensión. Si las respuestas no son correctas, se retrocede a la palabra y, si nuevamente son incorrectas, se retrocede hasta la prueba de sílaba. En la dimensión de matemáticas, se inicia con la prueba de resta; si se resuelven bien dos de tres intentos, se pasa a las pruebas de división y de problema; si, por el contrario, no se logran resolver las operaciones correctamente, se retrocede a la suma y, si continua fallando, a los números del 10 al 99. La calificación es correcto = 1, e incorrecto = 0.

Se realizó un muestreo representativo polietápico, probabilístico, estratificado, por conglomerados. Esto conlleva tres pasos: 1) la selección de conglomerados o unidades primarias de muestreo (UPM) mediante un muestreo aleatorio estratificado, sistemático y con probabilidad proporcional al tamaño; 2) la selección al azar de manzanas dentro del conglomerado; 3) la selección sistemática de viviendas en las manzanas (cada cuatro). Se entrevistó a todos los niños de entre 5 y 16 años de edad que viven en el hogar. Las unidades primarias fueron las secciones electorales.

La muestra estuvo conformada por 327 sujetos de entre 5 y 16 años de edad, con una media de 9.98 años (DE = 3.17). El 48.8% eran mujeres y el 51.2%, hombres. El 97.2% asistía a la escuela y el 2.8%, no asistía. El 5.7% eran alumnos de preescolar; un 70.3%, alumnos de primaria; el 21.1%, de secundaria y el 2.2%, de preparatoria. La distribución por grado se muestra en el cuadro 2.

Cuadro 2. Distribución de la muestra por grado escolar, tercera aplicación.

	Frecuencia	Porcentaje
Sin escolaridad	2	0.6
3º preescolar	18	5.7
1º primaria	49	15.5
2º primaria	35	11.0
3º primaria	40	12.6
4º primaria	32	10.1
5º primaria	32	10.1
6º primaria	35	11.0
1º secundaria	22	6.9
2º secundaria	18	5.7
3º secundaria	27	8.5
1º preparatoria	7	2.2
Total	321	100

Fuente: elaboración propia.

El 48.7% asistía a escuelas federales, el 46.5% a escuelas estatales, el 4.0% a escuelas multigrado, el 0.4% a planteles bilingües/biculturales y el 0.4% a escuelas especiales. La aplicación se realizó en los hogares seleccionados de los distritos electorales 8 y 10 de Veracruz (Xalapa rural y Xalapa urbano). Ante la ausencia de una unidad territorial homogénea determinada por las autoridades educativas mexicanas se seleccionaron los distritos elec-

torales como unidades mínimas de representatividad, dado que presentan cierta homogeneidad poblacional, están presentes en toda la república y poseen altos niveles de información estadística y cartográfica que facilitan el muestreo y la aplicación de campo (Instituto Nacional Electoral, 2014).

Se obtuvieron porcentajes y frecuencias de aplicación de cada versión, porcentajes de aciertos y fallas para corroborar el escalonamiento y los índices de dificultad y discriminación por reactivo; la consistencia interna se obtuvo por medio del coeficiente de Kuder-Richardson (alfa de Cronbach), la división por mitades (pares y nones) y las diferencias entre altos y bajos.

Resultados

Primera fase: creación de la primera versión

En esta primera fase se descartaron algunos reactivos. Al escalar los reactivos en orden de complejidad se enlistan a continuación los porcentajes de aciertos y fallas obtenidos en todas las pruebas (cuadro 3). Se observa que las dos últimas pruebas de lectura

Cuadro 3. Porcentaje de fallas y aciertos por prueba de la muestra total, primera versión.

Escala	Prueba	Acierto p	Falla q
Lectura	Letra	98.5	1.5
	Sílaba	94.5	5.5
	Palabra	91.2	8.8
	Enunciado	85.7	14.3
	Historia	70.7	29.3
	Comprensión	72.5	27.5
Matemáticas	Cantidades	98.2	1.8
	Números del 0 al 9	96.3	3.7
	Números del 10 al 999	91.9	8.1
	Suma	74.4	25.6
	Resta	39.9	60.1
	Multiplicación	48.0	52.0
	División	35.5	64.5
	Problema suma	62.3	37.7
	Problema resta	49.1	50.9
	Problema multiplicación	42.9	57.1

Fuente: elaboración propia.

y las seis últimas de matemáticas no escalaban adecuadamente y que había reactivos que no discriminaban por tener más del 90% en una sola opción de respuesta.

En esta primera versión, los índices de consistencia interna Kuder-Richardson (alfa de Cronbach) fueron de .87 para el instrumento total (16 reactivos), de .78 para la escala de lectura (seis reactivos) y de .83 para la escala de matemáticas (10 reactivos).

En cuanto a la validez interna de los reactivos, las correlaciones reactivo/calificación total para la dimensión de lectura estuvieron en su mayoría por arriba de .20, a excepción de las correlaciones de la prueba de letra (cuadro 4).

Cuadro 4. Correlaciones de Spearman entre reactivos de lectura (pruebas) y sumatoria de dimensión y total.

Lectura	Sumatoria de lectura	Sumatoria total
Prueba de letra	.129*	.054
Prueba de sílaba	.428***	.344***
Prueba de palabra	.554***	.491***
Prueba de enunciado	.675***	.561***
Prueba de historia	.839***	.561***
Prueba de comprensión	.792***	.608***

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Las correlaciones reactivo/calificación total para la dimensión de matemáticas fueron en su mayoría superiores a .20, a excepción de las pruebas de cantidades y números de 0 al 9 (cuadro 5).

De acuerdo con los resultados de esta aplicación, los autores, junto con el grupo de expertos, modificaron el instrumento para intentar escalonar los reactivos según la complejidad. Se eliminaron los reactivos que no discriminaban y aquellos que correlacionaban menos de 0.20 con la sumatoria total y la sumatoria de su dimensión. De la dimensión de lectura, se eliminó la prueba de letra y se modificó la de comprensión para aumentar su complejidad, específicamente, se cambió la pregunta de comprensión textual por una de comprensión inferencial. De la dimensión de matemáticas, se eliminaron la prueba de cantidades y la de números del 0 al 99 por ser muy sencillas para la muestra evaluada; se cambió la tercera prueba de números del 10 al 999 por números del 10 al 99; se homogeneizaron las restas, de modo que sólo quedaron las de decenas con acarreo; se eliminaron las multiplicaciones, ya que todos quienes las respondían correctamente también acertaban en las divisiones. Así, se decidió dejar sólo las divisiones, que se homogeneizaron a tres dígitos, con resultados

Cuadro 5. Correlaciones de Spearman entre reactivos de matemáticas (pruebas) y sumatoria de dimensión y total.

Matemáticas	Sumatoria de matemáticas	Sumatoria total
Prueba de cantidades	.025	.032
Prueba de números 0 al 9	.150*	.122*
Prueba de números 10 al 999	.405***	.397***
Prueba resta	.638***	.613***
Prueba suma	.718***	.703***
Prueba de multiplicación	.756***	.741***
Prueba de división	.784***	.769***
Prueba problema suma	.787***	.770***
Prueba problema resta	.723***	.715***
Prueba multiplicación	.767***	.755***

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

exactos y sin punto decimal. Por último, se eliminaron los problemas y, por consiguiente, quedó sólo uno que requería el uso de varias operaciones distintas.

Segunda fase: creación de la segunda versión, validación y creación de formas paralelas

En esta segunda fase se analizaron las propiedades psicométricas de la segunda versión de la MIA que resultó de los cambios arriba mencionados y las versiones paralelas creadas. Todas las versiones tuvieron reactivos escalonados correctamente. A continuación se muestran las frecuencias de errores y aciertos por versión (cuadro 6).

Los coeficientes de consistencia interna de Kuder-Richardson (alfa de Cronbach) para la escala total son buenos y, para las dimensiones, aceptables (cuadro 7).

Se observan correlaciones altas entre reactivos pares y nones altas (cuadro 8).

Se obtuvieron diferencias estadísticamente significativas entre altos y bajos en las medias de las sumatorias totales de las tres versiones (cuadro 9).

Para valorar las formas paralelas, se realizó una igualación horizontal de las tres versiones del instrumento MIA. Los reactivos de cada versión se obtuvieron aleatoriamente de un banco de reactivos elaborado por los expertos. Así, las tres versiones se comportan de manera similar: al aplicar la prueba ANOVA no

se encuentran diferencias estadísticamente significativas entre las sumatorias totales ni entre las dimensiones, independientemente de la versión (cuadro 10).

Se obtuvieron coeficientes de equivalencia muy altos (correlaciones entre 0.88 y 0.93) entre las tres versiones (cuadro 11).

Cuadro 6. Porcentaje de aciertos y errores: tres versiones paralelas.

Escala	Prueba	Versión 1		Versión 2		Versión 3	
		% Aciertos	% Fallas	% Aciertos	% Fallas	% Aciertos	% Fallas
Lectura	Sílaba	99.5	.5	99.0	1.0	99.0	1.0
	Palabra	97.6	2.4	97.6	2.4	96.7	3.3
	Enunciado	95.7	4.3	96.2	3.8	94.8	5.2
	Historia	91.4	8.6	92.4	7.6	91.9	8.1
	Comprensión	79.5	20.5	79.5	20.5	78.6	21.4
Matemáticas	Números 10 al 99	98.1	1.9	99.5	.5	97.6	2.4
	Suma	85.2	14.8	83.3	16.7	84.3	15.7
	Resta	72.4	27.6	70.5	29.5	70.0	30.0
	División	57.1	42.9	60.5	39.5	56.9	43.1
	Problema	25.2	74.8	34.9	65.1	34.9	65.1

Fuente: elaboración propia.

Cuadro 7. Índices de consistencia interna total y por escala de las tres versiones.

	Total (10 reactivos)	Lectura (5 reactivos)	Matemáticas (5 reactivos)
Versión 1	.81	.74	.73
Versión 2	.80	.74	.73
Versión 3	.82	.76	.77

Fuente: elaboración propia.

Cuadro 8. Correlaciones entre reactivos pares y nones en las tres versiones.

Escala	Versión 1	Versión 2	Versión 3
Total	.81***	.78***	.82***

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Cuadro 9. Pruebas t para la diferencia entre altos y bajos en las tres versiones.

	Media	t
Versión 1	Bajos=5.37	-17.76***
	Altos=9.38	
Versión 2	Bajos=5.44	-20.92***
	Altos=10.0	
Versión 3	Bajos=5.56	-19.65***
	Altos=10.0	

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Cuadro 10. Medias de las sumatorias totales y por dimensión de las tres versiones.

	Versión 1		Versión 2		Versión 3	
	Media	DE	Media	DE	Media	DE
Sumatoria total	8.02	2.01	8.13	2.02	8.04	2.13
Sumatoria total lectura	4.64	.87	4.65	.86	4.61	.92
Sumatoria total matemáticas	3.38	1.37	3.48	1.40	3.44	1.48

Fuente: elaboración propia.

Cuadro 11. Correlaciones de Pearson entre las sumatorias totales de las tres versiones.

	Versión 2	Versión 3
Versión 1	.91***	.88***
Versión 2		.93***

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Tercera fase: aplicación en hogares

En esta última fase se realizó la tercera aplicación en hogares de Xalapa, con el fin de observar el comportamiento del instrumento en la situación real de aplicación.

Como se mencionó en la metodología, se aplicaron las versiones en orden continuo al total de niños que hubiera en el hogar encuestado. En el cuadro 12 se presentan las frecuencias de aplicación de cada versión.

Todos los reactivos escalonaron adecuadamente en complejidad en las tres versiones, de acuerdo con el porcentaje de falla y

Cuadro 12. Frecuencias y porcentajes de aplicación de cada versión aplicada en hogares.

	Frecuencia	Porcentaje
Versión 1	133	40.7
Versión 2	89	27.2
Versión 3	82	25.1
Se desconoce la versión a la que respondieron	23	7.0
Total	327	100

Fuente: elaboración propia.

error y con los índices de dificultad y discriminación de los reactivos (cuadros 13, 14 y 15).

El índice de dificultad del reactivo (*p*) va de .00, que indica que nadie contestó el reactivo, a 1.00, que indicaría que todos lo respondieron adecuadamente. El valor ideal de *p* depende de varios factores, de modo que si únicamente se pretende detectar a algunos sujetos muy deficientes es mejor un valor promedio de *p* elevado.

El índice de discriminación del reactivo (*D*) nos habla de su eficacia para discriminar entre los individuos que obtienen bajas y altas calificaciones: mientras más alto es el indicador, más discrimina. Así, un valor de 1.00 significaría que ningún sujeto con calificación baja y todos los que obtuvieron calificación alta respondieron al reactivo de manera adecuada.

Cuadro 13. Porcentaje de aciertos y errores e índices de dificultad y discriminación, versión 1 N = 133.

Escala	Prueba	% Aciertos	% Fallas	P	D
Lectura	Sílaba	100.0	0.0	1.00	.00
	Palabra	91.7	8.3	.85	.31
	Enunciado	87.2	12.8	.76	.47
	Historia	73.7	26.3	.54	.92
	Comprensión	55.6	44.4	.50	1.00
Matemáticas	Números 10 al 99	97.7	2.3	.96	.08
	Suma	81.2	18.8	.65	.69
	Resta	54.9	45.1	.53	.94
	División	39.8	60.2	.50	1.00
	Problema	18.8	81.2	.35	.69

Fuente: elaboración propia.

Cuadro 14. Porcentaje de aciertos y errores e índices de dificultad y discriminación, versión 2 N = 89.

Escala	Prueba	Aciertos	Fallas	P	D
Lectura	Sílaba	100.0	0.0	1.00	.00
	Palabra	95.5	4.5	.92	.17
	Enunciado	92.1	7.9	.85	.29
	Historia	80.9	19.1	.67	.67
	Comprensión	65.2	34.8	.58	.83
Matemáticas	Números 10 al 99	100	0.0	1.00	.00
	Suma	86.5	13.5	.75	.50
	Resta	64.0	36.0	.50	1.00
	División	42.7	57.3	.50	1.00
	Problema	20.2	79.8	.33	.67

Fuente: elaboración propia.

Cuadro 15. Porcentaje de aciertos y errores e índices de dificultad y discriminación, versión 3 N = 82.

Escala	Prueba	Aciertos	Fallas	P	D
Lectura	Sílaba	100.0	0.0	1.00	.00
	Palabra	93.9	6.1	.89	.23
	Enunciado	85.4	14.6	.73	.55
	Historia	80.5	19.5	.64	.73
	Comprensión	62.2	37.8	.52	.95
Matemáticas	Números 10 al 99	98.8	1.2	1.00	.00
	Suma	86.6	13.4	.77	.45
	Resta	59.8	40.2	.55	.91
	División	39.0	61.0	.50	1.00
	Problema	15.9	84.1	.30	.59

Fuente: elaboración propia.

Los índices de consistencia interna totales mejoraron un poco en relación con la aplicación en las escuelas y son adecuados (cuadro 16).

En las correlaciones entre pares y nones se observan índices altos (cuadro 17), y son más elevados que los obtenidos en la aplicación en escuelas.

Se obtuvieron diferencias estadísticamente significativas entre los puntajes altos y bajos en las medias de las sumatorias totales de las tres versiones (cuadro 18), al igual que en la aplicación en escuelas.

Cuadro 16. Índices de consistencia interna total y por escala de las tres versiones.

	Total (10 reactivos)	Lectura (5 reactivos)	Matemáticas (5 reactivos)
Versión 1	.86	.77	.76
Versión 2	.81	.73	.72
Versión 3	.83	.77	.72

Fuente: elaboración propia.

Cuadro 17. Correlaciones entre reactivos pares y nones en las tres versiones.

Versión 1	Versión 2	Versión 3
.89***	.84***	.84***

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Cuadro 18. Pruebas t para la diferencia entre los puntajes altos y bajos de las tres versiones.

	Media	t
Versión 1	Bajos= 3.58	-26.01***
	Altos= 9.50	
Versión 2	Bajos= 4.43	-15.81***
	Altos= 9.48	
Versión 3	Bajos= 4.29	-14.69***
	Altos= 9.43	

*$p < .05$ **$p < .01$ ***$p < .001$
Fuente: elaboración propia.

Discusión y conclusiones

Los procedimientos aplicados permiten afirmar que el instrumento desarrollado (anexo 1) es confiable y válido para determinar, por medio de mediciones y de una manera sencilla, si los niños y adolescentes de entre 5 y 16 años pueden leer y hacer operaciones matemáticas básicas en México.

Se obtuvieron tres versiones paralelas del instrumento MIA, cada una con 10 reactivos escalonados según su complejidad, confiabilidades totales buenas obtenidas a partir de coeficientes de consistencia interna, correlaciones pares y nones y diferencias entre los grupos de puntajes altos y bajos. Todo ello tanto

en las aplicaciones en escuelas como en los hogares. De igual modo, se obtuvieron coeficientes de equivalencia muy altos para las tres versiones paralelas. Además, no se encontraron diferencias estadísticamente significativas entre las medias de estas tres versiones.

Cabe señalar que, a pesar de que en las tres versiones existen reactivos que podrían considerarse muy fáciles –como la identificación de los números del 10 al 99–, se decidió conservarlos para poder detectar a grupos extremos, es decir, a sujetos muy deficientes. Lo mismo sucede con los reactivos cuyos índices de discriminación son muy bajos: se decidió conservarlos, para poder hacer comparaciones con otros países.

Por último, se abre una agenda de investigación futura para cuatro áreas importantes. En primer lugar, la aplicación y el uso del instrumento para la Medición Independiente de Aprendizajes (MIA), ya validado, en diversos estados de la república mexicana y utilizando la metodología ASER. Esto implica reclutar a ciudadanos voluntarios que lo apliquen a una muestra estadísticamente representativa de hogares, de modo que podamos conocer el porcentaje de niños y adolescentes de entre 5 y 16 años que saben leer y hacer operaciones matemáticas básicas y compararlo con diversas variables (zonas geográficas, áreas rurales/urbanas, género) y distintas unidades territoriales, como los distritos electorales o entidades federativas. En este sentido, se espera que gracias a la simpleza del instrumento validado sea más fácil invitar a organizaciones civiles y reclutar a grupos de voluntarios para participar en estas evaluaciones ciudadanas en México y que los resultados que se presenten local y estatalmente sean más comprensibles.

La segunda área de desarrollo tiene que ver con la utilización de los instrumentos para realizar diagnósticos educativos comunitarios en diferentes unidades territoriales, como pueden ser el municipio, la localidad o incluso la escuela. El hecho de que este instrumento sea fácil de aplicar y de comprender permite pensar que podría usarse para llevar a cabo diversos diagnósticos con fines de investigación, pero también de intervención comunitaria y social. En este sentido, tanto las organizaciones civiles como los organismos públicos podrían generar diagnósticos complementarios a los existentes en la actualidad. Los instrumentos desarrollados son gratuitos y de acceso universal; pueden ser adaptados por otras regiones y países de América Latina y en sus lenguas nacionales, como en náhuatl, totonaco y popoluca, para el caso de Veracruz.

La tercera área de investigación tiene que ver con la posibilidad de realizar comparaciones entre países en vías de desarrollo que han aplicado una metodología similar, tanto a escala nacional como subnacional. Debido a que la metodología de levantamiento es equivalente en India, Pakistán, Kenia, Uganda y Tanzania,

donde existe información por distrito subnacional, es posible llevar a cabo algunas investigaciones comparativas en el territorio nacional, pero también en los subnacionales.

Por último, se abre una línea de investigación ligada al desarrollo y la evaluación de intervenciones innovadoras que tengan por objetivo regularizar las deficiencias de aprendizajes básicos detectadas por el instrumento en diversos contextos. Esto no sólo permitiría vincular las problemáticas identificadas con la investigación educativa, sino incorporar a mayores segmentos de la población –como las organizaciones sociales y civiles que participen como voluntarias en la implementación del proyecto– a la discusión educativa.

Referencias

Abad, F. J., Olea, J., Ponsoda, V., y García, C. (2011). *Medición en ciencias sociales y de la salud*. Madrid, ES: Síntesis.

Aboites, H. (2012). *La medida de una nación: los primeros años de la evaluación en México: historia de poder, resistencia y alternativa (1982-2012)*. México: UAM/CLACSO/Itaca.

Anderson, J. A. (2005). *Accountability in education*. Paris, FR: IIEP/Unesco.

Avilés, K. (2012, 6 de septiembre). Insostenibles, algunos resultados de Enlace, afirma Olac Fuentes. *La Jornada*, sección Sociedad y Justicia. Recuperado de: http://www.jornada.unam.mx/2012/09/06/sociedad/040n1soc

Backhoff, E., y Contreras Roldán, S. (2014). "Corrupción de la medida" e inflación de los resultados de ENLACE. *Revista mexicana de investigación educativa*, *19*(63), 1267-1283.

Backhoff, E., Sánchez, A., Peón, M., Monroy, L., y Tanamachi, M. de L. (2006). Diseño y desarrollo de los exámenes de la calidad y el logro educativos. *Revista Mexicana de Investigación Educativa*, *11*(29), 617-638.

Banerji, R. (2014). An intervention improves student reading. *Phi Delta Kappan*, *95*(6), 74-75.

Banerji, R., Bhattacharjea, S., y Wadhwa, W. (2013). The Annual Status of Education Report (ASER). *Research in Comparative and International Education*, *8*(3), 387-396. doi.org/10.2304/rcie.2013.8.3.387

Bautista, A. K. (2012). La desigualdad social bajo la prueba Enlace. *Reencuentro*, *64*, 27-45.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ: Wiley.

Campos, R. M., y Urbin, F. D. (2011). Desempeño educativo en México: la prueba Enlace. *Estudios Económicos*, *26*(2), 249-292.

CIESAS-UV (2014). Medición Independiente de Aprendizajes-MIA. Recuperado de: http://medicionmia.org.mx/

Corvalán, J., y McMeekin, R. R. (Eds.). (2006). *Accountability educacional: posibilidades y desafíos para América Latina a partir de la experiencia internacional*. Santiago de Chile: Cide/Preal.

Dreze, J., y Sen, A. (2013). *An Uncertain Glory: India and its Contradictions*. Londres, RU: Penguin Books.

Escudero Escorza, T. (2003). Desde los tests hasta la investigación evaluativa actual. Un siglo, el XX, de intenso desarrollo de la evaluación en educación. *RELIEVE*, *9*(1), 11-43.

Hanushek, E. A., y Raymond, M. (2004). *Does School Accountability lead to Improved Student Performance?* (National Bureu of Economic Research Working Paper No. 10591). Cambridge MA: National Bureu of Economic Research. Recuperado de: http://www.nber.org/papers/w10591.pdf?new_window=1

Hernández, R., Fernández, C., y Baptista, P. (2001). *Metodología de la investigación*. México: McGraw-Hill.

Hevia, F. J. (2014). *Peticiones, protestas y participación. Patrones de relación sociedad-gobierno en la educación básica en Veracruz a inicios del siglo XXI*. México: CIESAS.

Instituto Nacional Electoral (2014). Sistema de Información Geográfica Electoral. Recuperado de: http://cartografia.ife.org.mx/

Instituto Nacional para la Evaluación de la Educación (2014). Instituto Nacional para la Evaluación de la Educación. Recuperado de: http://www.inee.edu.mx/

Kane, T. J., y Staiger, D. O. (2002). The Promise and Pitfalls of Using Imprecise School Accountability Measures. *The Journal of Economic Perspectives*, *16*(4), 91-114.

Kerlinger, F. N., y Lee, H. B. (2002). *Investigación del comportamiento: método de investigación en ciencias sociales*. México: McGraw-Hill.

Kreiner, S., y Christensen, K. B. (2014). Analyses of Model Fit and Robustness. A New Look at the PISA Scaling Model Underlying Ranking of Countries According to Reading Literacy. *Psychometrika*, *79*(2), 210-231. doi.org/10.1007/s11336-013-9347-z

Latapí, P. (2004). *La SEP por dentro: las políticas de la Secretaría de Educación Pública comentadas por cuatro de sus secretarios (1992-2004)*. México: Fondo de Cultura Económica.

Levine, R. (2015, mayo 29). Friday Note: Making the Movement for Accountability and Learning/Hewlett Foundation. Recuperado de: http://www.hewlett.org/blog/posts/friday-note-making-movement-accountability-and-learning

Martínez, A., Bracho, T., y Martínez, C. (2010). Los Consejos de Participación Social en la Educación y el Programa Escuelas de Calidad: ¿mecanismos sociales para la rendición de cuentas? En A. Olvera (Ed.), *La democratización frustrada. Limitaciones institucionales y colonización política de las instituciones garantes de derechos y de participación ciudadana en México* (pp. 129-174). México: CIESAS/Universidad Veracruzana.

Martínez, N. (2014, 5 de febrero). ¿Por qué la SEP suspendió la prueba enlace? *El Universal*. Sección Nación. Recuperado de: http://www.redpolitica.mx/nacion/por-que-la-sep-suspendio-la-pruebla-enlace

McMeekin, R. R. (2006). Hacia una comprensión de la accountability educativa y cómo puede aplicarse en los países de América Latina. En J. Corvalán y R. R. McMeekin (Eds.), *Accountability educacional: posibilidades y desafíos para América Latina a partir de la experiencia internacional* (pp. 19-9). Santiago de Chile: Cide/Preal.

PAL Network (2015). People's Action for Learning Network. Recuperado el 2 de octubre de 2015 de: http://palnetwork.org/?lang=es

Ravitch, D. (2011). *The Death and Life of the Great American School System: How Testing and Choice are Undermining Education*. Nueva York, NY: Basic Books.

Rosenberg, T. (2014). In India, Revealing the Children Left Behind. Recuperado de: http://opinionator.blogs.nytimes.com/2014/10/23/in-india-revealing-the-children-left-behind/

Sánchez, C. (2014, 27 de abril). Lo que oculta el informe Pisa. Recuperado de: http://www.finanzas.com/xl-semanal/magazine/20140427/oculta-informe-pisa-7150.html

Santibáñez, L., Martínez, J. F., Datar, A., McEwan, P. J., Messan-Setodji, C., y Basurto-Dávila, R. (2007). *Haciendo camino: análisis del sistema de evaluación y del impacto del programa de estímulos docentes Carrera Magisterial en México*. Santa Mónica CA: Rand Education/Secretaría de Educación Pública.

Save the Children (2013). *The Right to Learn. Community Participation in Improving Learning*. Westport, CT: Save The Children.

SEP (2010). Lineamientos Generales para la Operación de los Consejos Escolares de Participación Social. *Diario Oficial de la Federación*, 8 de junio de 2010.

SEP (2015). Consejos Escolares de Participación Social. Recuperado de http://www.consejosescolares.sep.gob.mx/

Unesco (1990). *Declaración mundial sobre Educación para Todos y marco de acción para satisfacer las necesidades básicas de aprendizaje*. Jomtien, TH: Unesco.

Unesco (2000). *Marco de acción de Dakar: Educación para Todos*. Dakar, SEN: Unesco.

Unesco (2012). *Beyond 2015-Education for the Future. Key considerations for the Development of the Post-2015 Agenda*. París, FR: Unesco.

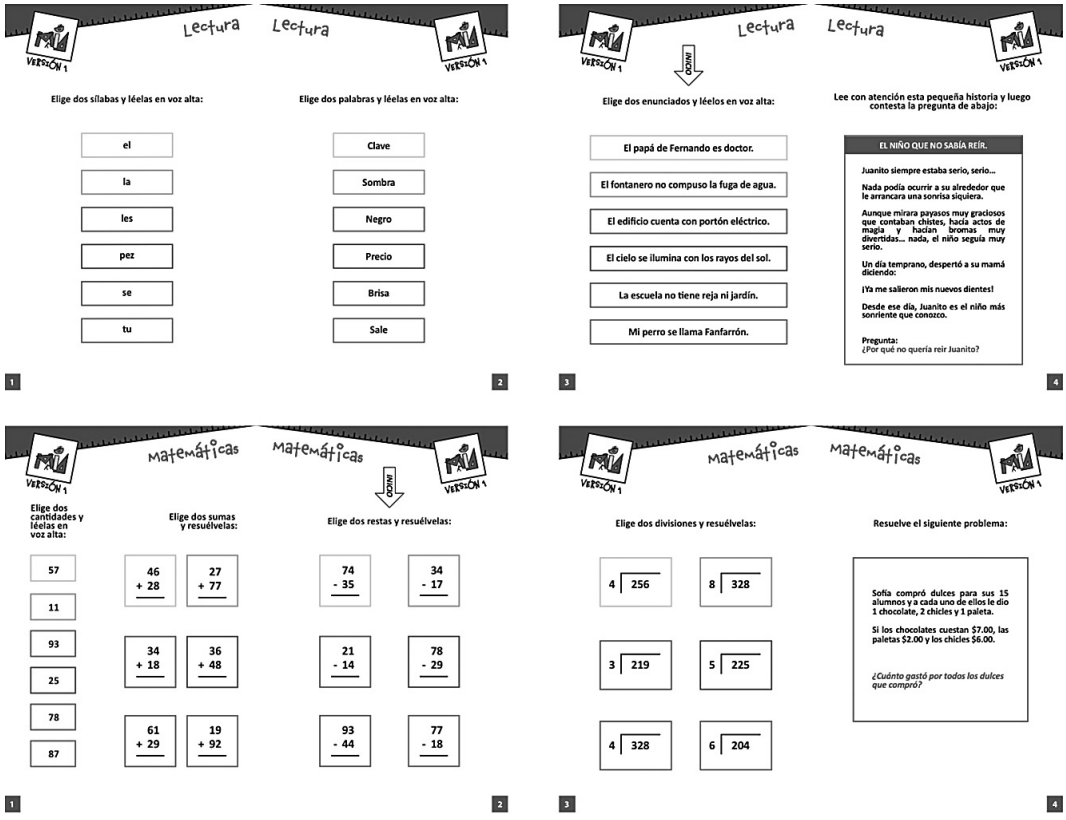
Unesco (2014). *Teaching and learning: Achieving quality for all*. París, FR: Unesco.

Zaidi, M. (2014, 14 de octubre). How Pakistan Fails Its Children. *The New York Times*. Recuperado de http://www.nytimes.com/2014/10/15/opinion/how-pakistan-fails-its-children.html

Zorilla, M. (Ed.). (2003). *La evaluación de la educación básica en México 1990-2000: una mirada contraluz*. Aguascalientes, MX: Universidad Autónoma de Aguascalientes.

Zurita, Ú. (2013). Paradojas y dilemas de la participación social en la educación básica en México. *Apuntes*, *XL*(72), 85-115.

Anexo 1.
Instrumento MIA-Medición Independiente de Aprendizajes



Fuente: elaboración propia.