

Acerca de la Validez de los Exámenes de la Calidad y el Logro Educativos (Excale)

María Araceli Ruiz-Primo* Jesús M. Jornet Meliá**
Eduardo Backhoff Escudero***

CUADERNO No. 20



Instituto Nacional para la
Evaluación de la Educación

COLECCIÓN CUADERNOS
DE INVESTIGACIÓN

ISSN 1665-9457

Acerca de la Validez de los Exámenes de la Calidad y el Logro Educativos (Excale)

María Araceli Ruiz-Primo* Jesús M. Jornet Meliá**
Eduardo Backhoff Escudero***

CUADERNO No. 20



Instituto Nacional para la
Evaluación de la Educación

COLECCIÓN CUADERNOS
DE INVESTIGACIÓN

ISSN 1665-9457

*Stanford University (USA) **Universitat de València (España) ***Instituto Nacional para la Evaluación de la Educación (México)

Este texto puede consultarse en: www.inee.edu.mx

MÉXICO, MAYO, 2006

CONTENIDO

Resumen	3
Introducción	5
Características de los Excale	6
● Objetivos y usos de los resultados del Excale	6
● Población a la que se dirigen los Excale	7
● Periodicidad de la administración de los Excale	7
● Dominios educativos de los Excale	8
● Estructura y unidad de análisis de los excale	9
Validez de los Excale en el contexto de la evaluación de sistemas educativos	10
Validez de los Excale como instrumentos de medida de la calidad y el logro educativos	13
Líneas de investigación para validar los Excale	16
● Marco conceptual para evaluar la validez de los Excale	17
● Marco de referencia para los estudios de validación	18
● Estudios dirigidos al diseño y elaboración de los Excale	19
Conclusiones	27

RESUMEN

Este documento trata de la validación de una nueva generación de pruebas de aprovechamiento desarrollados por el Instituto Nacional para la Evaluación de la Educación conocidos como Exámenes de la Calidad y el Logro Educativos (Excale). En el documento se discuten consideraciones conceptuales acerca de su validez, se propone un marco conceptual basado en premisas, y un marco de referencia para el diseño de estudios de validez. El marco de referencia se basa en la idea de que casi toda la información que se colecta durante el desarrollo de una prueba es útil para su validación. Este marco considera tres aspectos fundamentales: 1) las etapas de desarrollo de una prueba a gran escala (diseño y elaboración, administración, calificación, análisis e interpretación y usos), 2) la evaluación de argumentos confirmatorios (en favor) y no-confirmatorios (en contra) y 3) la evidencia lógica y empírica (uso de metodología cuantitativa y cualitativa). De acuerdo con este marco de referencia, el documento finaliza con una propuesta de líneas de investigación para la validación de los Excale.

ABSTRACT

In this document, we discuss issues around the validation of the tests, "Exámenes de la Calidad y el Logro Educativos – (Excale)," the new Mexican generation of large-scale assessment intended to measure educational achievement. We discuss conceptual issues around validation, propose an argument-based conceptual framework to validate the large-scale assessment, and propose a working framework for designing validation studies. The framework is based on the notion that almost all the information collected during the development of an assessment is helpful for its validation. Therefore, the working framework is based on three aspects: 1) the development stages of a large-scale assessment (i.e., design, administration, scoring, analyses and interpretation, and uses); 2) evaluation of confirmatory and disconfirmatory arguments (i.e., in favor and against); and 3) logical and empirical evidence (i.e., use of quantitative and qualitative methodology). Finally, the document suggests some studies to validate the Excale.



INTRODUCCIÓN¹

En México la cobertura educativa de la educación básica se ha ido resolviendo gradualmente y, desde hace una década, el interés se ha centrado en indagar qué es lo que efectivamente aprende la población escolar y cuáles son los aprendizajes básicos que forman parte del repertorio de los estudiantes.

El Instituto Nacional para la Evaluación de la Educación (INEE) se creó como una instancia, técnicamente independiente de la Secretaría de Educación Pública (SEP), para evaluar en forma válida, confiable y eficiente el logro escolar de los estudiantes mexicanos de educación básica y media superior. El propósito del INEE es retroalimentar al Sistema Educativo Nacional (SEN) y a las políticas que lo sustentan, así como dar cuentas a la ciudadanía de la calidad educativa del país (Poder Ejecutivo, 2002). Para lograr este propósito el INEE elaboró una nueva generación de pruebas nacionales que se conocen como Exámenes de la Calidad y el Logro Educativos (Excale). El diseño de estas pruebas inició en febrero de 2004 y su primera aplicación a nivel nacional se realizó en junio de 2005 (Backhoff y Díaz, 2005).

Siendo los Excale instrumentos con los que se pretende evaluar la calidad del SEN deben ajustarse a rigurosos estándares de calidad. Por el gran impacto social que tendrán sus características psicométricas son de gran relevancia, especialmente su validez (*American Educational Research Association, American Psychological Association, & National Council on Measurement in Education*, 1999). El tema central de este documento se enfoca precisamente en reflexionar sobre esta característica técnica: los elementos conceptuales que se deben considerar para evaluar la validez de los Excale. Para lograr este objetivo, el documento se dividió en cuatro grandes apartados: 1) Características de los Excale, 2) Validez de los Excale en el contexto de la evaluación de sistemas educativos, 3) Validez de los Excale como instrumentos de medida de la calidad y el logro educativos y 4) Líneas de investigación para validar los Excale. El trabajo termina con un apartado de conclusiones.

¹Nuestro agradecimiento al Dr. Guillermo Solano-Flores, de la Universidad de Colorado en Boulder, por sus sugerencias para mejorar este trabajo.

CARACTERÍSTICAS DE LOS EXCALE

Para orientar los estudios de validación es necesario considerar no sólo las características generales de los Excale, sino también las siguientes características específicas: objetivos y uso de los resultados, población a la que se dirige, periodicidad de la administración, dominios educativos que evalúa, y estructura y unidad de análisis. En las siguientes secciones describimos cada una de estas características generales y particulares.²

De manera muy general, y siguiendo la terminología de Madaus (1988), se puede decir que los Excale son pruebas externas, de gran escala, y de bajo impacto. *Externas*, porque las diseña, desarrolla y aplica una institución (INEE) cuya misión no es implementar la enseñanza, sino velar por su calidad a través de la evaluación del SEN. De *gran escala*, porque deben ser igualmente confiables y válidas para representar los niveles de logro educativo de los estudiantes de toda la nación, de todos sus entidades federativas y de las diversas condiciones educativas, culturales y económicas de los estudiantes. Finalmente, son de *bajo impacto*, porque los resultados no se utilizarán para tomar decisiones acerca de los alumnos, maestros o escuelas.

Asimismo, los Excale se diseñaron y construyeron tomando en cuenta las siguientes características técnicas: ser de tipo criterial, estar alineados al currículum nacional, y tener un diseño matricial. Son pruebas *criteriales* porque se diseñan para conocer con precisión el grado de dominio que el estudiante tiene sobre un conjunto de contenidos específicos (Popham, 1990). Están *alineadas al currículum* porque se elaboran con una metodología *ad hoc* para evaluar con mucha precisión los contenidos curriculares (por ejemplo: habilidades, conocimientos, y/o competencias) que se definen en los planes y programas de estudios nacionales (Nitko, 1994). Tienen un diseño *matricial* debido a que su desarrollo involucra la construcción de un gran número de reactivos para cubrir la mayor cantidad de contenidos curriculares, por lo que un estudiante determinado responde sólo a un subconjunto de ellos (Deng, Ferris y Hombo, 2003; Van der Linden, Veldkamp y Carlson, 2004).

Objetivos y usos de los resultados del Excale

La especificación clara de los objetivos de los Excale es de gran relevancia, ya que a partir de ésta se determina el tipo de evidencia que debe recogerse para validar las interpretaciones y usos de estos instrumentos. De acuerdo con Backhoff y Díaz (2005), los posibles usos de los resultados de los Excale son los siguientes:

- Construir una visión general de lo que los estudiantes aprenden como resultado de su escolarización formal.

²Para mayor información referirse a los trabajos de Backhoff et al. (2005).

- Conocer los puntos fuertes y débiles del aprendizaje de los estudiantes en distintos grados y áreas temáticas del currículum.
- Conocer las tendencias a lo largo del tiempo del aprendizaje de las Matemáticas, el Español, las Ciencias Naturales y las Ciencias Sociales.
- Permitir comparaciones del rendimiento escolar entre diversos estratos geográficos y económicos del país, así como entre distintas modalidades educativas.
- Identificar el crecimiento o estancamiento de los conocimientos en las disciplinas evaluadas desde la primaria hasta el bachillerato.
- Mejorar la evaluación de gran escala de las disciplinas que se evalúen.

Población a la que se dirigen los Excale

De acuerdo con Backhoff y Díaz (2005) los Excale se administran estratégicamente a los estudiantes de los grados terminales de cada ciclo escolar y que representen momentos de inflexión en el desarrollo del aprendizaje del sistema educativo mexicano: tercer grado de preescolar, tercer grado de primaria, sexto grado de primaria, tercer grado de secundaria, y tercer grado de bachillerato. Teniendo en cuenta lo anterior, el INEE se propuso un esquema de evaluación cíclico, con períodos que se repiten a lo largo de los años (ver tabla I).

Los Excale se aplicarán tanto a nivel nacional como estatal, de tal forma que permitan las comparaciones de mayor pertinencia para la toma de decisiones del sistema educativo: género, nivel socioeconómico, modalidad educativa, etcétera. Es importante mencionar que, debido a su diseño matricial, la unidad mínima de análisis es la escuela³ y no el estudiante.

Periodicidad de la administración de los Excale

Como se muestra en la tabla I, el INEE contempla un esquema de evaluación el cual permite hacer un seguimiento de las generaciones de estudiantes a lo largo de su escolaridad (Backhoff y Díaz, 2005). La periodicidad de las evaluaciones del aprendizaje se basa en ciclos de cuatro años en los grados ya mencionados. Este esquema permite dar seguimiento a las distintas generaciones de estudiantes, desde el tercer grado de preescolar al tercer grado de bachillerato (pasando por los grados intermedios evaluados) y comparar los mismos grados a través del tiempo, en períodos de cuatro años; tiempo suficiente para observar cambios significativos en los aprendizajes que se puedan atribuir al sistema educativo. De esta forma, se pretende que a partir de las pruebas de los Excale se puedan desarrollar análisis transversales y longitudinales de la evolución del SEN.

³Siempre y cuando la escuela tenga un número suficiente de estudiantes evaluados.

Tabla I. Programa de Evaluación del Aprendizaje del INEE

Grados	Años escolares											
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
tercer grado Preescolar			E, M				E, M				E, M	
tercer grado Primaria		E, M N, S				E, M N, S				E, M N, S		
sexto grado Primaria	E, M		E, M		E, M N, S		E, M		E, M N, S		E, M	
tercer grado Secundaria	E, M			E, M N, S		E, M		E, M N, S		E, M		E, M N, S
tercer grado Bachillerato				E, M			E, M N, S				E, M N, S	

Nota: E = Excale-Español; M = Excale-Matemáticas; N = Excale-Ciencias Naturales; S = Excale-Ciencias Sociales. Las celdas sombreadas indican evaluaciones con representatividad nacional, sin representatividad estatal.

Dominios educativos de los Excale

En la tabla anterior también se muestra que los Excale tienen la intención de proporcionar al país un conocimiento del rendimiento escolar que logran los estudiantes de cada grado evaluado en las asignaturas básicas del currículum nacional: Español, Matemáticas, Ciencias Naturales y Ciencias Sociales (Backhoff y Díaz, 2005).

Las características a tener en cuenta en este tipo de pruebas son aquellas que se aplican a grandes poblaciones (Jornet y Suárez, 1996); es decir, las pruebas deben estar referidas a los contenidos, competencias y/o habilidades, que puedan adquirirse en el sistema educativo mexicano con independencia del lugar de residencia (estado o localidad), de la modalidad educativa, así como de las condiciones económicas y culturales del alumno y su entorno.

De acuerdo con estas características, los Excale estarán siempre alineados al currículum, entendiendo éste como el conjunto de documentos y actividades escolares que dan racionalidad a la educación que se imparte en la escuela (planes y programas de estudio, libros para el profesor y el alumno, prácticas educativas, etcétera) (Backhoff et al., 2005). Es claro que si el currículum cambiara radicalmente en un futuro, las pruebas también deberán cambiar; y, en cualquier caso, los cambios o adaptaciones curriculares conllevarán necesariamente cambios importantes en los Excale.

El INEE ha desarrollado su propio modelo de alineamiento curricular para la elaboración de las pruebas que componen los Excale. En el documento que recoge el Plan General de Evaluación del INEE se indica que este modelo...

"... debe ser una herramienta fuerte, con reglas claras, que permita llegar a distintos comités a los mismos resultados. Para ello se deberá documentar todo el proceso de toma de decisiones, así como justificar los resultados que se encuentren, de tal mane-

ra que se pueda hacer público el análisis del currículum mexicano y los contenidos del sistema de evaluación de las pruebas nacionales". (Backhoff y Díaz 2005, p. 11)

El sistema de elaboración de las pruebas que componen los Excale se puede definir como un proceso sistemático de operacionalización del currículum en forma de instrumento de evaluación. Es decir, el proceso operacionaliza el currículum identificando sus elementos clave con base en un análisis sistemático y traduciendo estos elementos en reactivos. Para ello, se ha diseñado un procedimiento de trabajo que se sustenta en: a) comités de expertos de actuación sucesiva y b) análisis metodológicos rigurosos del contenido curricular. De acuerdo con estas características se puede decir que los Excale se sustentan en una orientación metodológica de alineamiento curricular (Nitko, 1994).

Estructura y unidad de análisis de los Excale

Atendiendo a las especificaciones discutidas en el Consejo Técnico, los Excale tienen las siguientes características (Backhoff, 2004): ciento treinta reactivos, en promedio, que se dividen en bloques de aproximadamente diez reactivos. Los bloques se construyen a partir de un muestreo estratificado de reactivos, considerando como estratos a las áreas de contenido o ejes curriculares de la asignatura a evaluar. Con la asignación aleatoria y estratificada de los reactivos se espera que los bloques sean equivalentes en cuanto a su extensión, dificultad y varianza. Cada bloque debe ser contestado en un período aproximado de treinta minutos. Cada estudiante responderá a dos bloques de reactivos por prueba. Los bloques de reactivos se combinan en pares formando distintos ordenamientos o bloques, tal y como se muestra en la tabla II.

Tabla II. Distribución hipotética de bloques de reactivos de los Excale

Alumnos	Bloques							
	a	b	c	d	e	f	g	h
A	1	2						
B			1	2				
C					1	2		
D							1	2
E		1	2					
F				1	2			
G						1	2	
H	2							1

Nota: los números indican el orden en que se presentan los bloques de reactivos a los alumnos.

Este diseño de pruebas basado en bloques implica que la unidad de análisis sea la escuela. Es decir, a partir de las respuestas de todos los alumnos de una misma escuela se puede establecer la estimación acerca del grado de dominio curricular que se tiene en ella. Por lo tanto, las inferencias acerca de los aprendizajes individuales de los estudiantes serán con mayor nivel de error, dado que cada alumno es evaluado parcialmente con respecto a la totalidad del currículum. Esta condición es necesaria considerarla tanto en el diseño de los estudio de validez, como en los usos e interpretaciones de los puntajes que se obtengan.

Una vez revisadas las características básicas de los Excale veremos cómo se puede entender su rol en el contexto del sistema educativo mexicano.

VALIDEZ DE LOS EXCALE EN EL CONTEXTO DE LA EVALUACIÓN DE SISTEMAS EDUCATIVOS

Los resultados que arrojen los Excale como indicadores escolares, en los planes de evaluación de sistemas educativos, deben considerarse en el contexto del Plan General de Evaluación del Aprendizaje del INEE en el que se insertan⁴.

Jornet, Suárez, & Pérez-Carbonell (2000) han definido la validez en el contexto de la evaluación de sistemas educativos como:

"...el grado en que un proceso de evaluación [...] atiende de forma adecuada los componentes metodológicos implicados con el fin de: representar de forma realista las características del objeto que evalúa, facilitar la consecución de los fines establecidos para el Plan de Evaluación, respetar los intereses de todos los implicados o afectados por la evaluación, representar el contexto en el desarrollo del Plan de Evaluación, en la modulación de la interpretación de resultados y en la adecuación de las consecuencias de la evaluación, y coadyuvar a la mejora o innovación de lo evaluado". (p. 343)

Con base en esta definición, es claro que la validez en el contexto de evaluación de sistemas educativos es un concepto *multidimensional* y *gradual*. *Multidimensional*, porque depende –y afecta– a todas las dimensiones implicadas en el plan de evaluación; y *gradual*, porque no se trata de una característica que se dé o no (todo/nada), sino que la validez se produce siempre en un grado determinado (y, claro está, nunca es completa). Para evitar que se omitan aspectos importante de esta validez, Jornet, Suárez y Pérez-Carbonell (*op. cit.*) proponen cinco dimensiones o facetas, con base en aquéllas identificadas en otros estándares, tales como los del *Joint Committee on Standards for Educational Evaluation* (1981, 1994), que pueden guiar las preguntas de validez en el contexto de evaluaciones de sistemas educativos.

⁴La elaboración de indicadores para evaluar sistemas educativos es siempre compleja y debe entenderse en relación a la utilidad del indicador dentro del plan de evaluación (véase a Selden 1990; Scheerens 1991; Landsheere 1996; Tiana 1993; Tiana y Muñoz 1997; García 2000).

Las dimensiones propuestas que se describen a continuación, son: representación del objeto evaluado, adaptación al proceso evaluado, integración contextual, participación e implicación de audiencias y compromiso con el cambio.

Representación del objeto evaluado. *“Se trata del grado en que, en la evaluación, se seleccionan indicadores y criterios que se ajusten a aquello que se desea evaluar. La identificación de instrumentos, fuentes de información, etcétera, forman parte de la misma” (ibidem, p. 346).* La calidad en la representación del objeto dependerá de la calidad de los Excale como instrumento de medida.

Adaptación al proceso evaluado. *“Se trata de llevar a cabo los diferentes elementos del proceso de la evaluación, respetando el proceso natural de desarrollo del programa educativo” (ibid, p. 347).* Esta adaptación implica atender diversos aspectos del proceso evaluado, tales como: a) la determinación del momento de aplicación de las pruebas; la aplicación de éstas no debe suponer una interrupción importante del desarrollo normal de la enseñanza y se debe realizar una vez que todos los contenidos incluidos en los Excale hayan sido trabajados en el proceso de enseñanza-aprendizaje, b) la eficiencia de la medida, de modo que se pueda garantizar la mejor calidad de la información, destinando el menor tiempo posible a su recogida, y c) la homogeneidad en el momento de aplicación de los Excale; esto es, no debe haber diferencias en el momento en que se recoge la información en cada una de las unidades educativas que se consideren en la evaluación, ya que esto puede conducir a un sesgo en la interpretación de los resultados.

Integración contextual. *“Otro elemento sustancial de la validez es la consideración de las diferentes variables que pueden afectar al programa evaluado y, por ello, deben ser atendidas como elementos de contextualización” (ibid, p. 347).* Los Excale deben aportar una información precisa y equiparable para todo el sistema educativo acerca de la calidad del aprendizaje de los estudiantes. Así, la representatividad muestral, el sistema de muestreo, los estratos a considerar en la configuración y definición de la muestra, entre otros elementos resultan clave en el proceso de validez. Asimismo, los Excale deben actuar sobre un estudio muestral convenientemente ajustado, no sólo en el diseño de la muestra, sino también en la muestra extraída.

Participación e implicación de audiencias. *“...Los fines de la evaluación sólo son posibles si se atiende adecuadamente la participación de los diferentes actores implicados en el programa. La validez de la evaluación, desde esta perspectiva, dependerá del grado en que se haya ajustado la participación de todos ellos, de acuerdo con su nivel de implicación y expectativas” (ibid, pp. 347-348).* Esta participación debe asegurarse desde el diseño de los Excale hasta la interpretación de la información. De este modo, es importante considerar la posibilidad de desarrollar estrategias de meta-evaluación con paneles de expertos y agentes sociales que, a partir de la información de los Excale y del Plan General de Evaluación del Aprendizaje del INEE, puedan establecer acuerdos acerca de las recomendaciones de mejora del sistema educativo (Jornet, Suárez y Pérez-Carbonell, 2000).

Compromiso con el cambio: credibilidad y utilidad. *“Esta faceta es probablemente la más determinante en cuanto a las consecuencias de la Evaluación. La Evaluación de Programas, desde la posición de que partimos, sólo es posible considerarla válida si aporta los elementos de mejora e innovación que se pretendía identificar y que justificaban la puesta en marcha del proceso” (ibid, p. 348).* Probablemente, este aspecto es uno de los más delicados, pues se trata de conseguir la credibilidad suficiente para que las recomendaciones hechas a partir de los resultados de los Excale sean útiles y tengan un impacto en el sistema educativo, ya sea en las acciones políticas, administrativas o docentes. Para poder tener impacto, es necesario que los Excale definan los elementos en los que pretende promover la mejora, qué información debe aportar y quién es el *usuario al cual se dirige*. En cualquier caso, la credibilidad y la utilidad es del plan de evaluación en su conjunto y, desde esta posición, **deben** analizarse las aportaciones y requerimientos de los Excale. Asimismo, debe entenderse que el INEE, como institución, es quien finalmente desarrolla un trabajo global orientado a asegurar estas facetas de la validez. No obstante, en el diseño de los Excale la estrategia de actuación de comités constituye una base importante para la consecución de esta faceta de la validez.

Si los Excale se consideran como medida de resultados escolares, producto o indicador acerca del aprendizaje de los estudiantes del SEN, entonces su validez deberá considerar las facetas discutidas. De estas consideraciones pueden derivarse algunas preguntas acerca de la validez de los Excale como indicadores del plan de evaluación del INEE. Estas preguntas proveen elementos de reflexión acerca de los factores que pueden coadyuvar en la validación de los Excale o atender contra ésta:

- ¿Existe una participación adecuada de los diferentes agentes implicados en el sistema educativo en el proceso de diseño del Excale?
- ¿Los Excale, como indicadores del plan de evaluación del INEE, aportan información de calidad?
- ¿El momento de aplicación de los Excale ha sido elegido atendiendo al aprendizaje desarrollado por los alumnos en cada grado?
- ¿Los Excale y el plan de evaluación del INEE se han diseñado de forma que la recogida de información sea operativa y no constituya una interrupción sustancial en el desarrollo habitual del programa educativo?
- ¿El plan de levantamiento de información a través de las diferentes escuelas se ejecuta en cada una de ellas en momentos equivalentes?
- ¿La muestra sobre la que se va a recoger información incluye alumnos representativos de todas las situaciones sociales, culturales y educativas que pueden darse en el sistema educativo mexicano y en la proporción adecuada a cada estrato?
- ¿La muestra utilizada se ajusta a la muestra proyectada?
- ¿En caso de desajustes, se utilizan criterios de ponderación de estratos, como elementos correctores?

- ¿Hay una participación adecuada de los diferentes elementos implicados en el sistema educativo en la interpretación de los Excale y en la elaboración de recomendaciones de mejora?
- ¿La información derivada de los Excale es creíble y utilizada por diferentes instancias?
- ¿La información derivada de los Excale es útil; es decir, tiene impacto para la mejora del sistema educativo?

A continuación revisamos el concepto de validez desde la perspectiva de los instrumentos de medida y derivamos, de igual modo, cuestiones acerca de la validación del Excale.

VALIDEZ DE LOS EXCALE COMO INSTRUMENTOS DE MEDIDA DE LA CALIDAD Y EL LOGRO EDUCATIVOS

La validez de un instrumento no es una característica inherente a éste. Lo que se valida no son los Excale en sí mismos, sino las *interpretaciones* que se les dan a sus resultados, de acuerdo con un propósito o uso determinado. Por consiguiente, el concepto de validez se refiere a la evaluación de la certeza de estas interpretaciones. Interpretar los resultados de una prueba es explicar el significado de sus resultados; es decir, clarificar sus implicaciones, o al menos algunas de las ellas. Messick (1989, p. 13) definió el concepto de validez como *"...un juicio evaluativo integral acerca del grado en que la evidencia empírica y las justificaciones teóricas dan sustento adecuado y apropiado a las inferencias y las acciones que se toman con base en los resultados de una prueba o de cualquier otro modo de evaluación"*.

El concepto de validez, entonces, se refiere al grado en que las inferencias hechas con base en los resultados de una prueba son justificadas, significativas y útiles. Messick (1989) dejó claro que la validez no es una cuestión de todo o nada, sino más bien de grado; todo depende de los usos específicos que se le den a los resultados de la prueba. Una prueba puede ser válida para un propósito, pero no para otro. La validez se establece tomando como base las evidencias acerca del grado de calidad de la información que proporciona la prueba para los propósitos y usos para los que fue construida. Toda validación requiere de la combinación de argumentos lógicos y evidencia empírica para sustentar las interpretaciones (Shepard, 1993).

Es claro que las demandas para justificar la validez de una prueba han aumentado. La visión tradicional que se tenía de la validez hace 25 años (validez de contenido, concurrente o predictiva) se ha reconceptualizado, haciendo ahora un énfasis especial en la validez de constructo (Cronbach & Meehl, 1955) como el concepto unificador para evaluar la calidad de una prueba (Messick, 1989). En esta reconceptualización, un aspecto crítico de la validez es el que se refiere a las consecuencias de la interpretación de una prueba. Messick (1989,

1995) argumentó que la evidencia acerca de la validez tiene que considerar las acciones que se generen tomando como referencia los resultados de una prueba, así como las consecuencias sociales relacionadas con los usos de los resultados.

La validación de cualquier instrumento de medida, de acuerdo con Haladyna y Downing (2004), es un proceso de investigación por medio del cual: 1) se crea un argumento plausible relacionado con la interpretación o el uso deseado de los resultados de una prueba, 2) se colecta y organiza la evidencia que sustente este argumento, y 3) se evalúa el argumento y la evidencia con respecto a la interpretación o uso deseado.

En el contexto de pruebas a gran escala, especialmente aquéllas de gran trascendencia social, la validez debe considerarse no como una lista de requerimientos, sino como un *argumento coherente* (Haertel, 1999; Kane 1992). El punto clave en un argumento coherente de validación radica en la evidencia en el cual se sustenta. Cualquier argumento de validez conlleva una serie de suposiciones o proposiciones para las cuales se requiere de un sustento empírico y lógico (Kane, 1992; Shepard, 1993). Por lo tanto, la validación debe ser un proceso de construcción y evaluación de argumentos *a favor* y *en contra* de las interpretaciones (que las apoyen o las refuten) y usos propuestos de la prueba en cuestión (Cronbach, 1988; Haertel, 1999; Kane, 1992; Shepard, 1993). Una evidencia masiva en favor de una suposición crítica no agrega mucho a la validez de la prueba si la evidencia para otra suposición crítica es escasa (Haertel, 1999). En el caso de las pruebas a gran escala es claro que ciertas evidencias son esperadas de forma habitual, como información técnica de la prueba, y se convierten entonces en un listado de requerimientos (por ejemplo, documentación del contenido cubierto, reporte de confiabilidades, errores estándares de medición, estimación de correlaciones con criterios relevantes). Sin embargo, el proceso de validación no es completo si no se pone atención a cómo las diferentes evidencias se vinculan y complementan para clarificar la validez de las interpretaciones y usos de la prueba.

“El propósito principal de la validez de constructo es justificar una interpretación particular del puntaje de una prueba, explicando lo que el puntaje de la prueba resume. La interpretación propuesta es el constructo de interés” (Moss, 1992, p. 233). Un plan de estudios de validez robusto requiere de la definición explícita de un *marco conceptual* que permita derivar hipótesis que sean comprobables a través de múltiples líneas de investigación que proporcionen evidencia acerca de las hipótesis planteadas (Moss, 1992, 1995; Shepard, 1993). Aunque en general se piensa únicamente en acumular evidencias para confirmar una hipótesis acerca de la validez de una interpretación o uso de los resultados de una prueba, es también importante recabar evidencias *no-confirmatorias* (Cronbach, 1989); es decir, evidencias acerca de hipótesis alternativas concernientes a la validez de interpretaciones o usos específicos de los resultados de la prueba. En otras palabras, en el proceso de validación es importante considerar factores que atenten contra la validez de la interpretación.

De acuerdo con Messick (1989, 1995), existen dos factores que atentan contra la validez del constructo que se mide: 1) la *sub-representación del constructo en la prueba* (SRC), y 2) la *varianza irrelevante al constructo* (VIC). El primer factor se refiere al grado en que una prueba es incapaz de considerar dimensiones importantes del constructo medido. El segundo, se refiere a todos aquellos aspectos de la prueba que pueden no estar directamente relacionados con el constructo medido, pero que pueden influir en la habilidad del estudiante para contestar una pregunta específica (por ejemplo: lenguaje, contexto cultural). La VIC es un error sistemático, no aleatorio, y se refiere a un proceso de contaminación de los resultados (por ejemplo: calificadores demasiado estrictos), aumentando o disminuyendo sistemáticamente las calificaciones de una prueba (Haladyna & Downing, 2004).

Los criterios que se han propuesto para sustentar la validez de una prueba han sido variados y cuestionados (Moss, 1995, Shepard, 1993). Messick (1989) por ejemplo, identificó las siguientes seis formas de evidencia para sustentar la validez: contenido, substantiva, estructural, generalizabilidad, aspectos externos y aspectos consecuenciales. Los criterios que Messick propuso a finales de los años ochenta han evolucionado y más recientemente se han considerado criterios nuevos. El Centro de Investigación y Evaluación de Estándares y de Estudiantes (Center of Research and Evaluation of Standards and Student Testing, CRESST) propuso en 1991 los siguientes criterios de validación (Linn, Baker, & Dunbar, 1991): representatividad del contenido/alineamiento, equidad, sensibilidad instruccional, transferencia y generabilidad, credibilidad, y practicalidad y costo.

Los *Estándares de Medición Educativa y Psicológica* (AERA, APA, NCME, 1999) sugieren obtener evidencia acerca de los siguientes cinco criterios:

- **Contenido.** Análisis de las relaciones entre el contenido de la prueba y el constructo medido.
- **Procesos cognoscitivos.** Análisis empírico y teórico del proceso de respuesta de los examinados y su correspondencia con el constructo medido.
- **Estructura interna de la prueba.** Análisis del grado en el que las relaciones entre los reactivos de la prueba y los componentes de la prueba corresponden al constructo medido (confiabilidad).
- **Relaciones de los puntajes de la prueba con otras variables.** Análisis de las relaciones con otras variables y su consistencia con el constructo medido.
- **Consecuencias.** Análisis de las consecuencias esperadas e inesperadas.

Un tipo adicional de evidencia que se ha empezado a considerar recientemente es el de la validez cultural (Solano-Flores y Nelson-Barber, 2001), la cual se refiere a los factores socioculturales que influyen en la manera en que los estudiantes que contestan una prueba interpretan y construyen un significado de los reactivos de una prueba.

“Estas influencias socioculturales incluyen los valores, creencias, experiencias, patrones de comunicación, estilos de enseñanza y aprendizaje, y epistemologías que son in-

herentes a las culturas de los estudiantes y a las condiciones socioeconómicas prevalentes en esos grupos culturales". (op. cit., p.555)

Los procedimientos tradicionales para tratar el efecto de la cultura en el desempeño en pruebas se basan en los estudios de sesgo con base en el funcionamiento diferencial de los reactivos (Camilli y Shepard, 1994), que se llevan a cabo al final o hacia el final del proceso de desarrollo de una prueba. Sin embargo, desde el punto de vista de los especialistas en validez cultural, tales procedimientos son necesarios pero insuficientes para asegurar la validez de una prueba y garantizar su equidad.

Debido a la sutileza de las influencias socioculturales y la complejidad del proceso de desarrollo de una prueba, para alcanzar su validez cultural es necesario cumplir con cuatro condiciones: 1) la colecta de datos sobre influencias socioculturales durante todo el proceso de desarrollo de la prueba (por ejemplo: en entrevistas con los estudiantes, para determinar cómo interpretan los reactivos), no simplemente al final del proceso; 2) la inclusión de muestras poblacionales de alumnos representantes de distintos grupos culturales y sociodemográficos a lo largo del proceso de desarrollo de la prueba (por ejemplo: en el piloteo de reactivos), 3) la participación de personas provenientes de estos grupos culturales y sociodemográficos (por ejemplo: docentes de escuelas rurales y urbanas, así como de todas las entidades federativas del país), y 4) la participación de especialistas en lingüística y antropología en los grupos de desarrollo y revisión de las pruebas.

LÍNEAS DE INVESTIGACIÓN PARA VALIDAR LOS EXCALE

Es claro que la búsqueda de evidencias para validar las interpretaciones de los resultados de una prueba es un proceso complejo. Más aún, conforme el impacto de una prueba aumenta, los requerimientos de validez también aumentan. Con base en esta información, como ya se sugirió anteriormente, los estudios de validez de los Excale tienen que considerar no solamente recabar evidencia confirmatoria, sino también no-confirmatoria de las interpretaciones y usos de sus resultados.

Como lo ha indicado Shepard (1993), la delimitación de un plan para la evaluación de la validez de una prueba requiere del establecimiento claro de los objetivos y propósitos de la misma (por ejemplo: qué es lo que la prueba pretende medir), así como de un marco conceptual o de referencia que guíe la investigación de su validez. En las siguientes secciones se presenta dicho marco conceptual para evaluar la validez de los Excale, el marco de referencia para los estudios de validación, así como los estudios dirigidos al diseño y elaboración de los Excale.

Marco conceptual para evaluar la validez de los Excale

Esta sección presenta algunas reflexiones de carácter conceptual acerca de los elementos implicados en la validación de los Excale. Las reflexiones se organizan en dos apartados: 1) Análisis de los objetivos de los Excale para su validez, y 2) Premisas para validar los Excale.

Análisis de los objetivos de los Excale para su validez

Una forma de analizar los objetivos del Excale es utilizar como referencia los objetivos propuestos por Haertel (1999) para las pruebas de gran escala: 1) proveer información para propósitos de rendición de cuentas, evaluación y comparación, 2) llamar la atención del público y los medios hacia cuestiones de educación y 3) cambiar la práctica educativa influyendo en el currículum y en la instrucción, o ejercer mayor influencia en el esfuerzo de administradores de escuelas, maestros y alumnos. El primer objetivo, proporcionar información útil, es el más tradicional y sin consecuencias; el segundo, mantiene la educación como centro de atención pública y, el tercero, intenta influir en la calidad de la educación.

Con base en los posibles usos descritos anteriormente, los Excale son pruebas de bajo impacto con objetivos similares al primero y segundo de los propuestos por Haertel. Esto es, su propósito principal es aportar información y mantener la educación en el centro de atención de la opinión pública de México y de sus dirigentes. Sin embargo, el tercer objetivo es más ambiguo y no están claros los límites de utilidad que se marcan para los Excale en su Plan General de Evaluación del Aprendizaje. Esta información es importante si se desea llegar a una especificación más precisa de las expectativas de validez pretendidas para los Excale respecto al tercer objetivo señalado. No obstante, más adelante incluimos unas propuestas básicas de preguntas a responder en su validación atendiendo al carácter curricular de las pruebas y a su uso genérico dentro de los planes de evaluación y misión del INEE.

Premisas para validar los Excale

Como ya se señaló, en el mejor de los casos, el nivel de análisis de los Excale sobre el que se establecen inferencias acerca del impacto del currículum es la escuela.⁵ De acuerdo con Haertel (1999), hay al menos una premisa importante en cualquier prueba de gran escala requerida para cualquiera de los tres objetivos mencionados anteriormente y que guía las evidencias de validez a considerar. La versión absoluta de esta premisa es: "*Las puntuaciones de la prueba indican el nivel de bondad del desempeño de las escuelas*" (Haertel, 1999, p.

⁵Al tener un diseño matricial los Excale, conviene agregar sus resultados en unidades mayores, como son las modalidades educativas y las entidades federativas.

7). La versión relativa de esta premisa es: “Las mejores escuelas deberán tener las calificaciones más altas”. Las ramificaciones de esta premisa son al menos dos: 1) la calificación de la prueba muestra qué tanto los estudiantes saben y pueden hacer, y 2) cuanto más sepan y puedan hacer los estudiantes de una escuela, mejor será la ejecución de la misma. De estas dos premisas se pueden derivar otras cinco, tres de la primera premisa y dos de la segunda (Haertel, 1999).

Las premisas que se derivan de la premisa primera son:

- Para contestar los reactivos de la prueba se requieren competencias escolares⁶ que están incluidas en el currículum.
- Las competencias demostradas en la prueba se pueden usar en otras tareas. Esto es, si los estudiantes pueden utilizar estas habilidades y conocimientos para contestar correctamente los reactivos, también pueden usarlas en otras tareas similares que se presentan en contextos diferentes.
- Una buena ejecución en las competencias evaluadas es señal de una buena ejecución de las habilidades y conocimientos no evaluados.

Las premisas que se derivan de la premisa segunda son:

- El propósito de la escuela es impartir conocimientos y desarrollar habilidades que estén incluidos en el currículum oficial. Así, las competencias evaluadas representan de alguna manera a las no evaluadas.
- La cantidad de conocimientos, habilidades y destrezas que los estudiantes poseen depende directamente de la instrucción que reciben en la escuela.

Con base en los objetivos de pruebas de gran escala propuestos por Haertel (1999), se concluye que los Excale tienen como propósito principal aportar información acerca del impacto del currículum en la ejecución de los estudiantes, no a nivel individual, sino a nivel de escuela u otras agregaciones mayores. Las premisas que se derivan de este objetivo determinan, de alguna manera, el tipo de evidencia que se necesita para validar los Excale.

Marco de Referencia para los Estudios de Validación

Casi toda la información que se colecta durante el proceso de desarrollo y uso de una prueba es relevante para la validez de constructo, siempre y cuando esta información sea evaluada usando el marco conceptual en el que se basa la interpretación propuesta (Moss, 1995). El marco de referencia que proponemos para guiar la validez del Excale se basa en la idea de que la validación de un instrumento deviene de la acumulación de evidencias en las diferentes etapas de su desarrollo y aplicación; desde el nacimiento mismo de las pruebas, hasta el uso de sus resultados. El marco de referencia propuesto considera, entonces, tres aspectos: 1) etapas de desarrollo de una prueba a gran escala, 2) evaluación de

⁶Entendiéndose por ellas como habilidades, conocimientos o destrezas que se adquieren en la escuela.

argumentos de validez a favor y en contra evidencias confirmatorias y no-confirmatorias) y 3) tipo de evidencia (lógica y empírica).

De acuerdo con este marco de referencia, las líneas de investigación propuestas se organizan en cinco etapas para la colección de evidencia de validez: *diseño y elaboración, administración, calificación, análisis e interpretación y usos*. Las líneas de investigación se enfocan en las dos perspectivas discutidas anteriormente, *estudios confirmatorios y no-confirmatorios* y contemplan una posición de complementariedad en cuanto a la metodología seguida; es decir, los estudios propuestos incluyen tanto *estudios cuantitativos como cualitativos* que permiten proporcionar evidencia lógica y empírica.

Todos los aspectos de validez propuestos en los estudios de acuerdo con las características de los Excale pueden alinearse con los criterios propuestos por los Estándares de Medición Educativa y Psicológica (AERA, APA, NCME, 1999), los Estándares de Calidad para Instrumentos de Evaluación Educativa (Martínez-Rizo et al., 2000), y los Criterios para la elaboración de estándares en evaluación educativa descritos por De la Orden (2000), aunque los estudios recomendados son más específicos que aquellos contenidos en los Estándares.

Los estudios propuestos se agrupan en cuatro de las cinco etapas descritas anteriormente: *diseño y elaboración de la prueba, administración de la prueba y recolección de información, análisis e interpretación de resultados, y usos de resultados*. Dentro de cada etapa se proponen preguntas acerca de la validez que tentativamente pueden ser contestadas. Por ahora, no se sugieren estudios de validez en la etapa de calificación, considerando que el tipo de reactivo que principalmente se utiliza en los Excale es de opción múltiple y no de respuesta construida. La lista de estudios podrá aumentar conforme los Excale se implementen y transformen a través del tiempo.

Estudios dirigidos al diseño y elaboración de los Excale

En esta etapa de desarrollo se han considerado dos tipos de estudios, aquellos que se refieren al diseño inicial de la prueba y aquellos que se refieren a la elaboración de los reactivos. Es importante señalar que los estudios en esta etapa tienen el propósito de presentar evidencia alrededor de dos grandes críticas que se han hecho a pruebas de gran escala (por ejemplo: NAEP; Champagne & Pearson, 2003): el alineamiento de las pruebas con el currículum y los requerimientos cognoscitivos para una ejecución exitosa.

a) Diseño Inicial. De acuerdo con las premisas de las pruebas a gran escala discutidas anteriormente, es importante que en los Excale se incluyan reactivos que requieran de las competencias que están incluidas en el currículum oficial de la Secretaría de Educación Pública (SEP). Por lo tanto, el alineamiento de los Excale con el currículum nacional implica proporcionar evidencia acerca de la relación que existe entre lo que se *intenta* que los estudiantes aprendan del currículum y su ejecución en los Excale.

Los estudios propuestos en esta etapa están encaminados a proporcionar evidencia acerca de la calidad del modelo utilizada para alinear los Excale con el currículum nacional; es decir, la aproximación utilizada para su diseño y elaboración. Una buena aproximación para alinear pruebas a contenidos curriculares no debe basarse únicamente en una lista de contenidos, sino también en la explicitación de todos los elementos que conforman el currículum (por ejemplo: ejes temáticos, subdominios, aproximación pedagógica), incluyendo las características cognoscitivas de las competencias involucradas en el currículum.

La primera serie de estudios sobre la validez de los Excale se dirigen a responder las siguientes preguntas: *¿En qué grado el proceso de desarrollo de la prueba es consistente con la práctica profesional?; ¿Los procedimientos para determinar los aspectos del currículum que se espera que los estudiantes sepan y conozcan son apropiados de acuerdo con el propósito de los Excale?; ¿El contenido de la prueba muestrea adecuadamente el contenido sobre el cual se sacarán conclusiones?*

El método más común para evaluar la integridad del modelo y su implementación es usar el *juicio de expertos*. Este tipo de evaluación se basa en un análisis *conceptual y lógico* de la aproximación y de los criterios para su aplicación. La tarea de los expertos es evaluar no sólo si los procedimientos para elegir el contenido representado en la prueba es adecuado, sino también, juzgar que los procedimientos a seguir sean lo suficientemente específicos como para que la implementación del modelo sea fidedigna.

En el caso de los Excale, desde el principio su finalidad principal fue establecer un plan de pruebas a largo plazo. En esta fase se definieron los propósitos, principios y referentes conceptuales de las pruebas, de los cuales dependen las fases restantes del proceso.

El proceso mismo de elaboración de los Excale ha incluido elementos dirigidos a la consecución de su validez. El diseño de las pruebas basado en la estrategia de comités de expertos de actuación sucesiva implica una garantía de procedimiento. Por ejemplo, la participación de especialistas en currículum, autores de libros de texto y especialistas en la enseñanza de la disciplina evaluada garantiza un análisis estructurado y plural de lo que se entiende por currículum en las diferentes áreas y grados. Esta estrategia de trabajo, que involucra alrededor de sesenta especialistas por prueba, tiene como objetivo garantizar la máxima pluralidad y objetividad, de forma que se dirige a aportar garantías para la credibilidad y utilidad del indicador.

El procedimiento de análisis curricular (basado en diversas etapas, como el análisis reticular, la elaboración de tablas de especificaciones) está diseñado de forma que intenta asegurar la máxima asociación desde la formulación teórica inicial del constructo hasta su especificación y operacionalización en forma de reactivos y de prueba.

Los estudios propuestos en esta etapa son los siguientes:

- *Revisión conceptual y lógica de los marcos de referencia de los Excale*. Se trata de averiguar si la fundamentación teórica o conceptual de las pruebas responde y es congruente con

sus propósitos, y si se basa en una aproximación actual de la evaluación educativa.

- *Revisiones lógicas y de objetivación del trabajo de los especialistas que diseñan y construyen los Excale.* Se trata de conocer la eficacia del trabajo de los distintos comités de expertos y la concordancia de los resultados obtenidos de estos comités con los resultados esperados. Asimismo, es importante valorar la pertinencia y representatividad de los especialistas que participan en cada una de las etapas de la construcción de los Excale.
- *Revisión de la documentación de las distintas etapas de diseño y construcción de los Excale.* El objetivo es analizar la pertinencia y congruencia de los documentos donde se consigna la información que da sustento a los exámenes, así como la utilidad y suficiencia de la información que éstos contienen.

b) Elaboración de reactivos y estructura de los Excale. Cualquier prueba es una simplificación de lo que se intenta medir, por consiguiente, es importante determinar qué distorsiones se pueden causar cuando se elige una forma específica de evaluación (Shepard, 1993). El análisis de contenido es un precursor necesario para hacer ciertas inferencias acerca de los resultados de una prueba, pero es insuficiente para defender su validez. Es necesario averiguar no sólo si los reactivos representan o no al contenido elegido, sino también si el tipo de reactivo elegido altera de manera alguna el constructo que se quiere medir (por ejemplo: evaluar la habilidad de redacción con un reactivo de opción múltiple).

En esta etapa es importante evaluar la calidad de los reactivos y la estructura interna de la prueba. Por ejemplo, es importante proporcionar evidencia de la relevancia del contenido curricular, de la representatividad del contenido y de la calidad de los reactivos de la prueba. Este análisis es primero lógico y después empírico. Algunas preguntas de validez relacionadas con esta etapa son: *¿La prueba es una muestra representativa del currículum que se enseña?; ¿En qué grado la conducta observada en la ejecución de la prueba constituye una muestra representativa de las conductas del dominio que se reflejan en currículum?; ¿Los Excale reflejan de forma adecuada las características del currículum?; ¿Qué aspectos del contenido del currículum no están representados en la prueba?; ¿Los reactivos desarrollados corresponden a las especificaciones de reactivos desarrolladas por los comités?; ¿El contenido y el formato corresponden a las especificaciones?; ¿La redacción de los reactivos es adecuada?; ¿El diseño de gráficos de los reactivos es adecuado? y ¿Los distractores de los reactivos representan errores plausibles y frecuentes en las posibles respuestas a los ítems?*

La evaluación inicial de la integridad de la implementación del marco conceptual y su traducción en reactivos y, finalmente en la prueba, es realizada también a través de una *revisión lógica* hecha por *expertos*. Aquí, es importante revisar la congruencia y consistencia del marco conceptual de la prueba, comparándola con otros similares. Así mismo, hay que recabar todas las evidencias documentales que soporten la fidelidad de su implementación.

Es de esperarse que un piloteo de la prueba ayude a determinar la estructura interna de los reactivos, de manera que el análisis lógico se confirme con un análisis empírico. En

el análisis empírico es importante proporcionar evidencia de que el tipo de reactivos elegido no impide a los estudiantes demostrar su conocimiento y sus habilidades, o por el contrario, les permite responder correctamente sin tener el conocimiento y las habilidades necesarias, debido a la familiaridad que se tiene con el tipo de reactivo (Shepard, Duran y Huges, 2003).

En el caso de los Excale, una vez que se realizó un análisis curricular de la asignatura y grado correspondientes, se definieron con mucha precisión los contenidos curriculares que se evaluarán en las pruebas y se construyeron los reactivos que se utilizarán para evaluar cada contenido curricular seleccionado. Los estudios de validez de mayor importancia en este aspecto son:

- **Correspondencia entre la estructura del examen y el currículum que se desea evaluar.** Se intenta conocer la representatividad de las habilidades y conocimiento que se evalúan.
- **Correspondencia entre los contenidos curriculares y la especificación de reactivos.** Se quiere saber si las especificaciones de los reactivos representan correctamente a las competencias curriculares.
- **Correspondencia entre las especificaciones de reactivos y los reactivos de los exámenes.** Se trata de conocer si las competencias que miden los reactivos representan fielmente a las competencias descritas en las especificaciones de los reactivos.
- **Ausencia de sesgo cultural y de género.** Se desea conocer la eficiencia de los mecanismos utilizados que aseguren que los reactivos no penalizan a los estudiantes por pertenecer a un género o a determinados grupos étnicos. Esto se hace por dos métodos: por el juicio de docentes en ejercicio y a través de análisis de DIF (por sus siglas en inglés, que aluden al comportamiento diferencial de reactivos) entre distintas poblaciones.
- **Comportamiento psicométrico de los reactivos.** A través del piloteo de reactivos se analizan sus parámetros estadísticos, tales como: la dificultad, la discriminación, el ajuste al modelo seleccionado, la funcionalidad de sus distractores, etcétera.

c) Estudios dirigidos a la administración de las pruebas y la recolección de información. Otro aspecto importante es el que se refiere a la representatividad de la población perteneciente al SEN. Las estimaciones de los Excale estarán basadas en la ejecución de una muestra de los estudiantes, y un aspecto fundamental para su validez es evaluar si el muestreo fue realizado de una manera que asegura que las estadísticas publicadas son representativas de la población de interés. Es esencial para las comparaciones longitudinales, entre estados, y para las mediciones de las brechas entre subpoblaciones, que cada grupo al ser comparado (por ejemplo: hombre y mujeres) esté apropiadamente representado. Desde el punto de vista estadístico, esto significa proporcionar evidencia de que todos los estudiantes del SEN tienen una probabilidad de ser elegidos en la muestra. De no ser cierta esta suposición, significaría que no todos los estudiantes están siendo representados en

los resultados publicados. Los estudios de validez en esta etapa están basados primero en un análisis lógico y luego en uno empírico. El primero se basa en una evaluación, nuevamente por expertos en estadística de la metodología utilizada en el muestreo. Sin embargo es de esperarse que durante la implementación y la administración de la prueba ocurran eventos inesperados (por ejemplo: un estudiante elegido en la muestra no se presente al examen) y la muestra ideal deje de serlo.

McLaughlin, Behuniak y Chromy (2003) proponen cuatro categorías de eventos que desvían el muestreo ideal del actual: listas incompletas de escuelas; no participación de escuelas; listas incompleta de estudiantes y no participación de estudiantes. Por lo tanto, una vez colectados los datos es importante realizar estudios de validez que se enfoquen hacia el impacto que cada uno de estos tipos de desviación, del plan ideal de muestreo, tienen en las inferencias realizadas.

Igualmente importante es la implementación de los exámenes, así como la lectura de la información colectada. La administración de los Excale debe ser uniforme y estándar para todos los estudiantes, independientemente de la escuela y de las condiciones contextuales de los alumnos. Asimismo, debe asegurar que los resultados de los exámenes se atribuyan solamente al nivel de aprendizaje o competencias académicas de los estudiantes y no a factores espurios de la administración de la prueba, tales como: diferencias en las instrucciones, condiciones adversas (falta de luz, ruido, condiciones físicas inadecuadas), etcétera.

Por el lado de la lectura de los resultados, hay que garantizar la precisión en la codificación y lectura de las respuestas de los estudiantes. Haladyna y Downing (2004) lo refieren cómo *saneamiento de las hojas de calificación*. Por ejemplo, borraduras incompletas pueden ocasionar *doble respuesta* y ello conduce a respuestas incorrectas cuando se utilizan lectoras ópticas de hojas de respuesta, conduciendo a una fuente de VIC.

Algunas preguntas de validez relacionadas con esta etapa son: *¿Cuál es el efecto de la participación o exclusión de ciertos grupos?; ¿Cómo afectan las alteraciones en la administración de la prueba la ejecución de los estudiantes?; ¿Se han seguido las normas especificadas en la Guía de administración de las pruebas? y ¿Se tiene garantía de que la lectura de resultados es confiable?*

Para el caso de los Excale, los estudios de validez de mayor importancia son:

- **Representatividad de la muestra poblacional.** Aquí se espera encontrar evidencias del uso de una metodología adecuada para el diseño de las muestras de estudiantes con una buena representatividad de los diversos estratos educativos y entidades federativas que se desean evaluar. Estas evidencias son de tipo documental, para lo cual se contrasta el marco y el diseño muestral, así como éste último con la muestra evaluada.
- **Estandarización de la aplicación de las pruebas.** Se intenta contar con evidencias que proporcionen información sobre las condiciones en que se administran los Excale y los

procedimientos que se siguen para asegurar que no haya sesgos ocasionados por diferencias en la aplicación de las evaluaciones. Aquí es importante considerar la forma en que se capacitó al personal a cargo de la aplicación de los Excale, los manuales y procedimientos utilizados, así como los problemas registrados en el proceso de aplicación de los Excale.

- **Captura y lectura de la información.** Se desea conocer la *calidad* de la información de las bases de datos que contienen los resultados de aprendizaje, para lo cual se deben utilizar procedimientos que aseguren la confiabilidad en la lectura y almacenamiento de la información de los Excale.

d) Estudios dirigidos al análisis e interpretación de resultados. Tradicionalmente, hay ciertos estudios de validez que se han considerado como necesarios para la validez de constructo. Este es el caso de la información acerca de la estructura interna de la prueba (por ejemplo: correspondencia entre la estructura de los puntajes con respecto a la estructura del constructo del currículum) y de sus relaciones externas con otras variables (por ejemplo: evidencia convergente y divergente) (Cronbach & Meehl, 1955; Messick, 1989). En el caso de la estructura interna de las pruebas se piensa no únicamente en las tradicionales preguntas psicométricas con respecto a la intercorrelación de los reactivos, sino en preguntas relacionadas con el peso de los diferentes componentes de la prueba (por ejemplo: ¿qué peso se le debe dar a geometría en una prueba de matemáticas?). Asimismo, es importante conocer la validez que tienen los niveles de logro utilizados en términos de su correspondencia con las competencias reales que manifiestan los estudiantes.

Algunas preguntas de validez en esta etapa son: *¿La prueba es adecuada considerando los criterios técnicos de un modelo de medida?; ¿Se aportan indicios suficientes basados en el ajuste de la prueba a diversos modelos de medida (Teoría Clásica, Teoría de la Generalizabilidad, Teoría de Respuesta al Ítem)? y ¿En qué medida los niveles de logro utilizados en los Excale representan las competencias de los estudiantes?*

En el caso de los Excale, esta etapa tiene como propósito la elaboración de los informes técnicos sobre los resultados de pruebas y cuestionarios de contexto, los cuales serán el sustento de los reportes anuales sobre el aprendizaje, que publique el INEE. Los estudios de validez de mayor importancia son:

- **Análisis métricos de ajuste a modelos de medida.** Dado que se utiliza el modelo de Rasch para calificar y escalar los puntajes, se debe verificar la unidimensionalidad de los Excale, el ajuste de los reactivos a la población estudiada, la confiabilidad y el poder de discriminación, entre otros aspectos.
- **Análisis del comportamiento diferencial de los ítems (DIF).** Se trata de conocer los sesgos empíricos de los reactivos que puedan afectar negativamente la interpretación de resultados de aprendizaje en ciertos grupos de interés educativo.
- **Evaluación de los niveles de logro.** Es importante conocer la medida en que son váli-

das las categorías utilizadas (y puntos de corte) para describir las habilidades y conocimientos que son capaces de realizar los estudiantes. Importa verificar que los estudiantes se agrupen empíricamente en estos conglomerados.

- **Interpretación de resultados.** Se desea conocer el grado en que los resultados de aprendizaje reflejan la situación real de los distintos grupos de estudiantes que se evalúan con los Excale. Esto se puede hacer comparando los resultados y conclusiones a las que llegan otras investigaciones nacionales e internacionales similares o equivalentes (por ejemplo: PISA, TIMSS, LLECE, ENLACE).
- **Estudios correlacionales y multidimensionales.** Se desea conocer el grado en que las variables contextuales, agrupadas en dimensiones o constructos, se relacionan y explican los resultados de aprendizaje proporcionados por los Excale. Al igual que en el punto anterior, es importante conocer la consistencia de los resultados comparándolos con otros estudios sobre logro académico de gran escala.
- **Validez de los niveles de logro.** Es importante conocer si las competencias escolares que se describen en los distintos niveles de logro de los Excale corresponden a los aprendizajes de los estudiantes. En otras palabras, es necesario conocer en qué medida estos niveles de logro son consistentes con otras medidas del rendimiento académico de los estudiantes.

e) Estudios dirigidos al uso de las pruebas. Al final, la validez y utilidad de los resultados reportados por los Excale reside en su capacidad para proporcionar información útil a las personas adecuadas, de manera que sea fácil su lectura y que minimice las interpretaciones inadecuadas de los resultados. Por lo tanto, la diseminación de los resultados y los reportes son tan importantes como la elaboración de la prueba, su administración y el análisis de la información colectada. No es fácil hacer comprender al público en general que cualquier prueba de aprendizaje es una medida imperfecta del aprovechamiento de los estudiantes, pero que, aún así, es posible alcanzar ciertas conclusiones válidas y confiables acerca del logro educativo de los estudiantes, así como de la calidad del sistema educativo de un país (Stancavage & Mullis, 2003).

En este sentido, es importante tener en cuenta que una de las premisas de los Excale es que las competencias escolares adquiridas por los estudiantes depende en gran medida de la instrucción recibida en su escuela. Sin embargo, es importante reconocer que entre el currículum formal y el aprendido, está el currículum *implementado*. Es de esperarse que en aquellas escuelas en donde el currículum implementado sea muy similar al currículum formal, el currículum logrado será mejor, lo cual se deberá reflejar en las puntuaciones de los Excale.

Por lo anterior, el impacto potencial que los resultados de los Excale puedan tener en las autoridades educativas, docentes, académicos y medios de información puede minimizarse, si la información no es útil, si no se presenta adecuadamente o si ésta es susceptible a ser malinterpretada.

Algunas preguntas de validez con respecto a esta etapa son las siguientes (Stancavage & Mullis, 2003): *¿Cuál es la mejor manera de reportar los resultados del Excale?; ¿Cómo se deben resumir las ejecuciones de los estudiantes de manera tal que su interpretación sea fácil y comprensible?; ¿Cuál es la mejor manera de diseminar los resultados?; ¿Cuál es la mejor manera de presentar los resultados a nivel de escuela?; ¿Cómo se puede proveer al público de un entendimiento substancial acerca de la ejecución de los estudiantes que vaya más allá de los puntajes numéricos?; ¿Cómo se puede explicar el error estadístico asociado a las pruebas y el tipo de inferencias que deben y que no deben hacerse de los resultados debido a este error? y ¿Cómo se deben interpretar los resultados en términos de lo que el estudiante aprende y no aprende del currículum formal?*

En el caso de los Excale, esta parte final del proceso evaluativo se relaciona con la utilidad de la información que se publica y el uso que distintos actores de la educación le dan, tales como: las autoridades, para proponer políticas educativas; los diseñadores del currículum nacional, para actualizar y mejorar los planes y programas de estudio; los profesores, para cambiar sus prácticas docentes; los académicos para documentar y explicar los fenómenos educativos del país; los medios para informar a la sociedad y la comunidad para que pidan cuentas a las autoridades. Así, proponemos los siguientes estudios de validez que den respuesta a estas interrogantes:

- **Análisis de oportunidades de aprendizaje.** Es importante conocer cómo la implementación del currículum formal en clase afecta el nivel de logro educativo de los estudiantes. Habrá que precisar qué contenidos se dominan, cuáles se revisan superficialmente y si éstos son los que se enseñan, o no, en las escuelas.
- **Análisis meta-evaluativos basados en paneles de expertos.** Es de interés conocer la importancia que distintos grupos de usuarios le dan a la información proporcionada por los Excale: las autoridades para la toma de decisiones; los docentes para cambiar sus prácticas educativas; los académicos para documentar; los medios de comunicación para informar; etcétera.
- **Análisis de trayectorias escolares.** Finalmente, es de suma importancia conocer si con los resultados de los Excale se puede evaluar el avance o retroceso de nuestro sistema educativo a través del tiempo, así como de los subsistemas que lo conforman, de manera válida y confiable.

CONCLUSIONES

En el contexto de los Excale, el concepto de validez se refiere al grado en que sus resultados comunican de forma *adecuada y precisa* la calidad del sistema educativo mexicano a políticos, administradores, educadores, académicos y al público en general. Por ejemplo, si el INEE con base en los resultados de los Excale, reporta que hay una brecha entre el nivel de aprovechamiento esperado en Matemáticas y el encontrado en los estudiantes de sexto grado de primaria, se pueden hacer muchas preguntas acerca del reporte, como son: *¿El tipo de habilidades matemáticas evaluadas en el Excale representan lo que les fue enseñado a los estudiantes?; ¿La manera en la que se evaluaron las habilidades matemáticas fue adecuada y precisa?; ¿La manera en la que se evaluaron las habilidades matemáticas fue justa para todos, o algunos estudiantes se vieron más favorecidos que otros?; ¿Los resultados reportados son representativos de la población del grado evaluado del país, así como de las entidades federativas y modalidades educativas?; ¿Los resultados reportados son lo suficientemente claros y precisos de manera que se eliminen posibles interpretaciones erróneas?* Estas constituyen tan sólo una muestra del tipo de preguntas que pueden hacerse con respecto a la validez de los Excale.

En este documento se han definido una serie de consideraciones conceptuales acerca de la validez de los Excale y se ha propuesto un marco de referencia para el diseño de estudios de validez que abarcan todas las etapas del desarrollo de estas pruebas. Sin embargo, es claro que existe una brecha entre la teoría de la validez y la práctica de su investigación (Cronbach, 1988; Messick, 1989; Moss, 1995; Shepard, 1993). El caso de los Excale, como el de muchas otras pruebas a gran escala en otros países, no es la excepción. Los requerimientos para una validación más o menos completa de una prueba son altos y difíciles de alcanzar. Por ejemplo, en este documento se ha reiterado la idea de que un buen plan de investigación de validez requiere no solamente de coleccionar evidencia confirmatoria, sino también no confirmatoria. Por desgracia, la mayoría de los reportes de validez son *confirmatorios* (Cronbach, 1989) y difícilmente se encuentra evidencia de estudios *no confirmatorios* que prueben la plausibilidad de hipótesis rivales. Más aún, un plan más o menos completo de investigación de la validez de una prueba es costoso y la limitación de recursos económicos, humanos y materiales afectan las decisiones de los estudios a realizar.

Creemos que el marco de referencia propuesto en este documento puede ayudar a decidir las prioridades de los estudios que han de realizarse para garantizar la validez de los Excale. Aunque aquí se han presentado ciertos estudios de validez, es claro que la agenda puede ampliarse considerablemente, si los recursos lo permiten. Por ejemplo, se puede pensar en estudios de validez para evaluar las siguientes premisas: para contestar los reactivos de la prueba se requieren competencias escolares que están incluidas en el currículum; las competencias evaluadas en los Excale representan de alguna manera a las no eva-

luadas; hay un alineamiento de las características cognoscitivas de las tareas involucradas en el currículum y en los Excale. Estudios para confirmar y refutar las dos primeras premisas requerirían de diseños experimentales que no se consideraron en los estudios propuestos. La tercera premisa requeriría de estudios de laboratorio donde se evaluara la validez cognoscitiva de las tareas de los Excale (Ruiz-Primo, Shavelson, Li, & Schultz, 2001).

Si esta agenda llegara a desarrollarse (ver Ruiz-Primo & Jornet, 2004), el establecimiento de las prioridades deberá tomar en cuenta el posible daño que pueda resultar, si cierto estudio de validez no se lleva a cabo, así como la probabilidad de ocurrencia (amenaza hipotética) de eventos que afecten la validez. El detrimento en la validez de los Excale sería serio si se descubre posteriormente que una amenaza hipotética a la validez fue real. Sugerimos, entonces, que el INEE decida las prioridades de los estudios que han de realizarse para establecer la validez de los Excale. Por ejemplo, siguiendo las recomendaciones hechas para el *National Assessment for Educational Progress* (NAEP) en los Estados Unidos, se podría pensar en los siguientes tipos de estudios:

- *Esenciales.* Estudios que si no se llevan a cabo permitirían que los Excale estuvieran sujetos a una crítica severa.
- *De alta prioridad.* Estudios que si no se realizan podrían conducir a críticas, algunas de ellas muy severas.
- *De mediana prioridad.* Estudios que si no se realizan podrían conducir a críticas no muy severas.
- *De baja prioridad.* Estudios enfocados a las amenazas de validez de baja probabilidad de ocurrencia, que el INEE puede justificar de manera contundente.

Creemos que hay tres preguntas que debieran guiar el establecimiento de la jerarquía de los estudios de validez sean las propuestas por Shepard (1993): *¿Qué es lo que los Excale pretenden evaluar?*; *¿Cuáles son los argumentos a favor y en contra de las interpretaciones esperadas de los Excale?*; *¿Qué hacen los Excale en el sistema educativo, para bien o para mal, además de lo que pretenden evaluar?*. Responder a estas preguntas requiere, necesariamente considerar los criterios propuestos por los *Estándares de Medición Educativa y Psicológica* (AERA, APA, NCME, 1999) y una larga agenda de estudios que se realicen en el contexto de un marco conceptual que explique claramente el argumento interpretativo de la prueba, que proponga claramente las premisas o suposiciones en los que se basa este argumento, y que proporcione evidencia a cada una de estas premisas, incluyendo la identificación y refutación de posible contra argumentos o hipótesis rivales (Kane, 1992).

Hasta ahora es claro que la validación de los Excale comenzó desde su diseño, desarrollando un modelo en el que se han tenido en cuenta diversos elementos dirigidos a asegurar su validez. No obstante, esta tarea de validación es una tarea compleja, que deberá llevarse a cabo durante toda la vida útil de los Excale en el sistema de evaluación que realiza el INEE, sustentada en un plan completo de investigación evaluativa, como el que aquí se sugiere.

REFERENCIAS BIBLIOGRÁFICAS

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- Backhoff, E. (2004). *Diseño matricial de las pruebas nacionales*. Documento mimeografiado. México: INEE, Dirección de Pruebas y Medición.
- Backhoff, E. y Díaz, M.A. (2005). *Plan General de Evaluación del Aprendizaje*. México: INEE.
- Backhoff, E., Monroy, L., Peón, M., Sánchez, A. y Tanamachi M.L. (2005). *Exámenes de la Calidad y el Logro Educativos (Excale): nueva generación de pruebas nacionales*. Cuaderno de investigación, N° 7. México: INEE.
- Camilli, D. y Shepard, L. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.
- Champagne, A. y Pearson, P. D. (2003). *Subject domain: What is being measured*. En: NCES Working Paper Series: NAEP Validity studies: An agenda for NAEP validity research.
- Cronbach, L. J. (1988). *Five perspectives on validation argument*. En H. Wainer y H. Brown (Eds.) *Test validity* (pp. 3-17). Hillsdale, NJ: Earlbaum.
- Cronbach, L. J. (1989). *Construct validation after thirty years*. En R. L. Linn (Ed.). *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.
- Cronbach, L. J. y Meehl, P. E., (1955). *Construct validity in psychological tests*. *Psychological Bulletin*, 52, 281-302.
- De la Orden, A. (2000). *Criterios para la elaboración de estándares en evaluación educativa*. Ponencia presentada en las I Jornadas de Medición y Evaluación Educativas. Valencia, España.
- Deng, H. Ferris, J. y Hombo, C. (2003). *A vertical Escheme of Bundling the NAEP Booklets*. Documento presentado en la Reunión Annual del NCME, Chicago.
- Haertel, E. H., (1999). *Validity arguments for high-stakes testing: In search of the evidence*. *Educational Measurement: Issues and Practices*, 19(4), 5-9.
- Haladyna, T., y Downing, S. M. (2004). *Construct irrelevant variance in high-stakes testing*. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Joint Committee on Standards for Educational Evaluation (1994). *Standards for evaluations of educational programs, projects, and materials*. New York: MacGraw-Hill.
- Joint Committee on Standards for Educational Evaluation (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: MacGraw-Hill.
- Jornet, J. M. Suárez, J. M., y Pérez Carbonell, M. A. (2000). *La validez en la evaluación de programas*. *Revista de Investigación Educativa*, 18, 2, 341-356.

- Jornet, J. M., y Suárez, J. M. (1996). *Pruebas estandarizadas y evaluación del rendimiento: usos y características métricas*. Revista de Investigación Educativa, 14(2), 141- 163.
- Kane, M. T. (1992). *An argument base approach to validity*. Psychological Bulletin, 112, 527-535.
- Linn, R., Baker, E. L. y Dunbar, S. B. (1991). *Complex, performance-based assessment: Expectations and validation criteria*. CSE Reporte técnico 331. Los Angeles, CA: UCLA, CRES-ST.
- McLaughlin, D.H., Behuniak, P. y Chromy, J. R., (2003). *Validity issues: Representing populations*. In NCES, Working Paper Series: NAEP Validity studies: An agenda for NAEP validity research.
- Madaus, G. F. (1988). *The influence of testing on the curriculum*. En L. N. Tanner (Ed.) *Critical Issues in curriculum. Eighty-seventh Yearbook on the National Society for Study of Education* (pp. 83-121). Chicago, IL: University of Chicago Press.
- Martínez, F. Backhoff, E., Castañeda, S., De la Orden, A., Schmelkes, S., Solano-Flores, G., Tristán, A. y Vidal, R. (2000). *Estándares de calidad para instrumentos de evaluación educativa*. México: CENEVAL.
- Messick, S (1989). Validity. En R. Linn (Ed.). *Educational Measurement* (pp. 201-219). New York, NY: Macmillan Publishing Company.
- Messick, S (1995). Validity of Psychological Assessment. *Validation of inferences from persons' responses and performances as scientific inquiry into score meaning*. American Psychologist, 50(9), 741-749.
- Moss, P. A. (1992). *Shifting conceptions of validity in educational measurement: Implications for performance assessment*. Review of Educational Research, 62, 229-258.
- Moss, P. A. (1995). *Themes and variations in validity theory*. Educational Measurement: Issues and Practice, 14(2), 5-13.
- Nitko, A. (1994). *A Model for Developing Curriculum-Driven Criterion-Referenced and Norm-Referenced National Examinations for Certification and Selection of Students*. Ponencia presentada en la Conferencia Internacional sobre Evaluación y Medición Educativas, de la Asociación para el Estudio de la Evaluación Educativa en Sudáfrica (ASSESA).
- Poder Ejecutivo (2002). *Decreto de Creación del Instituto Nacional para la Evaluación de la Educación*. México: Diario Oficial, 8 de agosto.
- Popham, J. (1990). *Modern Educational Measurement: A practitioner's perspective* (2a ed.) Needham Heights, M.A.: Allyn and Bacon.
- Ruiz-Primo, M. A. y Jornet, J.M. (2004). *Acerca de la validez de los Excale del INEE de México*. Documento mimeografiado. México: INEE, Dirección General de Pruebas del INEE.
- Ruiz-Primo, M. A., Shavelson, R. J., Li, M., y Schutlz, S. E., (2001). *On the validity of cognitive in-*

terpretations of scores from alternative mapping techniques. Educational Assessment, 7(2), 99-141.

Shepard, L. (1993). *Evaluating test validity.* In L. Darling-Hammond (Ed.) Review of Research in Education, 19, 405-450.

Shepard, L. Duran, R. P. y Hughes, G. (2003). *Subject domain: How is being measured?* En. NCES Working Paper Series: NAEP Validity studies: An agenda for NAEP validity research.

Solano-Flores, G., y Nelson-Barber (2001). *On the cultural validity of science assessments.* Journal of Research in Science Teaching, 38(5), 553-573.

Stancavage, F. B., y Mullis, I. V. S. (2003). *Validity and utility issues in NAEP reporting and data releases.* En NCES Working Paper Series: NAEP Validity studies: An agenda for NAEP validity research.

Van der Linden, W., Veldkamp, B. y Carlson, J. (2004). *Optimizing Balanced Incomplete Block Designs for Educational Assessments.* Applied Psychological Measurement, 28(5), 317-331.

TÍTULOS DE ESTA COLECCIÓN:

- 1.- *Los resultados de las pruebas de PISA*
Martínez Rizo, Felipe.
- 2.- *Factores externos e internos a las escuelas que influyen en el logro académico de los estudiantes de nivel primaria en México, 1998-2002*
Muñoz I, Carlos et al.
- 3.- *Contextualización sociocultural de las escuelas de la muestra de estándares nacionales (1998-2002).*
_____ *Determinantes sociales y organizacionales del aprendizaje en la Educación Primaria de México. Un análisis de tres niveles.*
_____ *Perfil de las escuelas primarias eficaces de México (2001)*
Fernández, Tabaré.
- 4.- *Tercer Estudio Internacional de Matemáticas y Ciencias Naturales (TIMSS): resultados de México en 1995 y 2000*
Backhoff, Eduardo y G. Solano.
- 5.- *Estudio Sobre las Desigualdades Educativas en México: la Incidencia de la Escuela en el Desempeño Académico de los Alumnos y el rol de los Docentes*
Treviño, Ernesto y G. Treviño.
- 6.- *Evaluación inicial de los procesos de calibración y equiparación de las pruebas del proyecto de estándares nacionales.*
Magriñá, Antonio.
- 7.- *Factores asociados al aprendizaje del lenguaje y las matemáticas en 13 estados de México.*
Cervini, Rubén.
- 8.- *Acciones de Evaluación en las Instituciones Públicas de Educación Media Superior*
Antonio, Rocío.
- 9.- *El proyecto PISA: su aplicación en México*
Vidal, Rafael, et. al.
- 10.- *La Comparabilidad de los Resultados de las Evaluaciones. Importancia y Dificultad de la Equiparación*
Martínez Rizo, Felipe.
- 11.- *Marginación y rezago educativo en México.*
Ávila, José Luis.
- 12.- *Pruebas y rendición de cuentas*
Martínez Rizo, Felipe.
- 13.- *Panorama Educativo 2004. La edición 2004 de Education at a Glance de la OCDE*
Martínez Rizo, Felipe.
- 14.- *El Diseño de Sistemas de Indicadores Educativos: Consideraciones Teórico-Methodológicas*
Martínez Rizo, Felipe.
- 15.- *La telesecundaria mexicana: desarrollo y problemática actual*
Martínez Rizo, Felipe.
- 16.- *Sobre la difusión de resultados por escuela*
Martínez Rizo, Felipe.
- 17.- *Exámenes de la Calidad y el Logro Educativos (Excale): nueva generación de pruebas nacionales*
Backhoff Escudero, Eduardo.
- 18.- *La Educación Mexicana en Education at a Glance 2005*
Martínez Rizo, Felipe.
- 19.- *Metodología para evaluar la calidad de las traducciones de las pruebas internacionales: el caso mexicano de TIMSS-1995*
Guillermo Solano-Flores, Eduardo Backhoff Escudero y Luis Ángel Contreras-Niño

LOS CUADERNOS DE INVESTIGACIÓN

Durante la década pasada nacieron y se fortalecieron en Latinoamérica los sistemas nacionales de evaluación educativa ante el desafío de contar con información apropiada sobre los conocimientos y competencias que los estudiantes adquieren en sus escuelas.

La mayoría de estos sistemas de evaluación han venido justificando su creación bajo la premisa de contribuir a la mejora de la calidad y equidad del sistema educativo. Así, el propósito fundamental es utilizar la información que arroja la evaluación, para rediseñar o ajustar políticas, planes, programas y prácticas pedagógicas y de gestión escolar.

Hacer bien la evaluación y difundirla suficientemente para que sus resultados sean utilizados en la toma de decisiones apropiadas, es de gran valor para el mejoramiento de las escuelas. Esto es lo que el Instituto Nacional para la Evaluación de la Educación (INEE) aporta a la educación mexicana.

Es necesario el diálogo entre personas e instituciones de diferentes sectores y de distintos países, capaces de desarrollar pensamiento crítico, promover debates, crear y fortalecer propuestas innovadoras, y unir esfuerzos encaminados a la búsqueda de una educación de calidad. A tal empeño contribuye también el INEE con la publicación de esta *Colección de Cuadernos de Investigación*, integrada por estudios técnicos, en los cuales convergen sustantivas aportaciones de especialistas en evaluación educativa de México y otros países.

