1 Las pruebas que tenemos

Alineación a los referentes

Para valorar las pruebas respecto de la alineación con su correspondiente referente curricular o de competencias, se emplearon 11 criterios y 22 subcriterios evaluativos. Para este informe sintético el contenido está organizado en tres apartados generales: el primero se refiere a la Planeación de las pruebas, e incluye su valoración en cuanto a los criterios que corresponden a la teoría de contenido curricular; a la determinación de la importancia relativa de los contenidos que evalúa; a la representatividad de ítems y subescalas respecto del dominio curricular y sus subdominios; y a la alineación de la prueba respecto al currículo en general. El segundo apartado, sobre Diseño y validación de especificaciones de ítems, comprende los criterios relativos a la definición de especificaciones para producir los ítems y a la complejidad cognitiva de los contenidos a evaluar. El tercer inciso Elaboración y validación de ítems, evalúa las pruebas a partir de los criterios relacionados con la existencia y contenido de un manual de diseño de reactivos; comité de redacción de ítems; manual de análisis de reactivos; comité de revisión de reactivos; y, sistema de revisión lógica de ítems.

Planeación de las pruebas

En pruebas de referencia criterial, como ENLACE-B, EXCALE y ENLACE-MS, el currículo es el referente para interpretar las puntuaciones que obtienen los estudiantes; por ello también es el referente principal para su planeación, diseño, construcción y validación. Al respecto, en el análisis de la información que aportaron las instituciones responsables de los tres instrumentos en sus manuales técnicos y en la documentación complementaria que nos fue entregada, se observaron problemas asociados con el currículo que sirvió de base para su desarrollo.

En efecto, tanto ENLACE-B como EXCALE tuvieron que esforzarse por alinear las pruebas a un currículo en continua transformación, en virtud del proceso que emprendió la SEP para integrar los niveles de la educación básica, y que culminó en 2011 con el acuerdo 592. Algo semejante sucedió en el caso de ENLACE-MS, que debió cambiar el foco inicial de la prueba y pasar de un esquema orientado al desarrollo de habilidades de lectura y matemáticas —que eran comunes en los perfiles de egreso de las instituciones de educación media superior—, a uno organizado por competencias, a partir del establecimiento en 2008 del acuerdo 442 para la adopción de un Marco Curricular Común (MCC) que buscó reducir la amplia dispersión curricular de la EMS en México.

A pesar de que la falta de estabilidad del currículo es una desventaja para cualquier prueba referida a un criterio orientado por dicho currículo, las tres pruebas trataron de superar dicha condición, para lo cual siguieron estrategias diferentes con logros distintos.

Planeación de las pruebas EXCALE

Para el caso de EXCALE, los especialistas del INEE siguieron una metodología que corresponde con los lineamientos técnicos y estándares que se mencionan en la literatura especializada. A partir de las definiciones de la SEP sobre el currículo de la educación básica se buscó definir un marco de referencia que orientara el desarrollo de las pruebas; después se procedió a efectuar un análisis formal del currículo para establecer su estructura (definición del universo de contenido) y, a partir de ella, determinar el contenido que era importante evaluar (determinación del universo de medida).

Los principales productos de estas acciones fueron: el modelo del logro educativo; el plan general de evaluación a largo plazo que permite ir abordando los diseños de las pruebas con un orden alterno, de forma que los cambios curriculares puedan atenderse con mayor oportunidad y precisión, a la vez que lograr una cobertura más amplia del contenido curricular a evaluar mediante un diseño matricial; y las retículas que muestran de manera gráfica los dominios, subdominios y contenidos curriculares de las asignaturas que se evalúan, así como las relaciones entre ellos; las retículas ayudan a determinar la importancia relativa de los contenidos y decidir cuáles de ellos serán objeto de evaluación en las pruebas.

Estos elementos quedaron plasmados en las tablas de contenidos o especificaciones de cada prueba, que posteriormente orientaron los trabajos de diseño de las tareas evaluativas, mediante especificaciones de ítems que detallan los atributos de los estímulos y las respuestas que debe exhibir cada uno.

Cabe señalar que dichos procesos fueron realizados y convalidados mediante operaciones de juicio por comités de especialistas con perfiles académicos y laborales adecuados, quienes además fueron formados para realizar acciones específicas; dichos comités contaron con protocolos de actuación, materiales de referencia, manuales de capacitación elaborados *ex profeso* y formatos apropiados para apoyar la realización de las acciones o, en su caso, para consignar sus juicios y adoptar sus decisiones.

En síntesis, el proceso de planeación que se siguió se ajusta a las prácticas de análisis curricular y detección y estructuración del contenido a evaluar en una prueba de las características referidas; los manuales técnicos particulares que se generaron fueron acompañados por una extensa documentación que aporta evidencias claras sobre los procesos que se siguieron y los productos que se obtuvieron en cada edición de las pruebas. Por lo anterior puede concluirse que el procedimiento seguido para el diseño de las pruebas EXCALE permite asegurar su alineación al currículum de referencia, así como la representatividad y relevancia de los contenidos cuyo dominio por parte de los sustentantes se pretende evaluar.

Planeación de las pruebas ENLACE para Educación Básica

En cuanto a las pruebas ENLACE-B, los especialistas de la DGEP planearon su desarrollo a través de una estrategia que consistió en conferir desde un inicio a los especialistas de la Dirección General de Desarrollo Curricular (DGDC), la responsabilidad de efectuar el análisis formal

del currículo para establecer la estructura del universo de contenido y, a partir de ella, determinar el universo de medida que identifica el contenido importante a evaluar en las pruebas.

Aunque la estrategia de división del trabajo entre los desarrolladores del currículo y los diseñadores de las pruebas no es la que se recomienda en la literatura especializada para la planeación y diseño de este tipo de instrumentos, permitió asegurar una continuidad entre el universo de medida y su explicitación en forma de prueba. De hecho, puede considerarse como garantía suficiente para contar, en cada edición de las pruebas, con un marco teórico actualizado que aseguró en buena medida la alineación entre el currículo y las pruebas.

No obstante, en la documentación revisada no fue posible identificar alguna evidencia relacionada con el proceso que siguieron los diseñadores del currículo de la SEP para representar la estructura del currículo, o para ponderar la importancia relativa de los contenidos a fin de determinar los blancos curriculares de primer orden cuyo dominio se evaluaría en cada una de las pruebas desarrolladas hasta el 2012. Tampoco se encontraron evidencias sobre la existencia de grupos independientes de especialistas que validaran las decisiones que se tomaron sobre tales asuntos. Una excepción a esta condición la encontramos en el proceso de planeación para desarrollar las pruebas en su edición de 2013, que estuvo a cargo de los especialistas de la DGEP, responsables de las pruebas. En el manual técnico correspondiente y en otros documentos revisados encontramos algunas evidencias de que se hizo una representación gráfica del dominio curricular completo y de los subdominios que lo constituyen, y que se identificaron los contenidos que se juzgó importante evaluar en cada prueba, con base en criterios técnicos para determinar la importancia diferencial de los contenidos.

Cabe señalar que en todas las ediciones de ENLACE-B se contó con un plan general de evaluación que orientó el desarrollo de los instrumentos y culminó con tablas generales de contenidos que incluyen áreas curriculares, subdominios y contenidos específicos a evaluar, así como las tablas de especificaciones de las pruebas, mismas que posteriormente fueron empleadas para elaborar los ítems de cada prueba. Sin embargo, hasta 2012 no se siguió un procedimiento homogéneo para construirlas, por lo que la estructura del dominio a evaluar se presenta en formatos diferentes, con elementos distintos y con niveles de desarrollo desigual, tanto en las materias de una misma asignatura, como entre las materias de asignaturas y años diferentes.

Las tablas en la edición de ENLACE-B 2013 son más homogéneas e incluyen los referentes del currículo adoptado en 2011, como son los aprendizajes esperados o los componentes que organizan los contenidos en ejes, ámbitos o temas de reflexión; algunas incorporan señalamientos sobre aspectos específicos a evaluar en cada contenido, el nivel de demanda cognitiva implicado o incluso una clasificación de su importancia relativa.

En síntesis, la situación descrita no corresponde plenamente con prácticas recomendadas de análisis curricular y detección y estructuración del contenido importante a evaluar en una prueba de las características de que se trata. Tampoco se ha basado en una estrategia de validación del análisis del universo de medida que incluya aportes de validación de grupos interdisciplinarios de especialistas que actúen de manera independiente. Por ello si bien el procedimiento seguido no permite asegurar la representatividad del contenido a evaluar en las pruebas, fue posible observar una evolución hacia mayores niveles de calidad técnica en cuanto al proceso de planeación de las mismas.

Planeación de las pruebas ENLACE para Educación Media Superior

La versión vigente hasta 2010 de ENLACE-MS evaluaba el dominio de habilidades básicas de lectura y matemáticas; a partir de 2011 evalúa indicadores de competencias de los campos disciplinares de Comunicación (comprensión lectora) y Matemáticas. Para transitar de una versión a otra, los especialistas del CENEVAL siguieron una estrategia de planeación que consistió en preservar, en la medida de lo posible, el proceso evaluativo original pero dando prioridad a las nuevas definiciones del Marco Curricular Común (MCC) establecido en la RIEMS.

En consecuencia, los referentes para la planeación de las pruebas mencionados en el Manual Técnico ENLACE-MS 2011-2012 fueron el MCC de la RIEMS, los de referentes de la versión anterior de la prueba ENLACE-MS, los de pruebas como PISA, TIMSS, SABER y ACREDITA-BACH, que se emplearon para propósitos específicos, y la documentación que se generó en el marco del proceso de adopción de la RIEMS por parte de las instituciones de educación media superior convocadas por CENEVAL para participar en la planeación de las pruebas. Estos elementos se integraron para constituir el marco de referencia de las pruebas.

En este contexto, el análisis del MCC, que funcionó como universo de contenido, permitió a los especialistas del CENEVAL identificar un universo de medida que sirvió de base para el desarrollo de las pruebas. No obstante, ENLACE-MS no evalúa el dominio de las cuatro competencias que establece el MCC de la RIEMS. La documentación disponible no permite precisar en qué medida esta aparente falta de representatividad del contenido a evaluar se debe a: 1) razones de conveniencia —por tratarse de una prueba de aplicación censal que emplea ítems de opción múltiple—; 2) fue ocasionada por la necesidad de dar continuidad al proyecto evaluativo anterior; o 3) se relaciona con las expectativas y propósitos definidos por la SEP para una prueba con las condiciones contextuales y características de ENLACE-MS.

> ENLACE-MS solo evalúa el dominio de las Competencias Disciplinares Básicas que son comunes a todos los egresados de la Educación Media Superior, y de ellas únicamente pone a prueba dos de los cuatro campos disciplinares básicos incluidos en la reforma: Comunicación (Comprensión lectora) y Matemáticas. Además, del primero, retoma solo 7 de las 12 competencias que establece el perfil de egreso en el MCC, y del segundo, únicamente 6 de las 8 competencias de dicho marco. El proyecto que desarrollaron los especialistas del Comité Académico Diseñador contó con el aval de los comités académicos validadores, del Consejo Técnico del CENEVAL, de las autoridades educativas de la SEP y de las instituciones de educación media superior en las que se aplica la prueba.

Los aspectos considerados en el marco de referencia de ENLACE-MS dieron lugar a dos tablas en las que se formaliza la planeación: la tabla que presenta la estructura de la prueba (con el número de reactivos para evaluar cada tipo de contenido de Comprensión lectora y Matemáticas, según el tipo de proceso cognitivo implicado) y la tabla con la taxonomía que define los niveles de complejidad por grupo de proceso cognitivo en cada campo. Aunque en el reporte técnico 2011-2012 estos componentes no se integraron en una tabla de especificaciones para cada prueba, en otro documento fue posible identificar los ítems específicos que correspondían a las definiciones contenidas en las mencionadas tablas.

Las decisiones del Comité Académico Diseñador sobre la planeación de cada prueba pasaron por un proceso de validación formal a cargo de los Comités Académicos Validadores. Sin embargo, el perfil de sus integrantes, su reducido número (3 en el comité de Comunicación y 2 en el de Matemáticas) y su poca representatividad (3 de la UNAM, 1 del INEE y 1 de AMAT), no corresponden con lo que sugiere la literatura, que enfatiza la necesidad de contar con expertos de varias áreas que representen no solo a la disciplina evaluada, sino también a la docencia y a la diversidad socioeducativa y cultural de quienes son evaluados.

Para explorar si los reactivos de las pruebas ENLACE-MS reflejan la estructura de contenidos planteada en su marco de referencia, y si están alineados con los aspectos del MCC que pretenden medir, tanto en relación al contenido como al nivel de demanda cognitiva previsto para el fin de la educación media superior, se hicieron dos estudios independientes. El primero se basó en estrategias de análisis de contenido, a partir de juicios formulados por comités de expertos (profesores experimentados, especialistas en las disciplinas, en medición, currículo e investigación educativa), para determinar en qué medida dichos juicios coincidían con los de los especialistas que diseñaron las pruebas en relación con la ubicación de los ítems en los mismos niveles de dificultad y en las mismas categorías y subcategorías del dominio cognitivo. El segundo estudio se basó en estrategias de análisis cognitivo, basadas en el proceso de respuesta evocado por los estudiantes ante los ítems de las pruebas, con el fin de verificar la congruencia entre dicho proceso de respuesta y el modelo teórico de las pruebas declarado en el Marco de referencia y en las especificaciones.

Los resultados del primer estudio mostraron coincidencias importantes entre los comités independientes y los responsables del diseño de las pruebas en la ubicación de los ítems en las mismas categorías del dominio cognitivo, tanto en Comprensión lectora como en Matemáticas. El estudio cognitivo encontró que en la prueba de Comprensión lectora, 15 de los 18 reactivos analizados presentaron congruencia entre el modelo subyacente y el modelo teórico; los tres ítems en que no hubo correspondencia evalúan otros procesos o niveles de complejidad cognitiva que no se declaran en su especificación. En contraste, en la prueba de Matemáticas solo en dos de los 18 reactivos analizados se observó la presencia de procesos de respuesta congruentes con la estructura declarada en el marco de referencia de la prueba; en los 16 restantes se observaron problemas de sobresimplificación o sobreestimación de la dificultad y complejidad cognitiva, el uso por parte de los alumnos de procesos de respuesta no declarados, distintos a los que establece la especificación, y errores en el diseño de los reactivos como la utilización incorrecta de términos o la redacción confusa de instrucciones.

Respecto a los procesos de capacitación de los comités académicos para analizar y estructurar el dominio curricular a evaluar, y los procedimientos o materiales y formatos utilizados para efectuar las operaciones de juicio y adoptar las decisiones, no se encontró información en el manual técnico 2011-2012. Un documento usado en la capacitación ilustra, de manera general, aspectos que se consideraron para analizar el MCC y la forma de proceder para establecer el perfil referencial y determinar la estructura de la prueba.

Diseño de especificaciones de ítems

Un aspecto crítico para tener evidencias de validez relacionadas con el contenido de una prueba referida a un criterio, es el análisis de la estructura del universo de medida que hace posible elaborar las especificaciones de contenido para la elaboración de ítems, las cuales incluyen

tanto la información estructural de la prueba como las especificaciones técnicas de unidades de dominio. En particular se detallan los atributos que deben tener los estímulos y las respuestas de cada ítem para probar el dominio de un contenido específico. De este modo, las especificaciones de ítems hacen posible contar con una visión precisa del universo de medida.

Diseño de especificaciones de ítems de EXCALE

Dentro del proceso de desarrollo de EXCALE que se describe en el manual técnico para el diseño de las pruebas, la elaboración de las especificaciones de ítems culmina la segunda fase denominada Estructuración de los EXCALE, y es la quinta etapa del proceso, que sigue al desarrollo de las tablas de contenidos o especificaciones de las pruebas.

En el Marco de Referencia, en los Manuales Técnicos e informes disponibles, especialmente en el propio Manual técnico para el desarrollo de especificaciones de reactivos, las especificaciones están claramente definidas; de manera que se homogeneiza todo el proceso de determinación de las unidades a evaluar.

El modelo para especificar los ítems de EXCALE es detallado e incluye procedimientos para contextualizar el contenido que se evalúa, la revisión de los documentos que justifican la selección del contenido (entre ellos la tabla de contenidos de la que forma parte), el análisis del contenido para determinar el nivel de dominio cognitivo implicado y la estrategia evaluativa que corresponde en cada caso, ya sea que se trate del dominio de un concepto o de un procedimiento, así como el desarrollo de la especificación. Ésta incluye secciones específicas para consignar la identificación del contenido, la descripción del aspecto del currículo a evaluar, la plantilla para especificar los ítems y uno como muestra para ilustrar la correcta aplicación de los elementos de la especificación, así como la bibliografía consultada para apoyar el rigor conceptual y disciplinario, o bien el apego al currículo.

La plantilla para especificar ítems incluye las características que debe tener cada uno: a) instrucciones particulares para responderlo; b) base del reactivo en cuanto a contenido, forma, extensión o redacción; c) información textual, gráfica o tabular que se puede utilizar; d) vocabulario que se debe o no usar; e) formato que debe tener cada ítem; f) características de la clave para ser considerada correcta o clarificar su enunciación, y de los distractores para que, siendo incorrectos, puedan considerarse plausibles por quienes no dominen el contenido que se evalúa.

El diseño de las especificaciones de ítems lo realizan los especialistas del Comité Elaborador de Especificaciones de Reactivos, quienes son capacitados con apoyo de un Manual elaborado *ex profeso* para dichas actividades. La capacitación incluye la revisión de la documentación generada por el comité técnico que previamente hizo la planeación de la prueba, en particular la retícula curricular de la materia y la tabla de contenidos o especificaciones; también incluye revisar el modelo para especificar los ítems, el procedimiento mismo de desarrollo de la especificación y la plantilla para cada ítem. Una vez desarrollada cada especificación de ítems, dos especialistas la analizan con el fin de detectar y corregir los errores que pudieran afectar la calidad de la construcción de reactivos. Los especialistas revisan la especificación en general, la plantilla y el reactivo muestra, y consignan sus observaciones sobre tales aspectos en un for-

mato que tiene 16 indicadores. El proceso se repite hasta que la especificación es aprobada por el coordinador y responsable académico.

Diseño de especificaciones de ítems de ENLACE para Educación Básica

No fue posible encontrar un apartado con especificaciones de ítems para orientar o normar su elaboración en los manuales técnicos de ENLACE-B y tampoco en la documentación proporcionada por la DGEP. Se pudo apreciar que en una misma tabla se presentan juntos elementos de la estructura del contenido a evaluar (tabla de especificaciones de la prueba) y de las tareas para evaluar el dominio de cada contenido (especificaciones de ítems). Aunque los formatos de especificaciones de ítems suelen incluir la estructura en donde se ubica el contenido a evaluar, en este caso el formato de las tablas no permite observar de manera independiente las características de los estímulos y las respuestas que deben tener los ítems.

> Hasta 2008 el manual técnico de ENLACE-B distinguía los dos componentes del desarrollo de las pruebas, que denominaba tablas de contenidos y tablas de especificaciones. En los manuales 2009-2012 ambos componentes se situaban en una misma tabla; y en el de 2013 las tablas de especificaciones, que ya consideran el currículo 2011, aparecen en una sección independiente. No obstante, en esta versión de la prueba no se usó un formato que homogenizara el diseño, por lo que los elementos de especificación difieren entre niveles, asignaturas y materias, e incluyen desde enunciados simples a manera de objetivos de aprendizaje asociados a temas o subtemas, hasta especificaciones más precisas que incluyen referentes del nuevo currículum como aprendizajes esperados, ámbitos o ejes curriculares en que se ubica el contenido a evaluar, o indicaciones particulares para su evaluación, así como descripciones de tipos de texto a emplear o atributos de la respuesta correcta y distractores, o incluso ejemplos que ilustran lo que se especifica.

Diseño de especificaciones de ítems de ENLACE para Media Superior

El manual técnico 2008-2010 de ENLACE-MS establece que las especificaciones de reactivos tienen como propósito proporcionar un marco normativo, claro y significativo, que aporte los elementos necesarios para que los elaboradores construyan reactivos adecuados para evaluar los contenidos y procesos cognitivos, así como los detalles técnicos para que los ítems resulten efectivos en la población objetivo y permitan generar interpretaciones válidas. Sin embargo en la documentación revisada no encontramos documento alguno que presente especificaciones para producir los ítems. Al parecer las tablas de especificaciones, además de precisar el número de reactivos requeridos para evaluar el dominio de subáreas y áreas curriculares, funcionan también como especificaciones de ítems, dado que incluyen ciertas características que deben tener los reactivos a elaborar, aspectos del contenido a evaluar o la precisión del nivel de dominio cognitivo asociado a cada contenido.

Aunque algunas columnas de dichas tablas incluyen información que comúnmente se presenta en una especificación de ítems, en otros casos aparece solo una delimitación general del contenido a evaluar. De hecho, en ninguno de los casos que revisamos se encuentran esos tipos de indicaciones de manera completa o sistemática. Tampoco se incluye algún ejemplo de ítem

que ilustre el cumplimiento de tales indicaciones. El único referente que aparece en todos los casos, es la mención del contenido que se evalúa, generalmente redactado como objetivo de aprendizaje.

En otras secciones del Manual Técnico de ENLACE-MS 2011-2012 aparecen también elementos con información que comúnmente incluye una especificación para producir un ítem. Por ejemplo, referentes para construir reactivos que miden competencias organizadas por niveles de complejidad específicos, o que permiten evaluar la variabilidad en el nivel de dominio de los sustentantes, según el proceso cognitivo que se moviliza, de conformidad con una taxonomía que se elaboró para cada campo disciplinar.

En síntesis, los elementos que se manejan en ENLACE-MS no corresponden con los procedimientos técnicos, las prácticas y los formatos que se detallan en la literatura especializada para la especificación de ítems, y que son necesarios para asegurar la producción de reactivos válidos, equivalentes y efectivos.

Elaboración y validación de ítems

Una etapa crucial del proceso de desarrollo de una prueba es la redacción de ítems. En ella convergen los demás elementos de la planeación del instrumento. Los ítems deben representar las unidades del universo de medida que se juzgan relevantes en la evaluación, como muestra del desempeño de los sustentantes. Para lograrlo se requiere un procedimiento formal de escritura de ítems, que asegure que representan el contenido a evaluar y que se adaptan al nivel de desempeño que se espera pueda darse en la enseñanza, como expresión de las oportunidades de aprendizaje brindadas a los alumnos.

La elaboración de ítems tiene dos requisitos fundamentales: primero, contar con un grupo de especialistas competentes en la disciplina de que se trate; en la operación del currículum en escuelas y aulas; en lingüística y teoría cognitiva, para cuidar que el leguaje no aumente la complejidad pretendida en los ítems; que incluya representantes de grupos que potencialmente pueden ser ofendidos o penalizados injustamente por los ítems; y especialistas en medición y evaluación. En segundo lugar, es fundamental que los redactores tengan formación adecuada y referentes claros que les permitan producir ítems de manera homogénea, ajustada a los niveles de calidad requeridos. El procedimiento adecuado para lograrlo es contar con un manual de redacción de ítems, diseñado según el propósito y contenido de la prueba que se construye, que tenga los elementos necesarios para que los diseñadores de ítems puedan interiorizar el tipo de producción que se espera de ellos.

Por otra parte, los miembros del comité a cargo de la validación de ítems deben tener un perfil semejante al de quienes los desarrollan. En esta etapa, los especialistas deben analizar la alineación de cada ítem con la especificación que lo produjo, y la correspondencia de ambos componentes con el plan general de evaluación, en particular con la tabla de especificaciones y la representación del universo de medida y, en general, con el currículo cuyo análisis sirvió de base para el desarrollo del instrumento. Además deben identificar y corregir posibles errores conceptuales, fallas al cumplir los lineamientos de redacción, sesgo, riesgo de ser ofensivos, complejidad cognitiva innecesaria, y falta de representatividad curricular, entre otros.

Elaboración y validación de ítems de EXCALE

Como parte del proceso de desarrollo de EXCALE, las etapas 6 y 7 de la fase III, "Construcción de reactivos de EXCALE", corresponden a la elaboración y validación de reactivos, respectivamente. La primera atañe a comités de constructores. El manual para construcción de reactivos que se usa para capacitar a dichos comités establece que la meta principal es asegurar que cada ítem represente la parte del currículo para la cual se especificó, por lo que debe garantizarse que tras revisar la retícula, la tabla de especificaciones de la prueba y la justificación del contenido a evaluar, junto con la correspondiente especificación de ítems y las normas de redacción estipuladas, se construyan tres ítems por cada especificación mediante la plantilla del reactivo elaborada previamente, a fin de seleccionar tras su pilotaje, el que posea mejores propiedades psicométricas y asegurar su validez.

Para proceder a la elaboración de los ítems el Comité Constructor debe recibir una capacitación que incluye revisar la documentación mencionada y luego realizar el trabajo, primero individualmente, para someterlo posteriormente a la consideración de otros miembros del comité en un trabajo colegiado.

En cuanto al proceso de validación de ítems, en el manual respectivo se explican los criterios para conformar los grupos de jueces que son especialistas en la enseñanza de las distintas disciplinas implicadas en las pruebas; los comités están conformados por docentes en ejercicio que provienen de las 32 entidades federativas, representando distintos estratos y modalidades escolares, y se procura que en su participación haya un equilibrio de género. Cada comité de validación está constituido por ocho personas que deben contar con un perfil que incluye dos tipos de características: indispensables (como ser profesor en ejercicio experimentado o conocer el currículo de la asignatura) y deseables (como estar inscrito en Carrera Magisterial o tener altas calificaciones en el Pronap).

La participación de los comités es precedida por una capacitación apoyada por el manual correspondiente y un sistema informático ad hoc. Los jueces desarrollan un trabajo tanto individual como colegiado al evaluar los reactivos de EXCALE. Deben hacer dos revisiones por ítem construido y dos reportes de validación por cada uno. Entre las funciones establecidas en el manual está la de revisar la documentación elaborada por los grupos que antecedieron su trabajo, (retícula, tabla de contenidos, especificaciones de ítems, reactivos elaborados, normas para su construcción que elaboró el INEE, claves de respuestas y en su caso rúbricas para calificar ítems de ejecución). Desde luego, su principal función es la validación técnica y cultural de los reactivos, lo que incluye revisar el grado en que los ítems del examen representan el dominio curricular a evaluar, juzgar la correspondencia del ítem con la especificación que guió su producción, evaluar el sesgo cultural y de género y detectar posibles fallas de construcción de los ítems como errores conceptuales, redacción compleja o proporcionar información innecesaria, entre otros.

Cada reactivo es evaluado por dos jueces independientes, quienes trabajan por rondas de reactivos antes de proceder a la evaluación colegiada que realizan los ocho profesores que integran cada comité. Para facilitar la revisión colegiada de los ítems el INEE desarrolló un programa informático con un formato electrónico de cinco páginas, el cual contiene la información sobre los reactivos y los indicadores para su validación. Además, se creó una página especial para el caso de la validación de los reactivos que se hace en la educación preescolar.

Elaboración y validación de ítems de ENLACE para Educación Básica

Existe información básica sobre las clases de reactivos que se elaboran para ser empleados como métodos de generación de ítems, pero no fue posible revisar los documentos que se citan en los manuales técnicos que se entregan a los diseñadores y revisores de ítems (Normas para la Construcción de Reactivos de Opción Múltiple; Normas de Presentación y Estilo; Normas para la Presentación y el Estilo en la Redacción de Reactivos de Opción Múltiple, Elaboración de Instrumentos de Medición). Se cuenta con recomendaciones para el diseño del ítem, su base y las opciones de respuesta, pero ubicadas en otros documentos normativos de carácter general.

Los manuales técnicos hacen referencia al análisis dimensional de los dominios en función de la taxonomía de Bloom, que se explica a los diseñadores y revisores de ítems, pero la relación de desempeños parece desvinculada o parcial. Esto es de importancia porque se indica que los ítems que se construyen deben corresponder al objetivo y nivel taxonómico asignados en la tabla de especificaciones. Sin embargo, aparentemente la tarea de definir el nivel cognitivo para el dominio de cada contenido parece haber recaído en los elaboradores de los ítems. De lo que sí hay evidencia en los manuales técnicos de ENLACE-B de 2009, 2010 y 2011 es de que cada ítem tuvo asignado un nivel taxonómico, el cual quedó registrado en la base de datos del Banco Nacional de Reactivos.

En general las evidencias disponibles muestran que las fases de desarrollo y validación de los ítems de las pruebas ENLACE-B son las mejor logradas por el grupo de especialistas de la DGEP. Sin embargo, existen aspectos particulares en los cuales se observan limitaciones importantes como la escasa información sobre los perfiles de quienes diseñaron los ítems o los validaron, que permita observar sus antecedentes, representatividad y nivel de pericia; el carácter genérico de los manuales, formatos y procedimientos utilizados para capacitar a ambos grupos o para realizar sus actividades; y la insuficiente descripción de los procedimientos técnicos que operaron al realizar sus actividades.

Respecto a validación, el manual técnico incluye una descripción sobre el detalle de la revisión de ítems por jueceo, incluyendo el perfil de los revisores y la frecuencia de revisiones. Aunque no se cuenta con la lista de los participantes ni con los formatos que utilizaron para consignar sus juicios, se menciona que en el jueceo intervinieron docentes de las 32 entidades federativas, además de profesores del SNTE y de las Áreas Estatales de Evaluación, así como especialistas de la Sociedad Matemática Mexicana.

En relación con la capacitación de los comités de revisores, no se detalla el proceso pero se indica que las áreas de contenido que se revisan incluyen la representatividad del dominio de contenidos curriculares, la formulación de cada reactivo y la presencia de sesgos. También se menciona que para la revisión de los ítems se utilizaron criterios de congruencia y correspondencia con los contenidos curriculares de los programas. Además, si bien se mencionan los criterios para aceptar, modificar o rechazar los ítems, en 2012 ya no se realizó el taller de jueceo, pues los reactivos fueron sometidos a un proceso de validación directa por los especialistas de la DGDC.

En el manual se muestra una tabla que especifica las ponderaciones que los jueces hacen de los atributos que se revisan en cada reactivo, aunque no se indica la forma en que se establecieron los acuerdos entre los jueces. Por su parte, los resultados de confiabilidad se estimaron con el coeficiente alfa de Cronbach, pero no se indica el error de medida.

En cuanto a la valoración del alineamiento de las pruebas con el currículo de referencia, se observó un claro interés por buscar la correspondencia y armonización de los ítems con la prueba y el currículo, sin embargo las evidencias disponibles no son suficientes para asegurar que ello se logró.

Elaboración y validación de ítems de ENLACE para Media Superior

En relación con el perfil de elaboradores de ítems de ENLACE-MS, un documento interno menciona nombre e institución de adscripción de casi 150 personas que han participado en la elaboración o validación de ítems, pero no se da información sobre su especialización académica, laboral o su representatividad respecto a la diversidad del país. Tampoco se indica quiénes de ellos elaboraron los ítems, y quiénes los validaron. El manual técnico de 2011–2012 señala que a los talleres de capacitación para redactores de ítems asistieron docentes y especialistas de asignaturas afines a los campos disciplinares básicos, y que el requisito principal fue que contaran con experiencia en el aula y, de ser posible, en la implementación de la RIEMS. También se menciona que en esos talleres los especialistas aprenden a elaborar reactivos de opción múltiple. No se cuenta con elementos suficientes para saber si los miembros de estos comités cubren de manera adecuada el perfil necesario.

Respecto a los referentes que permitan a los elaboradores de ítems homogeneizar su construcción y ajustarla a los niveles de calidad requeridos, hay un manual de redacción de ítems que contiene lineamientos para la construcción de reactivos de opción múltiple, que describe y proporciona ejemplos de los tipos de reactivos que tienen las pruebas que desarrolla el CENEVAL, e indica cómo clasificarlos y justificarlos. Sin embargo, este manual no se refiere en especial a ENLACE-MS, por lo que no aporta evidencias sobre la relevancia de las respuestas de los examinados para el dominio pretendido por esa prueba. Además, los lineamientos que aparecen en el documento resultan incompletos y poco explícitos para orientar el desarrollo de ítems efectivos.

En la documentación disponible no pudimos encontrar una guía o formato que oriente el diseño de los dos tipos de ítems que contiene la prueba operativa de 2014 (opción múltiple y multiítem de base común). No obstante, tanto el Manual Técnico 2011-2012, como el documento de lineamientos mencionado hacen una breve referencia al uso de la plataforma informática denominada Sistema de Administración de Bancos de Exámenes y Reactivos (SABER), para la elaboración y organización de reactivos, y para la elaboración de los cuadernillos de examen que se aplican.

De lo que sí hay evidencia es de que para guiar el desarrollo de los ítems se contó con una taxonomía del dominio cognitivo de cada campo disciplinar. Dicho sistema clasifica los procesos cognitivos a través de los cuales los sustentantes exhiben su nivel de dominio de distintas competencias disciplinares básicas. A cada ítem se le asigna un nivel taxonómico y un nivel de complejidad específico. Para el campo disciplinar de Comunicación (comprensión lectora), la taxonomía considera las categorías de Extracción e Interpretación (subcategorías Desarrollo de la comprensión y de la interpretación) y Reflexión y evaluación (subcategorías de la forma y el contenido). En cuanto al campo disciplinar Matemáticas, la clasificación considera las categorías Reproducción, Conexión y Reflexión.

En cuanto a la correspondencia de los ítems elaborados con el currículo, el manual comenta que el criterio principal fue asegurar que se obtuviera una muestra representativa de lo que todo bachiller debe dominar, en congruencia con el MCC y el perfil de egreso de la RIEMS, para lo

cual los reactivos de la prueba cubren toda la gama de procesos cognitivos que se indican en la tabla que presenta la estructura de la prueba y así evalúan contenidos que exigen un desarrollo básico, intermedio y avanzado de las competencias disciplinares básicas. Los resultados de los estudios adicionales mencionados en el inciso 1.3 parecen confirmar en buena medida lo que se dice en el manual sobre la cobertura de la gama de procesos cognitivos que suscitan los ítems.

Conclusión

En síntesis, los principales hallazgos de este apartado son los siguientes:

- Hay diferencias importantes entre los tres instrumentos revisados. ENLACE-B y ENLACE-MS son pruebas basadas en formas que se aplican a todos los alumnos, permitiendo dar cuenta del logro de cada uno de ellos, pero sobre un dominio curricular reducido. EXCALE se diseña y administra con una modalidad matricial, lo que hace posible una valoración más amplia del dominio curricular que se evalúa, pero solo permite dar cuenta del logro del conjunto de alumnos que respondieron la prueba. Estas diferencias remiten a una problemática general de política evaluativa que debe resolverse: la necesidad de decidir qué, para qué, cómo y con qué periodicidad se debe evaluar el dominio del currículo de referencia por parte de los estudiantes del sistema educativo nacional.
- También observamos coincidencias importantes entre los tres instrumentos: son pruebas de referencia criterial alineadas al currículo nacional (ENLACE-B y EXCALE) o a un marco curricular común (ENLACE-MS), y por ello tuvieron un reto similar: el referente para su diseño y desarrollo ha sido un currículum móvil al que han tenido que alinearse continuamente. Esta condición ha hecho difícil garantizar la oportunidad, relevancia, representatividad y alineamiento de las pruebas respecto al currículo que sirve de referencia para su construcción.
- Las evaluaciones se han administrado al final del ciclo escolar, lo que ha contribuido a sobredimensionar la función de rendición de cuentas y minimizar los usos diagnósticos, pedagógicos y formativos. Con mayor o menor detalle, en el proceso de diseño de las tres pruebas se hizo referencia a la utilidad pedagógica de sus resultados y a otros posibles usos educativos. Sin embargo, más allá de las declaraciones, no se explica de manera específica cómo se relacionan los aspectos del dominio curricular que se evalúa con la forma en que se implementa, enseña o evalúa en las escuelas, o la manera en que los alumnos, sus padres y profesores, o las escuelas y las autoridades educativas podrían utilizar los resultados de las evaluaciones para mejorar su participación, su función educativa o, en general, la calidad del servicio educativo que reciben los estudiantes.
- Una observación relevante que puede matizar algunas de las apreciaciones que hemos hecho, tiene que ver con el ciclo de desarrollo de cada instrumento. El proceso de planeación, construcción, validación, aplicación, análisis y reporte de resultados de una prueba censal que se administra anualmente, como ENLACE, impone retos distintos a los de una prueba como EXCALE, que opera bajo un plan cíclico y se administra con muestreo matricial. Ya se han comentado implicaciones como la capacidad de respuesta a cambios curriculares, pero se deben considerar otros aspectos como el traslape del ciclo de desarrollo de las pruebas de gran escala, que requiere más de un año, con ciclos educativos anuales o con los tiempos ceñidos que en ocasiones se tienen por presiones para dar a conocer los resultados. Todo esto puede ocasionar reducción de los tiempos necesarios para desarrollar actividades complejas. Algunas de las observaciones que he-

mos hecho parecen reflejar desfases entre ambos tipos de ciclo que deben ser considerados con cuidado.

- Aunque hay diferencias considerables entre ellos, los informes técnicos de las tres pruebas no documentan debidamente los procesos que siguieron y las evidencias que obtuvieron para asegurar la validez en cada edición. También pudimos apreciar diferencias notables en cuanto a calidad técnica, suficiencia y condición explícita de la documentación sobre las pruebas.

Además de consideraciones generales que se harán en el segundo capítulo y en la conclusión, de este apartado se desprenden recomendaciones particulares:

- Respecto a las dificultades que identificamos para asegurar la oportunidad, relevancia, representatividad y alineamiento de las pruebas a un currículo en movimiento, si se quiere desarrollar una prueba de referencia criterial que tenga calidad, es necesario que el referente tenga estabilidad y, más importante aún, que los sistemas de evaluación entren en acción una vez que el ciclo de operación curricular en las escuelas haya culminado con formación docente apropiada y con la disponibilidad de libros de textos, guías y demás materiales didácticos necesarios. La estabilidad del currículo evitaría la necesidad de reelaboración y adaptación periódicas que caracterizó a las pruebas valoradas.
- Se podrían explorar formas de combinar el muestreo matricial de contenidos y aplicaciones censales, buscando ampliar al máximo el acervo de contenidos curriculares cubiertos sin dejar de evaluar en cada ocasión los aspectos centrales del currículo. Una prueba censal podría incluir un núcleo de ítems que exploren el dominio de contenidos curriculares de primer orden de importancia que tienen los estudiantes, y a la vez incluir grupos adicionales de ítems diseñados y administrados en forma matricial, que exploren el grado en que los estudiantes en conjunto logran aprender del resto de contenidos curriculares.
- En atención a las diferencias que encontramos en cuanto a la calidad técnica, suficiencia y condición explícita de la documentación sobre las pruebas, así como los puntos débiles que pudimos observar en relación conlos procesos que siguieron y las evidencias que obtuvieron para asegurar la validez de las pruebas, se desprende la recomendación de instrumentar programas de alto nivel para la formación de los grupos de especialistas que desarrollan los instrumentos.

Muchas de las observaciones sobre las pruebas, hallazgos y recomendaciones de este apartado están estrechamente vinculadas. Algunas se refieren a fortalezas que pueden ser aprovechadas para generalizarse a todas las pruebas y otras a debilidades cuyo conocimiento también debería ser difundido, y que podrían superarse a partir de sugerencias como las que formulamos o por medio de una gestión más eficaz; otras implican retos y oportunidades comunes a los sistemas de evaluación. En todo caso su atención requiere de una capacidad de articulación, un poder de convocatoria, una capacidad de transferencia y un nivel de autoridad que consideramos debería aportar el INEE.

Aspectos psicométricos

Las tres pruebas analizadas presentan diferentes aproximaciones al diseño y concepción de los aspectos psicométricos. Evidentemente, por tratarse de instrumentos de alcance nacional, aunque con distintos propósitos, hay tópicos comunes, como la validez y la confiabilidad, pero cada institución desarrolladora acopió las evidencias con criterios propios y diseños diversos, justificados de maneras igualmente diversas, a fin de atender los propósitos, al menos iniciales, de las pruebas.

La solidez psicométrica de las pruebas debe tomarse con gran responsabilidad y seriedad por la institución o agencia que tenga a su cargo el diseño de un proyecto de evaluación educativa. El compromiso de cumplir con estándares psicométricos es de la mayor importancia cuando las pruebas son de interés nacional, tanto de alto como de bajo impacto, porque las implicaciones se reflejan en importantes consecuencias para los involucrados en el proceso, para la población en general o para estratos específicos, regionales, étnicos, culturales, por género o por otras características.

Pudo constatarse en estos proyectos que la presión administrativa y logística por producir, aplicar y calificar las pruebas con una frecuencia, a nuestro juicio innecesaria e inadecuada, obligó a los responsables —principalmente en el caso de ENLACE y en menor medida en el caso de EXCALE— a atender los puntos más urgentes e inmediatos de implementación, para cumplir con los tiempos y las exigencias operativas, dejando para otra oportunidad el cumplimiento integral de los aspectos psicométricos y atendiendo de manera parcial, incompleta o deficiente, los principales requerimientos que garantizan la calidad de las pruebas.

Manuales técnicos de las pruebas

El Manual Técnico (MT) es indispensable para todo proyecto de evaluación pues documenta, entre otros elementos, el perfil, las especificaciones y las tablas de validez, los modelos utilizados para la calibración de ítems y pruebas, los estudios hechos para determinar el funcionamiento diferencial y la forma de recopilar evidencias en los análisis realizados por jueces.

Si bien no hay modelo estático para el MT, éste debe documentar con suficiente detalle los procesos seguidos en el diseño e implementación de una evaluación a gran escala. No se espera que contenga una recopilación de ideas generales sobre el proyecto de evaluación, sino que permita al público en general y a especialistas en particular, conocer la fundamentación del proyecto. El MT debe cubrir varios elementos consensuados en estándares internacionales que abarcan desde la descripción general del proyecto, la definición del perfil o propósito a evaluar, el diseño de los instrumentos y la forma de producir los reactivos, hasta las normativas de aplicación y calificación, y la forma de reportar e interpretar los resultados. También debe mostrar los detalles de los procesos seguidos para su implementación y, en ese sentido, servir de memoria para poder replicar las actividades y apoyar la capacitación del personal. Por su importancia, el MT debe acompañar la publicación de los resultados y estar disponible al público.

Los hallazgos de este rubro pueden sintetizarse diciendo que ENLACE-B mantuvo actualizado un manual técnico durante sus primeros tres años, incluyendo información sobre las actividades y resultados de las aplicaciones; sin embargo, quedó pendiente la construcción de materiales y estudios documentales, principalmente sobre el proceso y los modelos utilizados. Esta práctica fue diferente en ENLACE-MS, que produjo dos manuales técnicos (uno cada dos o tres años) sin actualizar toda la información, en unos casos aportando ideas sobre el diseño de pruebas en general, en ocasiones con documentación con fines de divulgación y otras con enfoque técnico. En EXCALE solo se dispuso de un primer manual técnico al inicio de proyecto, pero no fue actualizado con cada aplicación anual. Puesto que incluyó las especificaciones de diseño, sirvió más bien de marco de referencia al establecer criterios y elementos considerados pertinentes para documentar las pruebas, sugiriendo formas para demostrar evidencias de validez y de confiabilidad, además de incluir normativas logísticas y de aplicación, junto con líneas de trabajo y de investigación.

De lo anterior se desprenden recomendaciones particulares: la sistematización de contenidos, actualización y organización del manual técnico de cada prueba es una necesidad evidente y un área de oportunidad para mejorar los proyectos a corto plazo. El MT debe cubrir al menos: a) la descripción general del proyecto; b) el diseño de pruebas de cada área; c) el diseño de cuestionarios de contexto; d) las adaptaciones y consideraciones culturales de los instrumentos del estudio: e) la descripción de los operativos de campo; f) los procesos de control de calidad; g) los procesos de captura, codificación y validación de los datos; h) el procesamiento de datos; i) el diseño muestral y cálculo de ponderaciones; j) la selección de muestras y tasas de participación; k) la evaluación de la calidad de la muestra y varianza muestral; l) el escalamiento y validación de las pruebas cognitivas; m) el establecimiento de puntos de corte; n) la construcción de índices, escalamiento y validación de preguntas en el cuestionario de contexto; o) la preparación de las bases de datos; y p) la publicación de resultados. Algunos de los apartados mencionados se pueden agrupar, pero es importante que se trate cada uno de los temas presentados en esta lista.

Desarrollo y construcción de las pruebas

Las pruebas en su modalidad censal (ENLACE) y muestral (EXCALE) y su diseño —versiones completas por asignaturas o modelo matricial de bloques incompletos—, se organizan en pruebas operativas y piloto. Para garantizar que se satisfacen los estándares de diseño, es primordial justificarlo y documentarlo, así como verificarlo con resultados de las aplicaciones, por medio de la calibración de las pruebas y la equivalencia de las versiones, la adecuación de la escala con los ítems de anclaje, el análisis de funcionamiento diferencial de los ítems y otros estudios.

La definición de las tablas de validez de contenido es heterogénea. En ENLACE-B se detalla el número de reactivos por cada subtema; en ENLACE-MS se definen por combinación de subtemas y tipos de aprendizaje, mientras que en EXCALE, se cuenta con una descripción con altísimo nivel de detalle de las especificaciones de los reactivos, que seguramente podría reducirse.

Un acierto de ENLACE-B es usar un modelo de pre-test muestral matricial para experimentar con más contenidos y competencias. Los reactivos del pre-test se calibran con el fin de obtener valores

de referencia para diseñar las pruebas operativas. No obstante, la información sobre el procedimiento para el diseño de las pruebas debe actualizarse; también requiere justificarse la formulación del modelo de muestreo (estratificado y por conglomerados), y describirse los criterios de elección de escuelas y estudiantes del Distrito Federal y la zona conurbada del Estado de México, que se justifica por costo y facilidad logística, pero no por representatividad de diversos grupos culturales.

Asimismo, falta documentar el proceso de construcción de las pruebas a partir de los bancos de reactivos, con información del sistema (computarizado o manual) que justifique las cualidades métricas de la prueba, así como corregir la práctica de construcción que se basa en el criterio de un diseñador que se sigue en ENLACE-B. El MT debe informar sobre las calibraciones de las preguntas utilizadas en el pre-test y sobre la calidad de las que pasan a formar parte de la prueba operativa en la siguiente aplicación. Esta retroalimentación ayudaría a refinar la elaboración de preguntas por quienes diseñen futuras aplicaciones.

Para ENLACE-B se llevó a cabo un esquema de "equiparación vertical" combinado con formas de pre-test (denominadas beta y gama); este mismo planteamiento fue seguido en ENLACE-MS. La explicación del modelo es clara en el MT de ENLACE-B, aunque sin evidencias suficientes sobre los resultados, valores experimentales y conclusiones del proceso de equiparación. De hecho, las justificaciones con base en correlaciones o gráficas son incompletas o inapropiadas, porque pruebas de distintas áreas pueden correlacionar sin implicar que miden lo mismo. Igualmente dudosas son las correlaciones con estudiantes de grados superiores para analizar los ítems y la calidad de la prueba, sin la justificación que se esperaría por tratarse de poblaciones focales diferentes.

La correlación entre pruebas es evidencia necesaria pero no suficiente para justificar una equiparación entre pruebas o su dimensionalidad. En general, se observa una correlación bastante alta entre pruebas de rendimiento, por tratarse de una medida combinada de los contenidos a evaluar y la capacidad de los estudiantes de leer y comprender lo que se les pide en la prueba. Para justificar una equiparación debe contarse con evidencia estadística, en paralelo a la comparabilidad del marco teórico a medir, y sobre el que se hacen las inferencias.

EXCALE dispone de un análisis reticular del currículo acompañado de los criterios para las especificaciones y la organización de las tablas de validez. El MT describe las fases de producción de las pruebas y el modelo muestral para las aplicaciones piloto en algunos estados del país. Se esperaría contar con mayor documentación y respaldo de los cambios del modelo muestral, a lo largo del tiempo. Sería útil justificar los cambios en el diseño muestral con un análisis de error de muestreo, precisión y otras consideraciones, al igual que presentar un análisis de las ventajas y desventajas de los nuevos procedimientos que se establezcan. De particular importancia es el efecto que pudiesen tener los cambios del proceso de muestreo en los resultados de la evaluación y la medición de cambios a través del tiempo.

Todas las pruebas utilizan reactivos de anclaje. EXCALE usa un bloque para centrar la escala de la prueba con ayuda del software de calificación con el modelo de Rasch, sin embargo falta detallar el proceso de anclaje y ajuste de la escala. En el caso de ENLACE-B la justificación es menos clara, sobre todo al dejar tácito el uso del anclaje y su nexo con el modelo de la Teoría de Respuesta al Ítem (TRI).

El armado de las pruebas presenta el reto de aplicar en un tiempo razonable una gran cantidad de preguntas que cubran los contenidos y habilidades a medir y sobre los que se quiere reportar resultados. Al reto se añade la necesidad de aprovechar la aplicación definitiva para probar preguntas

que se utilizarían en futuras aplicaciones, evitando operativos adicionales. El plan de aplicación debe considerar que el tiempo disponible en cada escuela y de cada alumno es limitado.

Una solución posible es utilizar diseños matriciales de muestreo múltiple, administrando a cada alumno un subconjunto de preguntas; con ello cada pregunta es respondida por un subconjunto de estudiantes. En otras palabras, nadie responde todas las preguntas y ninguna pregunta es respondida por todos. Estos diseños, bien concebidos, son muy eficientes para obtener información acerca de los alumnos y las preguntas, así como para explorar efectos de posición y contexto de las preguntas o de fatiga de los estudiantes. Los diseños matriciales forman parte del mundo de la evaluación educativa y se recomienda su utilización, ya que permiten incrementar la cantidad de información evaluada y contribuyen a la validez de la medición a través de la representatividad de los contenidos evaluados. El uso de las técnicas modernas de escalamiento, como la TRI, ya utilizadas en los programas evaluados, permiten trabajar con este tipo de diseños.

Los puntos fuertes del diseño de estas pruebas se centran en la pertinencia de los modelos matriciales y su combinación con las pruebas operativas temáticas. Las pruebas cuentan con una base de tablas de especificaciones que deberá mejorarse para reducir la heterogeneidad de presentación entre pruebas, e incluso dentro de ellas. Un acierto de ENLACE-B es el uso del modelo de pre-test muestral matricial, para experimentar con más contenidos y competencias, que amerita contar con mayor detalle y justificación metodológica. Todos los proyectos disponen de un sistema informatizado para administrar los bancos de reactivos, pero su descripción es incompleta y no logra destacar sus posibles ventajas y operatividad. Es necesario que los manuales técnicos informen sobre las calibraciones de las preguntas utilizadas en el pre-test y sobre la calidad de las que pasan a formar parte de la prueba operativa. EXCALE dispone de un análisis reticular del currículo que se acompaña de los criterios para las especificaciones y la organización de las tablas de validez, lo cual puede ser una opción de mejora para ENLACE-MS. Aunque no se cuenta con documentación suficiente para juzgar la calidad de los resultados y la estabilidad que aportan a la prueba operativa, es importante que las pruebas continúen con la práctica de utilizar reactivos de anclaje, en forma aislada o con un bloque específico del diseño matricial.

Se recomienda homogeneizar el diseño de las tablas de especificaciones, con descripciones en términos de competencias y niveles de complejidad taxonómica, que permitan hacer el seguimiento longitudinal de la calidad del sistema educativo como proyecto de gran visión. Se debe documentar tanto el proceso de construcción de pruebas como la información con la que se cuenta en el sistema administrativo de bancos de reactivos, justificando las cualidades métricas de las pruebas producidas. Deben detallarse los modelos de equiparación entre versiones de las pruebas, aportando evidencias experimentales debidamente documentadas. Se debe dar cuenta de la metodología de anclaje entre pruebas y su función en el ajuste de la escala, ofreciendo evidencias de los resultados. Deberán realizarse estudios que muestren la dimensionalidad de los bloques matriciales y el diseño de las pruebas operativas, detallando las investigaciones que se hagan sobre posición de los reactivos y fatiga de los alumnos, entre otros elementos.

Calibración y análisis psicométrico de las pruebas

ENLACE y EXCALE deben fundamentarse en datos objetivos, con respaldo estadístico que soporte cualquier auditoría técnica, la cual, a su vez, sirve de respaldo a los miembros de la sociedad que estén legítimamente interesados en obtener conclusiones constructivas a partir de los resultados de las pruebas. Es evidente que no se pueden obtener tales conclusiones si no se cuenta con instrumentos sólidamente construidos.

Para reportar los resultados de las pruebas es aceptable la Teoría Clásica de los Tests (TCT), aunque dada la estructura, diseños y requerimientos de las mismas son deseables otros modelos, en general más flexibles y poderosos, basados en la TRI. Se requieren estudios descriptivos del comportamiento de los ítems que integran las pruebas, así como del de los sustentantes, pues ello permite valorar la pertinencia de la prueba para medir de forma objetiva (y justa) a diferentes grupos socioeconómicos, regionales, étnicos y con discapacidades.

Las pruebas ENLACE utilizan modelos de tres parámetros de la TRI; EXCALE utiliza el modelo de Rasch combinado con el cálculo de valores plausibles. Son enfoques justificados por el ámbito de interés de cada prueba (censal en ENLACE, muestral en EXCALE), pero hay implicaciones que no parecen apropiadas al combinarse con modelos de la TCT sin correspondencia entre los valores de las teorías empleadas. Esto es muy evidente en el caso de los intervalos de aceptación para los parámetros de diseño de la prueba, que no corresponden entre TCT y la TRI, lo cual revela una falta de integración de los elementos de las teorías involucradas o un intento injustificado de emplear todos los modelos disponibles para cubrir los aspectos de diseño de las pruebas.

En EXCALE se presentan resultados de estudios de sesgo y funcionamiento diferencial de las pruebas, que no se tienen en ENLACE. EXCALE no reporta de forma clara el uso de los valores plausibles y la manera de utilizarlos en los trabajos de análisis y dictamen de la calidad de la prueba para describir a la población evaluada. Los estudios de seguimiento longitudinal que se hacen en ENLACE no están suficientemente respaldados por documentos técnicos.

Igualmente hay limitaciones en los estudios de calidad de la prueba atendiendo a factores diversos que pueden influir en las respuestas de los estudiantes. Por ejemplo, debe justificarse detalladamente el tiempo de respuesta y su relación con factores como fatiga, uso de términos potencialmente regionales y velocidad de lectura. Ninguna de las pruebas cuenta con normativas para las adaptaciones a alumnos con algún tipo de discapacidad, lo cual es una omisión importante.

Las deficiencias encontradas permiten identificar áreas de oportunidad para las pruebas, en particular, relacionadas con las normativas de aplicación y de interpretación o uso de los resultados. Las pruebas deben establecer las especificaciones de diseño y análisis de calidad, coordinando apropiadamente los modelos clásicos y logísticos, para contar con el conjunto de características y parámetros de diseño que permitan juzgar el comportamiento de la prueba y ubicar los estratos poblacionales evaluados dentro de la escala. Las explicaciones y reportes deben ser satisfactorios para los distintos públicos, contar con las bases de datos de los valores plausibles y la documentación de su construcción y uso; también se esperan tablas con los valores estadísticos del instrumento y su comparación contra la población focal de cada nivel respecto del marco teórico del perfil académico referido a criterio.

Es recomendable continuar utilizando modelos de la TRI que permiten expresar los resultados de los estudiantes en escalas independientes de la muestra en la que se aplicó la prueba. La gran ventaja de utilizar estos modelos es que brindan información acerca de lo que los alumnos saben y pueden hacer en términos probabilísticos. Por ejemplo, dado el contexto escolar en el que se espera un dominio de contenido específico por parte de los alumnos, la TCT reporta tanto la proporción que responde cada pregunta de manera correcta como la cantidad de preguntas que responde cada alumno acertadamente. En el contexto de la TRI, se expresan los resultados en una escala que permite calcular la probabilidad de responder correctamente a una pregunta de determinada dificultad. Esto se convierte en información útil para fundamentar las decisiones de política educativa, al calcular la probabilidad de que los alumnos cumplan con criterios de desempeño establecidos por expertos.

Otra ventaja de los modelos de la TRI es que pueden ser combinados con otros de estimados poblacionales que resultan en el cálculo de valores plausibles; éstos, ayudan a corregir sesgos que se observan en la comparación de resultados grupales, al igual que a obtener estimadores del error de medida observado en los resultados de cada estudiante. Es importante señalar que los valores plausibles no deben ser utilizados para reportar resultados individuales.

Hay pocos estudios de impacto de factores asociados. Solo EXCALE cuenta con estudios de sesgo y análisis de funcionamiento diferencial. Las bases de datos con valores plausibles de EXCALE no parecen estar utilizándose con fines de investigación. Ninguna prueba cuenta con normatividad para adaptaciones para alumnos con algún tipo de discapacidad, lo cual es una omisión importante.

Se recomienda homologar los criterios de los modelos de la TCT y del TRI utilizados en las pruebas, para eliminar contradicciones en el dictamen de calidad que se aplica con cada modelo y evitar se asuma que la calidad de la prueba está garantizada por el hecho de usar un cierto modelo o un programa de calificación en particular. En cada aplicación deben sistematizarse los estudios de sesgo y funcionamiento diferencial de las pruebas y los ítems, para diversos estratos socioeconómicos y otros factores contextuales. Es imprescindible contar con normativas para las adaptaciones de las pruebas a distintos grupos socioculturales y alumnos con discapacidades, evitando que los aplicadores las definan de manera personal no controlada desde el diseño. Deben sistematizarse los reportes estadísticos de los instrumentos y de los grupos de la población focal en cada nivel evaluado.

Confiabilidad de las pruebas

Las pruebas deben demostrar niveles aceptables de precisión, como un requisito necesario, pero no suficiente. La calidad de una prueba se respalda sobre todo con evidencias de validez y niveles aceptables de precisión, que se traducen en un error de medida asociado con la prueba para medir los constructos previstos. A menor error de medida se tiene mayor precisión. Pruebas con idéntica confiabilidad pueden tener diferente error de medida, en función de dos grupos de factores asociados con: 1) la cantidad, calidad y distribución los ítems, en la escala de medida; y 2) la dispersión (varianza) de medidas de los estudiantes.

A partir del error de medida se determina el intervalo de confianza para la puntuación de cada persona, para la media de todos los estudiantes o los puntos de corte que separan niveles de desempeño. Así, para conocer la precisión de la prueba (especialmente si es de alto impacto) es requisito indispensable reportar el valor del coeficiente de confiabilidad y, sobre todo, el error de medida del instrumento completo y de sus subescalas (temas, áreas o componentes). Si bien es factible obtener los intervalos de confianza a partir del error de medida y de la varianza de las personas, es deseable que éstos se reporten. El coeficiente de confiabilidad es un número adimensional que se interpreta de forma general (independiente de cada prueba); es común considerar que una confiabilidad de 0.8 es aceptable. Este coeficiente debe tomarse con reservas porque no informa de manera suficiente sobre la calidad de una prueba específica. El error de medida se expresa en unidades (número de ítems, porcentaje de aciertos, unidades estandarizadas o logísticas) de la escala de medida de la prueba específica; constituye un valor propio de la prueba que lo hace de mayor interés.

Los modelos para reportar confiabilidad incluyen el que se basa en correlaciones entre mitades y fórmulas de consistencia interna (como Alfa de Cronbach) y los que se asientan en modelos logísticos y la teoría G. El error de medida puede calcularse con el modelo clásico que lo considera único y fijo para cada forma de la prueba, o con TRI que lo reporta para cada ítem y persona. Con modelos de TRI, la confiabilidad es propia de los ítems aplicados a un individuo. Los modelos logísticos proporcionan, además, índices de separación, al igual que la cantidad de información a lo largo del continuo de medidas de habilidades, convertible al error de medida relacionado con la medición en un punto específico de la distribución.

El concepto de confiabilidad y del error de medida de la TCT ha evolucionado con el desarrollo de la TRI, pues de ser un concepto único y específico para una forma aplicada a una muestra particular de la población, ha pasado a uno más general. Bajo la TCT, la confiabilidad es un número único que aplica por igual a la prueba y a todos los estudiantes que la toman, independientemente de dónde se encuentren en la distribución de habilidades y conocimientos. Con los principios de la TRI, no desaparece el concepto de confiabilidad única de la prueba, pero pasa a ser menos importante, pues es posible calcular el error de medida como una función del conjunto de preguntas en la prueba, y la habilidad del estudiante que la toma. Es decir, se conceptualiza que la precisión de la medida varía dependiendo de dónde se encuentre el estudiante, y el tipo de preguntas que se le aplican. En este sentido, incrementar el número de preguntas no necesariamente ayuda a disminuir el error de la medición, a menos que éstas aumenten la información proporcionada. Utilizando estos principios —y suponiendo que el banco cuenta con suficientes reactivos entre los cuales escoger—, se pueden armar pruebas que optimicen la medida en ciertos puntos de la distribución de habilidades y conocimientos; por ejemplo, medir los conocimientos de los alumnos en los niveles bajos de rendimiento, al igual que conformar pruebas que optimicen la medición alrededor de los puntos de corte para tener mayor certeza en la clasificación que se hace de los alumnos según niveles de rendimiento.

Complementariamente existen fórmulas sugeridas para determinar la confiabilidad de las decisiones que se toman respecto de los puntos de corte para pruebas referidas a criterio, como es el caso de las analizadas en este trabajo o su equivalente en términos de la función de información dentro de la TRI. Junto con esta información se reporta el error de medida, único y homogéneo en el caso de la TCT y punto a punto, tratándose de la TRI. No es suficiente tener valores de alfa superiores a algún número definido en forma subjetiva o como dato tradicional

(es muy común postular como intervalo aceptable de 0.8 a 0.95), sino que debe haber una justificación metodológica que permita reconocer los valores esperados para la prueba en su conjunto o para cada sección o subescala en particular.

El reporte de confiabilidad cuenta con poco detalle en los manuales técnicos y en los reportes. Las tres pruebas disponen de resultados del alfa de Cronbach, y obtienen valores dentro de los rangos esperados, pero no debidos necesariamente a la calidad del diseño, sino posiblemente por el número de ítems. No se reporta la confiabilidad con los modelos logísticos, ni tampoco se tiene evidencia de que se hayan hecho estudios de generalizabilidad (teoría G), especialmente importantes para el grupo de validez cultural que se reporta en el siguiente apartado. En ningún caso se cuenta con valores de confiabilidad criterial en los puntos de corte, ni con estudios piloto o experimentales que correlacionen aplicaciones sucesivas. En la documentación evaluada se presentan gráficas esporádicas de la función de información, pero no se utilizan de forma sistemática ni se dan interpretaciones apropiadas para su uso en los puntos de corte. No se presentó evidencia de uso de la función de información para seleccionar las preguntas que conforman la prueba, ni para seleccionar y validar los puntos de corte.

Ninguna de las tres pruebas reporta de forma sistemática el error de medida (global o puntual); no lo utilizan al emitir las calificaciones finales de los estudiantes, ni interpretan las implicaciones del error en la definición de los niveles de desempeño. El error de medida es imprescindible para revisar la distribución de ítems en cada subescala, y reforzar las decisiones acerca de la ubicación de los sustentantes en los niveles de desempeño ante un escaso número de reactivos.

En el futuro, será conveniente respaldar las pruebas con estudios exhaustivos de confiabilidad general y de cada sección o subescala. Igualmente, en los casos de reactivos de grupo —v. gr. una lectura de la cual se derivan varios ítems "hijos"—deberán hacerse los análisis de confiabilidad e interdependencia de los ítems.

Los modelos de valores plausibles no solo permiten obtener una medida adicional de la confiabilidad de la prueba, sino también incorporar la información de los cuestionarios de contexto de los alumnos en el cálculo de sus puntajes, para posteriormente calcular el error de la medición para los grupos que aquellos representan. De esta manera se disminuyen los sesgos que pudiera haber en las comparaciones entre grupos o estratos por razones de regresión hacia la media, producidos por la presencia del error de la medición. En este sentido es deseable y recomendable que las pruebas analizadas continúen utilizando modelos de la TRI en el armado de las pruebas y en el cálculo de los puntajes, al igual que modelos de valores plausibles al calcular los puntajes de los alumnos a fin de reportar resultados grupales.

Como hallazgos a destacar tenemos que se ofrece poco detalle sobre la confiabilidad de las pruebas y las descripciones son superficiales. En los manuales técnicos no se reportan de manera sistemática el error de medida, la confiabilidad criterial o el uso de la función de información de la TRI, entre otros aspectos que inciden en el dictamen de calidad de las pruebas y de los puntos de corte para los niveles de desempeño. Es pertinente el uso de los valores plausibles en EXCALE, pero falta mayor detalle para facilitar el aprovechamiento de las bases de datos, así como explicación sobre las formas de interpretarlos.

Se recomienda realizar estudios con TCT o TRI de la confiabilidad (normativa y criterial) de cada prueba desde el punto de vista global, por subescala (tema o área) y por grupos de reactivos, combinados con estudios de dimensionalidad de cada parte citada y las posibles dependencias funcionales entre ítems. Debe detallarse y reportarse el cálculo del error de medida por escala, subescala y puntual, especialmente en los puntos de corte. En caso de usar la función de información para determinar el error de medida, se deberá especificar su interpretación para los potenciales usuarios. Se recomienda el uso exhaustivo de modelos de la TRI al armar las pruebas y al calcular los puntajes de los estudiantes, junto con valores plausibles en el caso de las pruebas muestrales, detallando las bases de datos y la forma de utilizarlas.

Calificación en niveles de desempeño e interpretación de resultados

Cada prueba debe justificar el modelo de calificación de los estudiantes y la escala para reportar los resultados. En las pruebas referidas a criterio se debe asociar la calificación con desempeños establecidos por puntos de corte, con un modelo definido de antemano, el cual debe contar con una interpretación para resultados inconvenientes o inesperados y explicaciones acerca de la forma de cálculo y manejo de los niveles de desempeño. Esto no implica que una persona pueda reproducir los cálculos pues necesitaría contar con los modelos matemáticos, el sistema de cómputo, las claves de respuestas correctas y la función (o tabla) que brinde la equivalencia entre puntaje bruto y medida o calificación final.

Para EXCALE se utiliza el modelo de Rasch con el que puede obtenerse la curva característica de la prueba, que relaciona el puntaje bruto con las medidas siempre y cuando a los alumnos se les apliquen las mismas preguntas. Esta curva no puede construirse con los modelos de 2 o 3 parámetros, debiendo utilizarse para ello programas especializados que estiman la medida de cada persona con modelos de máxima verosimilitud, en función de cuántos y cuáles ítems responde cada estudiante, de tal modo que dos estudiantes que tienen un mismo número de respuestas correctas en una prueba, pudiesen terminar con distintos puntajes. Si bien esto puede parecer un poco injusto, es algo común en otros campos en los que el "puntaje" asignado no solo se basa en la cantidad de las preguntas respondidas, sino también en la dificultad de las preguntas. Esta diferencia entre los modelos de Rasch —y de manera más general los modelos de la TRI— tiene implicaciones en la calificación y la emisión de reportes par los estudiantes.

Por tratarse de una prueba muestral, EXCALE no tiene el propósito de reportar calificaciones para cada estudiante, sino a nivel de grupos de interés en la población. Para ello se adoptó la metodología de los valores plausibles que reportan para cada estudiante un conjunto de calificaciones que simulan la distribución representativa de su participación dentro de la población evaluada.

Las pruebas ENLACE utilizan modelos de la TRI de tres parámetros, dado que todos los alumnos responden a las mismas preguntas. Los puntajes individuales se asignan *a posteriori*, utilizando métodos de estimación de verosimilitud máxima marginal. Dado su propósito de reportar resultados a nivel individual, las pruebas ENLACE-B hacen análisis de copia utilizando dos modelos computarizados, práctica que no se sigue en las otras pruebas analizadas.

Tanto la emisión de calificaciones en la escala (factor de escala y sumando), como la combinación con los ítems de anclaje y otros modelos que permiten la igualación de versiones, descansan en las cualidades de los programas de calificación, lo cual deberá documentarse a futuro. Igualmente debe establecerse el modelo de cálculo para asociar calificaciones con los niveles de desempeño, donde intervenga el intervalo de error de medida o correcciones en función de la confiabilidad o precisión de los puntos de corte. En el caso de ENLACE-B se usan valores establecidos en las primeras aplicaciones, que no son corroborados en las subsiguientes.

Respecto del origen de los puntajes, deben reforzarse los análisis de copia e investigaciones de prácticas fraudulentas en cada aplicación, emitiendo reportes informativos pertinentes que, por un lado, muestren a los usuarios que efectivamente pueden detectarse dichas prácticas y, por otro, brinden sugerencias y recomendaciones para evitarlas en el futuro. Estos reportes deben ser indicativos para las autoridades nacionales o locales, independientemente de otras decisiones o sanciones administrativas que se adopten (como invalidar el aula o la sede).

Son varias las consideraciones acerca del cálculo de los puntajes individuales. La primera se refiere a que, si bien se reconoce el esfuerzo que se hace para construir pruebas que midan de la mejor manera el propósito previsto, es ingenuo pensar que el resultado de aplicarlas no tiene error. Por ejemplo, en una prueba de 40 preguntas, un estudiante contesta correctamente 25 y otro contesta 24. Más allá de afirmar que uno contestó una pregunta más que el otro, convendría analizar la relevancia de la pregunta que hace la diferencia.

Vale la pena preguntarse cuáles son las diferencias de dificultad, calidad y contenido en las preguntas contestadas correctamente y si dichas diferencias son significativas en el estimador de lo que los alumnos saben o pueden hacer. Con los modelos de la TRI se puede calcular el puntaje estimado de la habilidad más probable que generó el patrón de respuestas observado; este cálculo se realiza con la mayor precisión permitida por el programa de cómputo utilizado, que puede arrojar varios decimales que resultan insignificantes al estar dentro del margen de error, por lo que es recomendable redondear estos resultados a un número que no sobreinterprete la precisión obtenida. Por ejemplo, pruebas reconocidas de los Estados Unidos se reportan en una escala que va de 200 a 800 puntos, con puntajes de 10 en 10, lo que significa que solo hay 61 puntajes posibles (200, 210, etcétera), al grado que podrían hacerse coincidir con igual número de preguntas que cada alumno debería responder, dentro de un mismo rango de dificultad. Pero sería ingenuo pensar que el continuo de conocimientos o habilidades tiene solo 61 categorías posibles. Para llenar los espacios intermedios, se utilizan los valores plausibles que calculan aleatoriamente puntajes de la distribución de posibles habilidades que habría podido generar el patrón de respuesta observado.

Estas situaciones llevan a la segunda consideración de importancia: el proceso de rellenar los espacios entre puntajes contradice la primera consideración, pero ilustra precisamente el conflicto que existe entre asignar puntajes óptimos para describir individuos y óptimos para describir grupos de individuos.

La tercera consideración se refiere a tomar en cuenta el error de la medición. Si se asume la confiabilidad como la proporción de la varianza total que es varianza verdadera, se ve que los puntajes obtenidos tienen una regresión hacia la media de magnitud del error de la medición. Una solución a esta restricción de rango es añadir varianza de magnitud igual al error de la medición a fin de que la varianza de los puntajes refleje la varianza verdadera, lo que es deseable si el propósito es estudiar la distribución de las habilidades y conocimientos en

la población, y las diferencias entre grupos o estratos. Pero de nuevo, este principio colisiona con el de asignar puntajes individuales.

A partir de estas tres consideraciones puede apreciarse la divergencia que se tiene al calcular los puntajes dependiendo de la finalidad del reporte: individual para cada estudiante o grupal para los estratos sociales a los que estos mismos individuos pertenecen. En el primer caso son útiles los puntajes redondeados procedentes de estimados *a posteriori;* en el segundo, los valores plausibles son óptimos. Conviene hacer notar que estos valores se obtienen seleccionando al azar una distribución de posibles puntajes, a fin de rellenar los espacios entre los puntajes observados y contrarrestar la regresión hacia la media. Como no se obtiene uno sino varios puntajes posibles para cada individuo, es imposible su uso para reportar resultados individuales.

Un puntaje en una prueba informa sobre el desempeño de los estudiantes que la toman y, si se usa la TRI, estima la probabilidad de que una persona responda correctamente a preguntas de dificultad conocida. Los puntajes se pueden convertir en información más útil en términos educativos o pedagógicos si se categorizan en niveles de rendimiento o de desempeño.

Una vez seleccionados los puntos de corte se procede a detallar cada nivel en términos de lo que los alumnos ubicados en ese nivel saben o pueden hacer. La descripción se hace clasificando las preguntas por niveles y resumiendo los comportamientos y conocimientos que cubren. En algunos proyectos de evaluación se opta por escoger los puntos de corte puramente con base en criterios estadísticos, y tomando percentiles de la población o puntajes en referencia a la distribución de los resultados. Por ejemplo, en el caso del TIMSS —prueba internacional de matemáticas y ciencias en la que no hay un referente curricular único— se toman los percentiles 25, 50, 75 y 90. En el caso de PIAAC —prueba de alfabetismo de lectura y numérico en el que no existe referente curricular alguno— se establecen niveles que abarcan una desviación estándar de la distribución de puntajes, y se les define. Utilizar la desviación estándar para establecer los puntos de corte es análogo a elegir percentiles en la población, ya que aquella medida corresponde a percentiles en la distribución normal correspondiente.

Acordados los puntos de corte y definidos y descritos los niveles, se clasifican los estudiantes, pudiéndose etiquetar o describir su desempeño con base en el nivel alcanzado. Esto facilita la interpretación y utilidad de los datos ya que el desempeño de los alumnos no se describe en puntajes brutos, sino en términos de haber alcanzado o no niveles deseados de desempeño.

Las pruebas utilizan diferentes métodos para definir los niveles de desempeño. En ENLACE-B se utiliza un método desconocido en la literatura. ENLACE-MS y EXCALE utilizan básicamente el de *Bookmark*. Más allá del método, importa que se valide constantemente el procedimiento y la descripción de puntos de corte y niveles. Para esas validaciones se recomienda que:

- Los niveles de desempeño sean distinguibles, mostrando diferencias que puedan describirse, por lo que no conviene utilizar más niveles de corte de los que sean interpretables. Generalmente, cuatro niveles son suficientes para clasificar a los alumnos (por ejemplo: deficiente, básico, competente y avanzado).
- La categorización en niveles de desempeño tenga relevancia educativa y provea información útil a los usuarios. Por ello, es importante validar las descripciones de los niveles y actualizarlas constantemente a fin de que contengan información relacionada con las exigencias curriculares.

- Los niveles sean apropiados a lo que se puede esperar en forma realista, dadas las exigencias curriculares y las expectativas de la población. No es raro que en un primer momento se establezcan niveles inalcanzables por los alumnos.
- Puesto que con las pruebas se pretende medir tendencias, es importante validar la relevancia y el significado de cada uno de los niveles a lo largo del tiempo, dados los cambios en el marco de referencia, y los posibles cambios curriculares que sean implementados e incorporados en las pruebas.
- Se debe tener en cuenta el error de clasificación, en función de dos errores que pueden tener un efecto aditivo: uno relacionado con la elección de un punto de corte y otro con la asignación del puntaje. Ambos errores pueden cancelarse o sumarse, por lo que deben considerarse, sobre todo al reportar resultados individuales.

Los principales hallazgos incluyen que los modelos de calificación son diferentes: TRI de 3 parámetros en ENLACE y Rasch con valores plausibles en EXCALE. No se aprecia ningún intento de homologar los modelos, lo cual es deseable para que los usuarios sepan cómo utilizar los resultados. Las pruebas cuentan con modelos diferentes para definir los niveles de desempeño; que EXCALE y ENLACE-MS utilicen el método Bookmark, puede ser un área de oportunidad para su revisión y uso en otras pruebas, o para hacer investigación sobre las que se aplican a nivel nacional tomando en cuenta el error de medida y la confiabilidad criterial en los puntos de corte. También puede ayudar a mejorar los criterios descriptivos frente a las especificaciones de cada prueba y del currículum educativo. En ENLACE-B se usan valores fijos, establecidos en las primeras aplicaciones, lo cual es criticable porque al cambiar las especificaciones deben verificarse los parámetros de la escala. Solo ENLACE-B realiza análisis de copia, con un procedimiento apenas esbozado en el Manual Técnico, pero aparentemente sin consecuencias administrativas.

Se recomienda documentar la forma en que se utilizan los valores plausibles en las pruebas, para que sea clara su inclusión en las bases de datos y su interpretación para los usuarios. Debe documentarse el algoritmo o forma de obtener la calificación de los estudiantes al hacer intervenir los ítems de anclaje, junto con la ubicación en los niveles de desempeño y las posibles correcciones tomando en cuenta el error de medida en los puntos de corte. En cada aplicación deberá sistematizarse el análisis de patrones de respuesta para dictaminar la presencia de prácticas fraudulentas (incluyendo copia y dictado de respuestas), así como para producir reportes con fines informativos desde el punto de vista académico y administrativo. Deberá revisarse y validarse el procedimiento para establecer los niveles de desempeño con sus descripciones y la determinación de los puntos de corte, tomando en cuenta que no es aceptable mantenerlos fijos si las especificaciones de las pruebas cambian con el currículo y otras consideraciones en el tiempo.

Bancos de reactivos y calidad psicométrica de los ítems

La materia prima para la construcción de las pruebas son los ítems; éstos son diseñados siguiendo especificaciones claramente definidas y calibrados con un conjunto de criterios de dictamen de su calidad para aceptarlos, modificarlos o eliminarlos, según sea el caso. Para cada prueba

debe demostrarse que los ítems miden con la mayor precisión posible lo que se pretende medir. La calidad de la medición, sin embargo, no reside solo en parámetros estadísticos de los ítems, sino también, y quizá más fundamentalmente, en la representatividad de los contenidos y habilidades a medir con la prueba.

Los modelos de calibración de reactivos disponibles en la literatura cubren desde la teoría clásica de los tests hasta modelos de la TRI, siendo común una formulación en la que ambos métodos son utilizados para emitir reportes enfocados a diversos usuarios. Los criterios de aceptación de los ítems deben ser exhaustivos y estar debidamente fundamentados para justificar que los criterios se mantengan estables o requieren ser modificados o actualizados.

Para la calibración de los reactivos, las pruebas ENLACE utilizan el modelo de tres parámetros de la TRI mientras que EXCALE sigue el modelo de Rasch; en ambas se complementan los análisis con TCT. Pero es de observar que los criterios no son homogéneos entre modelos; por ejemplo, se privilegia el uso de la correlación punto-biserial como criterio de calidad de los ítems, inclusive en casos en que el ajuste al modelo logístico es deficiente. ENLACE-B contó con un juego de valores bastante comprensivo en las primeras ediciones del Manual Técnico, pero éstos ya no se utilizaron en las siguientes aplicaciones o no se documentaron debidamente. No se cuenta con una explicación que justifique se siguieran criterios diferentes en ENLACE-MS. Los criterios seguidos por EXCALE difieren entre reportes y no se cuenta con un documento conductor para las aplicaciones a lo largo del tiempo.

Es recomendable mantener estables o constantes los criterios de selección y aceptación de los reactivos, a fin de asegurar la comparabilidad entre las aplicaciones o, por lo menos, controlar la calidad de los instrumentos a lo largo del tiempo. Los cambios en los criterios de selección de las preguntas, aunque pudieran catalogarse como menores, tienen efectos acumulativos en las pruebas que podrían afectar los resultados de forma y en magnitud desconocidas.

Los análisis de funcionamiento diferencial de los ítems (DIF) son obligatorios en estas pruebas dado que tienen implicaciones a nivel regional y nacional; por ello deben sistematizarse los estudios en cada aplicación anual, explorar patrones de respuesta y localizar las fuentes de variación en el funcionamiento de los ítems considerando los estratos específicos. También debe reportarse la evolución de los parámetros psicométricos de los ítems, para determinar su estabilidad comparando su funcionamiento entre las aplicaciones muestrales matriciales y las de las pruebas operativas.

Aunque el análisis DIF penas comienza a implementarse en EXCALE (que privilegia el estudio de sesgo de la prueba), éste por lo menos ofrece un reporte con un procedimiento emitido por el software de análisis. EXCALE también cuenta con estudios puntuales de DIF, más cercanos al análisis cualitativo, como es el caso del trabajo realizado sobre una población focal para español y lengua maya.

Ninguna de las modalidades de ENLACE-B o MS cuenta con evidencias de estudios sistemáticos de DIF, ni produce explicaciones con base en modelos estadísticos o matemáticos acerca de diferencias en los patrones de respuesta en los ítems por motivo de género, etnia, región, o cualquier otra clasificación. Tampoco hay estudios de DIF entre una aplicación y otra.

En el caso de EXCALE se cuenta con una normativa para la permanencia de los reactivos en los bancos. Pero, a ENLACE, la práctica de distribuir los cuadernillos una vez administradas las

pruebas le ocasiona volatilidad porque los reactivos tienen una permanencia limitada a una sola aplicación, salvo los que integran los módulos matriciales experimentales; lo anterior es un factor que dificulta o impide los estudios de funcionamiento de los ítems. Independientemente de ello, y de que pudiera parecer un trabajo inútil porque no se utiliza el mismo ítem en aplicaciones subsecuentes, se requiere de este tipo de estudios para verificar que las pruebas se comportan de acuerdo con lo planeado.

En ENLACE, los parámetros de las preguntas son estimados en la aplicación experimental y se utilizan sin modificación en la operativa, por lo que se recomienda verificar que las características de las preguntas no varién de una aplicación a la siguiente. Pueden producirse reactivos de tipo "espejo" o "clones", similares a los aplicados porque, bajo condiciones controladas de diseño, se comportan como los ítems que sirven de "patrón". De este modo, el trabajo de construcción y calibración de los ítems tiene mucho sentido y utilidad y debe aprovecharse en estas pruebas.

Además de contar con los ítems como materia prima, se requiere de un sistema de gestión de los bancos (computarizado, o manual) que facilite que los instrumentos: se construyan en apego a las tablas de especificaciones y criterios de validez de contenido; cuenten con los niveles de complejidad definidos y explicitados en los marcos de referencia; y sus ítems se distribuyan en la escala de la forma prevista por los parámetros de diseño. El sistema informatizado facilita la edición y revisión de los ítems y la producción de las pruebas.

Salvo los primeros manuales técnicos de ENLACE-B, no se reporta de forma sistemática el inventario de cada banco. Hay algunas tablas con datos de los reactivos de alguna aplicación, pero falta mayor información acerca de la distribución y cantidad de ítems disponibles. En todos los casos se cuenta con un sistema informático de gestión, pero sus características no están completamente detalladas.

Encontramos que no es suficiente la documentación sobre los modelos utilizados en las pruebas. No es un problema que, además del modelo clásico, se usen modelos de tres parámetros de la TRI en el caso de ENLACE y de Rasch, en el caso de EXCALE; el inconveniente radica en que los criterios empleados para juzgar la calidad de los ítems y de la prueba no son homogéneos o equiparables entre modelos. Se dispone de algunos análisis de sesgo y de funcionamiento diferencial de los ítems en el caso de EXCALE, con modelos cuantitativos y cualitativos. En cambio, no se cuenta con evidencias de que se haya realizado este tipo de análisis en ENLACE. Salvo los primeros manuales técnicos de ENLACE-B, no se reporta de forma sistemática el inventario de cada banco, excepto algunas tablas con datos de los reactivos usados en alguna aplicación. Falta mayor información acerca de la distribución y cantidad de ítems disponibles. Todas las pruebas cuentan con un sistema informático de gestión de los bancos de reactivos, pero no se presenta documentación detallada sobre ellos.

Se recomienda mantener estables o constantes los criterios de diseño, selección y aceptación de ítems para conservar la comparabilidad entre aplicaciones o, al menos, para controlar la calidad de los instrumentos a lo largo del tiempo. Dado que los ítems se liberan al final de cada aplicación, se sugiere producir reactivos tipo "espejo" o "clones", para las nuevas pruebas. En el caso de que se propongan cambios a los criterios citados, debe proporcionarse documentación suficiente que los justifiquen. Deben sistematizarse los estudios de DIF y de sesgo en cada aplicación anual, junto con los análisis de copia y prácticas fraudulentas para explorar patrones de respuesta y localizar las fuentes de variación en función de estratos poblacionales específicos. Debe reportarse la evolución de parámetros psicométricos de ítems para determinar su estabilidad comparando entre aplicaciones piloto y las de las pruebas operativas. Debe sistematizarse la producción de los inventarios de los bancos, mejorar las descripciones y clasificaciones taxonómicas. También deben documentarse los sistemas de gestión de los bancos de reactivos, para respaldar los procesos e indicar la potencialidad de edición, revisión y producción de las pruebas de acuerdo con las tablas de especificaciones.

Evidencias de validez

El proyecto de evaluación educativa debe mostrar que cuenta con suficientes evidencias sobre validez, recabadas de muy variadas formas y en todos los factores posibles. Conceptualmente es un tópico complejo en el que los autores no concuerdan completamente respecto a lo que debe interpretarse como validez, ni al conjunto de evidencias requeridas para que un proyecto esté debidamente fundamentado. Pero sí hay acuerdo en que la validación no es un proceso con un resultado binario, sino más bien de grado; es decir, la validación de una prueba no se emite como "sí" o "no", sino que se plasma en una presentación de evidencias que apoyan las inferencias que se hacen con los resultados de las pruebas.

Este apartado considera aspectos enfocados a la medición y los modelos matemáticos o estadísticos utilizados para asignar los puntajes a los estudiantes que toman las pruebas; a la validación del uso de los métodos; y a sustentar la forma de seleccionar las preguntas que integran la prueba y sirven de fundamento acumulativo al calcular puntajes y resultados. Estos elementos técnicos permiten justificar que los estudios y análisis pertinentes proporcionan suficientes evidencias de validez para reducir y estimar imprecisiones en las interpretaciones que pueden emitirse con los resultados de las pruebas. El conjunto de evidencias solicitadas se refiere a: estudios con modelos de correlación o multivariados sobre validez de criterio; el análisis estadístico de los factores que representan las áreas a evaluar; la verificación de la calidad dimensional de las escalas utilizadas; y el respaldo analítico sobre las decisiones tomadas por especialistas o jueces relativas a validez de constructo.

En general, los aspectos teóricos y del marco conceptual de la validez de contenido están vinculados con el currículo prescrito por la SEP. La elección de los referentes presenta un criterio general de atención a áreas de la Matemática (mayormente centrada en aritmética y geometría) y del Lenguaje (con énfasis en comprensión lectora), alineados al grado escolar y al perfil de competencias esperadas. Al parecer hubo algunos intentos de incluir otras temáticas o áreas del currículo pero , por apreciación de los especialistas o diseñadores de las pruebas, se revelaron

como un reto mayor para medirlas en aplicaciones masivas con reactivos de respuesta cerrada, por lo que las escalas están menos desarrolladas o son inexistentes.

Respecto de la validez de criterio, se cuenta con algunos estudios de calidad variable de validez concurrente, correlacionando resultados de la prueba PISA con los de las pruebas ENLACE-B (especialmente en la cohorte 2012) y EXCALE. Se reportan algunos resultados y problemas enfrentados por la diferencia de esquema de aplicación de las pruebas (muestrales y censales). Se sugiere sistematizar estos estudios por nivel , incluyendo las especialidades menos exploradas (ciencias naturales y ciencias sociales, por ejemplo), el desarrollo de competencias estratégicas y aspectos actitudinales con los cuales se identifiquen opciones para incrementar el desempeño académico.

Deben documentarse más a fondo los análisis sobre validez de constructo, a través de esquemas que muestren cómo se llegó al consenso entre especialistas de currículo, expertos en didáctica y autoridades. En ENLACE-B, los constructos y los contenidos a evaluar fueron definidos por responsables de currículo, mientras que en EXCALE se realizó un trabajo conjunto de especialistas, de diversos puntos de la geografía nacional, y en ENLACE-MS por comités específicos, cuyo proceso de análisisdebe documentarse más.

La validez de escala se trató en los primeros manuales técnicos de ENLACE-B con la distribución de ítems en forma gráfica, que no formó parte de los de otros años. El Manual cita, sin demostrar, elementos básicos que se supone está cumpliendo la prueba (distribución de dificultad de los ítems, el error de medida general de la escala en TCT, punto a punto en modelos de TRI, Mapa de Wright con Rasch). La escala debe tener reactivos para medir a las personas con la mayor precisión posible a lo largo de todo el continuo, principalmente porque se utilizan niveles de desempeño en intervalos definidos por puntos de corte fijados por especialistas. Se requieren análisis multivariados, factoriales y otros para justificar que las respuestas y resultados de los alumnos se organizan a posteriori en factores similares a los definidos por especialistas, en función de los contenidos y de los constructos definidos a priori. Este tipo de análisis se menciona en la documentación de las pruebas, pero sin demostración en reportes formales, los cuales son indispensables en el caso de las tres pruebas nacionales analizadas.

Independientemente de las restricciones de marco teórico del proyecto, el acopio de evidencias para la justificación técnica de la elección de perfiles y contenidos es heterogéneo entre las pruebas, lo cual sugiere la necesidad de un trabajo para alinearlas e identificar un modelo de evaluación longitudinal apto para analizar la evolución del sistema educativo en todos los grados, así como para identificar las áreas de oportunidad para el desarrollo del currículo; actualizar el proyecto educativo para que no privilegie solo dos áreas del conocimiento; o diseñar esquemas de apoyo psicopedagógico que incidan en una mejora de las prácticas de enseñanza. Ahora que se están rediseñando las pruebas, el desarrollo de una escala de alcance nacional sería más que oportuno, junto con estudios que permitan revisar los constructos con análisis factoriales o multivariados y trabajos por juicio de expertos.

El proceso de validación debe tener lugar a lo largo de la trayectoria de cualquier proyecto de evaluación del rendimiento educativo; hay muchas oportunidades de implementar actividades para obtener evidencias de apoyo a los usos que se pueden dar a los resultados. Una recomendación concreta consiste en conformar un panel de expertos que analice las evidencias de validez de las pruebas, las evalúe y recomiende estudios de validez que provean nuevas

evidencias. También es apropiado mencionar la pertinencia de atraer a investigadores (a través de programas de becas, incentivos o premios) para que hagan uso de las bases de datos a fin de alimentar las evidencias de validez de las pruebas.

Se encontró que la documentación que justifica y respalda las evidencias de validez es heterogénea y se revelan insuficiencias para las tres pruebas, aunque en menor medida para EXCALE que tiene una cantidad importante de publicaciones, pero que necesita un Manual Técnico actualizado. Hay algunos estudios parcialmente realizados sobre validez concurrente de las tres pruebas, en particular contra resultados o algunos reactivos de la prueba PISA, con las limitaciones implicadas en la comparación de pruebas muestrales y censales. La validez de escala solo se presenta en las primeras versiones del Manual Técnico de ENLACE-B, siendo un elemento a completar de manera sistemática.

Se recomienda demostrar que la escala tiene ítems a lo largo del continuo de conocimientos o de habilidades, para tener la mayor precisión posible en la estimación de medidas de los estudiantes respecto de los puntos de corte que definen los niveles de desempeño. Deben hacerse sistemáticamente estudios de validación, de alineación longitudinal de las especificaciones, por consenso entre jueces y análisis multivariado factorial, entre otros, para demostrar que los resultados experimentales de la prueba corresponden con factores similares a los definidos a priori por los especialistas. Se recomienda sistematizar la validación de las evidencias de validez por medio de un panel de expertos que proponga estudios o investigaciones que incidan en la mejora de las pruebas y retroalimenten al currículo y a las políticas educativas. Una última recomendación es promover el desarrollo de proyectos de investigación que usen las bases de datos de resultados de las pruebas.

Conclusión

Hemos presentado nuestras observaciones y juicios con respecto una selección de aspectos psicométricos y la forma en que éstos han sido abordados en las pruebas ENLACE-B, ENLACE-MS y EXCALE. En particular, hemos discutido los que tienen que ver con la elaboración y actualización de los manuales técnicos de las pruebas; las evidencias psicométricas de validez recolectadas y presentadas en apoyo al proceso de desarrollo, construcción y uso de las pruebas; la calibración, análisis psicométrico y confiabilidad de las pruebas; el análisis psicométrico y de calidad de los ítems; la calidad y gestión de los bancos de reactivos; la calificación y asignación de puntajes a los estudiantes; y, por último, pero no menos importante, el establecimiento de los niveles de desempeño y la interpretación de los resultados.

Observamos distintos niveles de desarrollo, actualización y atención de cada uno de estos aspectos, al igual que distintas respuestas ante los retos que presentan para un programa de pruebas. Si bien entendemos que cada reto, dadas las circunstancias específicas de un programa, requiere o conlleva a respuestas individualizadas que se adapten a su realidad y necesidades, consideramos fundamental que estas respuestas estén documentadas de manera detallada. Dicha documentación permitirá que los procesos empleados en la implementación de estos

programas puedan ser evaluados por actores externos, al igual que servir como evidencia de la calidad de la información recolectada. La documentación del proyecto se vuelve, por tanto, un instrumento didáctico para las generaciones futuras que tomen las riendas en estos procesos, así como en respaldo de las inferencias que se hacen con base en los resultados de las pruebas.

A riesgo de ser reiterativos, conviene insistir en que la heterogeneidad en la documentación que se observa entre las pruebas y, con mayor gravedad, dentro de las pruebas, hace difícil la lectura no solamente para los responsables de este estudio, sino para cualquier persona interesada en informarse acerca de la manera en que se desarrolla cada prueba. Esto impide tener una visión clara y completa del modelo de evaluación, así como de la exigencia de cada proyecto en relación con los aspectos psicométricos. Por otro lado, hace a cada proyecto vulnerable ante críticas y análisis , puede propiciar usos indeseables, además dar una apariencia de desvinculación no solo entre las agencias evaluadoras, sino entre los proyectos. Esta heterogeneidad se vuelve un área de oportunidad para la mejora que puede llevarse a cabo en el corto plazo.

Hemos hecho sugerencias y comentarios que creemos pertinentes para cada área específica. Estas sugerencias deben ser siempre consideradas y estudiadas cuidadosamente a fin de evaluar su pertinencia al programa. Como aclaramos en el párrafo anterior, no hay respuestas únicas, pero sí existen alternativas que son recomendables, dadas ciertas circunstancias. En este caso entran los estudios de sesgo y funcionamiento diferencial, los análisis de equiparación de pruebas, los estudios de confiabilidad por diversos medios, y los análisis de respaldo para la validez de constructo asociada con los niveles de desempeño en la escala de calificación.

Cabe resaltar que observamos un gran cuidado por parte de los técnicos de atender las necesidades de los programas, a la vez que deficiencias tanto en los procesos, como en la documentación de los mismos, que se deben no tanto al desconocimiento de proceso y técnicas, sino más bien a la premura con la que se llevan a cabo muchas de las actividades de estos programas. Estas deficiencias son áreas de oportunidad para encargar a otros investigadores o asesores externos su colaboración en la formalización de los documentos y estudios necesarios para dar certidumbre a cada proyecto.

Observamos también que las exigencias, frecuentemente poco ajustadas a la realidad educativa, y en desfase con las posibilidades de cualquier programa de evaluación, representan el mayor reto para los aquí analizados. Observamos que frecuencia y magnitud, en particular ENLACE-B y ENLACE-MS, no guarda proporción con la información y utilidad extraída de estas pruebas. Un redimensionamiento de la frecuencia y magnitud de las pruebas permitirá extraer con más facilidad elementos de retroalimentación y mejora al diseño del currículo, a los procesos educativos y a los usos de los resultados que pueda hacer la sociedad.

Atención a la diversidad cultural

El término *cultura* se refiere al conjunto de valores, experiencias, patrones de comunicación, formas de socialización y circunstancias históricas que comparten los individuos de un grupo social. Aunque el término en ocasiones evoca imágenes de sociedades antiguas o ajenas a la propia, en realidad todas las personas viven en comunidades de práctica que implican modelos culturales

diferenciados. El tipo de actividad económica, el nivel educativo, el socioeconómico, el tamaño y tipo de localidad (por ejemplo, rural, o urbana), la etnicidad, y el idioma, o las variedades de idioma que usa una persona, son factores que moldean una cultura y su combinación influye en la manera en que los sujetos dan sentido a las experiencias y perciben el mundo.

Como cualquier otro producto de la actividad humana, las pruebas de logro académico son productos culturales. Su formulación refleja las características de la cultura de los individuos que las crean. De tal suerte, la información contextual incluida en muchos ítems de prueba que buscan hacerlos significativos, refleja un contexto social específico y asume una serie de experiencias culturales compartidas por los alumnos examinados. La facilidad con que cada uno confiere sentido a ese contexto depende en gran medida de sus experiencias culturales. Tómese como ejemplo el siguiente ítem hipotético:

> Juan va a cenar con sus papás. Su papá deja $17.00 de propina. ¿Cuál es el total de la cuenta de la cena, suponiendo que se agrega el 10 % de propina a la cuenta total?

El reactivo no pregunta simplemente cuál es el total del que 17 es el 10 por ciento. El contexto de la cena en un restaurante hace que el problema requiera la aplicación de un concepto en una situación concreta significativa. Tal contexto no es el contenido evaluado; simplemente sirve como base para plantear un problema. El ejemplo planteado asume que el alumno está familiarizado con una estructura familiar (padre y madre) en la que el padre es el proveedor de recursos (es el papá quien paga), con cuentas de cena de una cierta magnitud ($170.00, en este caso), y con la práctica de pagar 10% de propina en restaurantes.

Cuando el contexto de un ítem asume ciertas experiencias culturales ajenas a las de un alumno determinado, este puede llegar a responder de manera incorrecta no necesariamente porque desconozca el contenido evaluado, sino por su falta de familiaridad con ese contexto.

En el año 2001 se planteó el concepto de "validez cultural" para referirse a la efectividad con que una prueba toma en cuenta las influencias socioculturales y lingüísticas que influyen en la manera en que los alumnos dan sentido a los ítems de una prueba y los responden. En una sociedad como la mexicana, con contrastes socioeconómicos acentuados y múltiples grupos lingüísticos y étnicos, el concepto de validez cultural es especialmente importante cuando se piensa en pruebas a gran escala que se aplican a la población estudiantil.

Desde hace tiempo se ha reconocido que las diferencias culturales son parte de los factores que atentan más seriamente contra la validez de la interpretación de los resultados de una prueba. El análisis de sesgo es una estrategia creada para asegurar una evaluación válida para distintos segmentos poblacionales. Este proceso busca analizar estadísticamente las diferencias en el desempeño de grupos poblacionales en cada uno de los reactivos de una prueba.

En la práctica, el análisis de sesgo es costoso y difícil de realizar con todos los reactivos de una prueba; además es una estrategia que se emplea al final del largo proceso su desarrollo de una prueba, en un punto en que es difícil hacerle mejoras significativas. Por tal razón, desde la

perspectiva de validez cultural, el análisis de sesgo es una estrategia necesaria pero no suficiente para asegurar la evaluación válida y justa para múltiples grupos culturales.

Para ello debe ser considerada la diversidad cultural en todas las etapas del desarrollo de una prueba, no solo en las finales. Esto implica el muestreo adecuado de la población objetivo (a fin de considerar, por ejemplo, distintos grupos étnicos y socioeconómicos); la participación de profesionales de diversas disciplinas (antropólogos, lingüistas, etcétera) en los equipos de elaboradores de pruebas; y la inclusión de muestras de distintos grupos lingüísticos y socioeconómicos y alumnos de diversas áreas geográficas en las etapas piloto de las pruebas (por ejemplo, cuando se evalúa si los alumnos entienden los reactivos de manera adecuada , a fin de realizar modificaciones apropiadas en su redacción).

La evaluación desde la perspectiva de validez cultural de EXCALE, ENLACE-B y ENLACE-MS se realizó con base en doce criterios, que reflejan tanto la práctica actual estándar en sistemas de pruebas a gran escala, como lo que la literatura moderna señala sobre validez, equidad, y diversidad cultural y lingüística. En otras palabras, los doce criterios con que se evaluaron las tres pruebas se basaron tanto en el conocimiento como en la práctica. Se consideró que establecer criterios basándose solo en lo que otros sistemas evaluativos hacen, contribuiría a perpetuar prácticas de efectividad limitada. Se sabe que las prácticas concernientes a la relación entre validez y diversidad cultural y lingüística empleadas actualmente por los sistemas evaluativos más importantes en el mundo tienen serias limitaciones. Al utilizar criterios que incluyen tanto la práctica actual como la necesaria, se establece un estándar alto de calidad y equidad y se contribuye al desarrollo de un sistema mexicano de pruebas con un nivel de atención a la diversidad cultural sin precedente.

Vale la pena señalar que para cada uno de los tres conjuntos de pruebas EXCALE, ENLACE-B y ENLACE-MS se realizaron microanálisis de conjuntos de reactivos seleccionados aleatoriamente.

Estos microanálisis incluyeron una revisión lingüística del fraseo de cada reactivo; la forma en que presentan las posibles respuestas (en el caso de los reactivos de opción múltiple); y si existe la posibilidad de encontrar más de una respuesta correcta (o ninguna). Este microanálisis puede definirse como un conjunto de razonamientos efectuados con el propósito de examinar la manera en que las características de los reactivos y los factores lingüísticos, culturales y socioeconómicos de los estudiantes se combinan y moldean la manera en que éstos interpretan los reactivos, a menudo con base en factores que no son los que los elaboradores de pruebas tienen en mente. Considera un análisis de los contenidos a partir de criterios de tipo cultural (por ejemplo, si es probable que los estudiantes estén familiarizados con los contextos de cada reactivo) y académico (si la información que se incluye en la evaluación corresponde a contenidos curriculares).

El microanálisis de reactivos de las tres pruebas reveló en algunos casos, aspectos problemáticos diversos, tales como falta de concordancia morfosintáctica, errores en las marcas de puntuación, o imprecisión en los contenidos. En casos extremos, se identificaron reactivos con más de una posible respuesta correcta y algunos sin respuesta correcta. Especialmente en el caso de ENLACE-MS, el microanálisis reveló que en ocasiones se usa información contextual que potencialmente puede ser desconocida o ajena para estudiantes de zonas rurales o de nivel socioeconómico bajo. Estas deficiencias parecen ser una consecuencia de la ausencia de marcos conceptuales que formalicen los contenidos y que aborden aspectos epistemológicos disciplinarios tales como la relación entre contenido, conocimiento y cultura.

En seguida se presentan y discuten los criterios empleados, con un resumen de las conclusiones obtenidas para cada una de las tres pruebas.

Marco conceptual de la prueba

El *marco conceptual de una prueba* es un documento que formaliza la estructura del conocimiento que se pretende evaluar; identifica los tópicos (por ejemplo, fracciones, punto decimal) y las habilidades (pensamiento crítico, solución de problemas, aplicación de procedimientos, etcétera) que deben tenerse en cuenta al elaborar una prueba. La combinación de cada tópico con cada una de las habilidades permite determinar de manera sistemática el contenido que debe tener la prueba. El marco conceptual de la prueba debe establecer los contenidos por materia (por ejemplo, matemáticas, comunicación) y por grado. Un marco de alta calidad discute los factores culturales y lingüísticos que influyen en el aprendizaje, la enseñanza y la evaluación de los contenidos objetivo y proporciona directrices para que estos aspectos sean considerados adecuadamente.

En ninguna de las tres pruebas se encontró un documento que las conceptualice; sin embargo, en el caso de EXCALE se identificó que aunque gran parte de la información que debe contener el marco conceptual se ha formalizado, se encuentra dispersa en diversos documentos. En ENLACE-B y ENLACE-MS se encontró poca evidencia de un trabajo conceptual profundo que sustente los contenidos de las pruebas. Aunque ENLACE-MS usa una matriz que establece los contenidos, no hay una discusión profunda de ellos.

La falta de marcos conceptuales de las pruebas impide identificar con claridad el vínculo entre sus contenidos y los factores culturales y lingüísticos que deben considerarse para su evaluación, y sustentar sólidamente los criterios que se siguieron para su elaboración.

> Las pruebas revisadas no cuentan con un marco conceptual explícito con base en el cual pueda asegurarse una adecuada alineación de los reactivos a una población social, cultural y lingüísticamente diversa.

Especificación de las poblaciones

Especificar la población a la que se le aplica una prueba consiste en identificar la estructura sociodemográfica de la población estudiantil y las cohortes de interés, de acuerdo con variables sociodemográficas tales como: tipo de actividad económica, escolaridad de padre y madre, nivel socioeconómico, tamaño y tipo de localidad, etnicidad y primera lengua.

El *marco muestral poblacional* es un documento —producto de la especificación de poblaciones— que identifica los porcentajes de alumnos dentro de cada una de estas cohortes. Ayuda a determinar las muestras de alumnos (y su tamaño) que deben incluirse al efectuar ensayos piloto de una prueba, con el fin de ajustar su contenido, refinar su redacción y otras características.

En la documentación de las tres pruebas analizadas se encontró evidencia de una atención incipiente a la diversidad cultural y lingüística del país. Sin embargo, no se especifica con detalle las poblaciones a las que se aplican.

En el caso de ENLACE-MS, aunque los cuestionarios de contexto colectan información sobre las variables mencionadas, no la emplean en los análisis de los resultados de las pruebas.

Por otro lado, llama la atención que, si bien los cuestionarios de contexto de las tres pruebas recogen información lingüística de los alumnos o de sus padres, no recaban datos que permitan determinar qué lengua específica habla el estudiante cuando responde afirmativamente a la presencia de hablantes de lenguas indígenas en su escuela u hogar.

Se encontró que, en general, se presta atención a las diferencias poblacionales solo con el propósito de reportar resultados de logro escolar para grandes grupos, pero sin el nivel de detalle que requiere un marco muestral poblacional. Ello da lugar a conclusiones sobre el rendimiento diferenciado por sector (*vgr.* primaria indígena frente a primaria general) cuya naturaleza es descriptiva pero no diagnóstica, e impide utilizar los resultados para mejorar la calidad de la educación en los grupos más desfavorecidos.

Estrategia para tratar diversidad cultural, lingüística y socioeconómica

Para abordar adecuadamente la diversidad cultural y lingüística es necesario adoptar una posición teórica que permita tomar decisiones sobre distintos grupos culturales. Esta posición debe ser consistente con el pensamiento actual de disciplinas como la antropología sociocultural y la sociolingüística y estar basada en evidencia empírica.

Por ejemplo, para tomar una decisión informada respecto de la inclusión en un sistema de pruebas de estudiantes de grupos indígenas cuya primera lengua no es el español, es necesario tener claridad conceptual de los factores que definen cada grupo indígena y de la diversidad étnica en el país, así como de las características tipológicas de cada lengua. En México, el Instituto Nacional de Lenguas Indígenas (INALI) contabiliza 364 variantes lingüísticas agrupadas en 68 lenguas que a su vez pertenecen a 11 familias lingüísticas diferentes, lo que implica una gran variedad en cuanto a sus sistemas gramaticales y a la manera en que éstos reflejan una forma particular de organizar el mundo a través del lenguaje.

En la documentación a la que se tuvo acceso, no se encontró evidencia de que los sistemas de pruebas hayan desarrollado una estrategia sistemática para abordar conceptualmente la diversidad cultural, lingüística y socioeconómica del país. Sin embargo, tanto EXCALE como ENLACE-B han efectuado análisis que comparan las calificaciones obtenidas por estudiantes de grupos por tipo y tamaño de localidad, nivel socioeconómico, y tipo de escuela. En el caso de ENLACE-MS, las comparaciones se realizan por modalidad educativa, sin especificar el tipo de población.

Especificación de ítems

En la práctica contemporánea, los sistemas de pruebas desarrollan un documento llamado *Especificación de Ítems* que describe en detalle los distintos tipos de reactivos que han de incluirse en las pruebas, su estructura, y otras características formales como el número y la extensión de las oraciones, la presencia de ilustraciones, etcétera. Un documento de especificación de ítems de alta calidad permite que dos equipos de elaboradores de pruebas, construyan reactivos de complejidad y aspecto similares para un mismo tipo de ítem, trabajando independientemente.

Cuando no existe un documento de especificación de ítems, o es de baja calidad, las características de los reactivos terminan siendo determinadas por factores idiosincráticos de los elaboradores de pruebas (por ejemplo, sus preferencias o estilos), lo que produce un error de medición considerable.

El análisis de la documentación disponible permitió concluir que, en el caso de EXCALE, existen documentos que guían a los elaboradores de pruebas acerca de la estructura y características que han de tener los ítems que construyen. Sin embargo, el nivel de especificidad podría ser más alto. En el caso de ENLACE-B, las descripciones de las características de los ítems son muy superficiales; se describen los porcentajes de ítems de cada tipo que han de constituir las pruebas, pero no se proporcionan detalles estructurales. En el caso de ENLACE-MS, no existe ninguna descripción formal y detallada de la estructura de los ítems; los reactivos de la prueba de un año se reemplazan al siguiente con otros de características similares, pero éstas no se justifican.

La carencia de criterios para la especificación de los ítems es una limitación seria que afecta no solo la calidad de la evaluación de poblaciones cultural y lingüísticamente diversas, sino la de la prueba en general, independientemente de los grupos poblacionales a los que se aplique.

Profesionales involucrados en el desarrollo de los ítems

En el proceso de desarrollo de las pruebas, la mayoría de los sistemas emplean profesionales de diversas áreas, que en ocasiones incluyen especialidades como antropología y sociolingüística. Idealmente tales expertos debieran participar en todas las fases del proceso, por ejemplo, formando parte de los equipos que redactan los ítems, y no simplemente como sus revisores al final del proceso de desarrollo de una prueba.

La documentación no incluyó evidencia de que profesionales de disciplinas que atiendan características sociales, culturales y lingüísticas del alumnado mexicano, sean incluidos sistemáticamente en el proceso de desarrollo de las pruebas.

ENLACE-MS tiene documentos en los que se da crédito por su participación a diversos profesionales, pero no incluye información sobre el tipo de participación.

Representación de poblaciones diversas en muestras piloto de alumnos

Una manera muy simple de promover la validez cultural de una prueba consiste en asegurarse de que los distintos grupos culturales, lingüísticos y socioeconómicos del país estén debidamente representados en las aplicaciones piloto de las pruebas; esta es la etapa en que se refinan sus características, basándose en las respuestas o comentarios de los alumnos.

En la documentación disponible se encontró evidencia limitada de que tal representación tenga lugar.

Sin embargo, debe mencionarse que en el caso de EXCALE y ENLACE-B se evalúa a un espectro muy amplio de escuelas clasificadas por nivel socioeconómico y localidad, aunque los análisis se efectúan al final del proceso de elaboración de las pruebas, y no como parte de su desarrollo. En la documentación de ENLACE-MS no se encontró evidencia de que tal representación tenga lugar, pese a que algunos de sus cuestionarios de contexto recogen información sociodemográfica. En ninguna de las tres pruebas hay evidencia de que el pilotaje de instrumentos se haya efectuado en diversas localidades, lo cual puede ser consecuencia de la ausencia de un marco muestral poblacional detallado.

Validación cognitivo-cultural

Hace alrededor de 25 años que se emplean protocolos verbales y entrevistas cognitivas para determinar si, al contestar los reactivos de las pruebas, los estudiantes usan los conocimientos y habilidades que se pretende evaluar. En tales protocolos y entrevistas, los estudiantes piensan en voz alta a medida que resuelven un problema o explican cómo entienden o interpretan los ítems.

Esta forma de validación, a la que se le llama *validación cognitiva*, puede ser enriquecida con preguntas para los alumnos sobre la manera en que relacionan el contenido de los ítems con sus experiencias culturales. Esta segunda forma de validación puede ser llamada validación *cognitivo-cultural*. Ambos procedimientos deben ser empleados en combinación con otras formas de validación.

En la documentación disponible no se encontró evidencia de que alguno de los tres sistemas de pruebas emplee alguna forma de validación cognitivo-cultural.

Revisión

En los sistemas más importantes de pruebas a gran escala, se aplican rutinariamente procedimientos de revisión para prevenir que las pruebas tengan sesgos culturales, antes de que éstas se apliquen. Uno de esos procedimientos consiste en conducir grupos focales con individuos de diferentes especialidades (por ejemplo, maestros, lingüistas, especialistas en contenido), para examinar si los reactivos podrían ser difíciles de entender para estudiantes de ciertos grupos socioeconómicos o culturales.

En la documentación consultada no se halló evidencia de que los procedimientos de revisión inherentes al proceso de desarrollo de pruebas consideren sistemáticamente fuentes de sesgo cultural, lingüístico y socioeconómico.

Solo en ENLACE-B se encontró información sobre una revisión de esta naturaleza, pero no existe evidencia de que se hayan tomado en cuenta las observaciones realizadas por profesores bilingües y otros profesionales del campo, expertos en contextos escolares culturalmente diversos.

Análisis de sesgo

El análisis de sesgo consiste en comparar el desempeño global en la prueba de dos poblaciones (una de referencia y otra objetivo) en un ítem específico, una vez que se controla por diferencias poblacionales. Si las muestras A y B contienen proporciones similares de estudiantes en cada uno de los niveles de desempeño en la totalidad de una prueba, y a pesar de ello la calificación obtenida por B es sustancialmente más baja que la obtenida por A en ese ítem, se dice que éste tiene un sesgo en contra de la población B.

A pesar de que el análisis de sesgo, basado en la TRI, es una de las técnicas mejor conocidas para examinar las propiedades técnicas de una prueba, solo se encontró evidencia de su uso en EXCALE, aunque no es claro si los análisis de sesgo se han efectuado de manera sistemática. Tampoco hay información sobre el tamaño de las muestras de ítems analizados.

Estudios de generalizabilidad

La Teoría de la Generalizabilidad (TG) permite analizar la estructura del error de medida en pruebas, mediante la identificación del porcentaje de error que aportan distintas fuentes de variación de puntajes, como los ítems empleados, los tópicos evaluados o los formatos de ítem aplicados. Los estudios de TG permiten determinar si, a partir de los resultados de una prueba, pueden hacerse generalizaciones válidas sobre las habilidades de los estudiantes.

A menudo, la generalizabilidad de una prueba puede aumentarse simplemente mediante el incremento del tamaño de la muestra de ítems; es decir, incrementando el número de reactivos de una prueba. La TG permite determinar el número de ítems necesarios para hacer generalizaciones válidas sobre el rendimiento académico de los estudiantes.

Aunque la TG existe desde hace más de cuatro décadas, su uso en la evaluación de poblaciones con gran diversidad cultural y lingüística es reciente. La investigación revela que las minorías culturales y lingüísticas son extremadamente heterogéneas. Debido a esta heterogeneidad, puede ser necesario emplear más reactivos en las pruebas de los que normalmente se incluyen para hacer generalizaciones válidas de las calificaciones obtenidas por los alumnos.

Generalmente, el número de ítems necesarios para garantizar una prueba válida se determina sin desagregar a las muestras de estudiantes, según nivel socioeconómico, grupo lingüístico, cultural, etcétera. Sin embargo, el empleo de estudios de generalizabilidad con muestras desagregadas de grupos poblacionales permitiría a los sistemas de pruebas evaluar si su calidad técnica es similar o diferente para distintas poblaciones.

Debido a su reciente aplicación en el campo de validez cultural, no sorprende que en la documentación analizada no se haya encontrado evidencia de que se efectúen estudios de generalizabilidad para ninguna de las pruebas, aunque su implementación sería de muy bajo costo y alto valor informativo.

Tiempos y calendarios

Uno de los problemas que obstaculizan más seriamente la evaluación válida de minorías culturales o de grupos de nivel socioeconómico bajo, consiste en que cuando se realizan acciones encaminadas a promover su evaluación justa (tales como la adaptación de las características de los ítems), éstas se efectúan al final del proceso de desarrollo de las pruebas.

Cuando tales acciones están consideradas en los calendarios de actividades de desarrollo de pruebas generalmente se les asigna poco tiempo. A menudo cualquier ajuste en el calendario se hace a costa del tiempo asignado a tales acciones.

En ninguna de las tres pruebas se encontró evidencia, de que estuvieran calendarizadas actividades orientadas a asegurar una evaluación válida y justa para los segmentos minoritarios de la población estudiantil.

Esta limitación parece derivar del hecho de que no existe un marco muestral poblacional que identifique segmentos poblacionales específicos, pues al no haber una formalización de los distintos grupos definidos por nivel socioeconómico, etnicidad tipo y tamaño de localidad, etcétera, es difícil que los sistemas de pruebas programen actividades relacionadas con su atención.

Mecanismos de corrección

Un reto fundamental para una evaluación justa y válida de minorías culturales es la implementación efectiva de mecanismos de corrección; los elaboradores de pruebas debieran saber exactamente qué hacer con los ítems en los que se detecta sesgo.

En la documentación disponible no se encontró evidencia de que exista una formalización de las acciones que deben tomarse para corregir aspectos de las pruebas identificados como desfavorables para las minorías culturales.

Como en el criterio anterior, esta limitación parece deberse a la ausencia de marcos muestrales poblacionales que consideren grupos específicos de interés definidos por nivel socioeconómico, etnicidad, tipo y tamaño de localidad, etcétera.

Conclusión

Las deficiencias detectadas en cuanto a los criterios analizados incluyen que la conceptualización de los contenidos evaluados no se detalla suficientemente, ni considera cómo influyen en los resultados las particularidades lingüísticas o culturales de la población escolar. La información sobre perfil de alumnos, modalidad educativa y tamaño de localidad, permite análisis que consideren esos factores, pero la organización de la pruebas no refleja un diseño que tome en cuenta expresamente la diversidad del país. En el desarrollo de la pruebas no se considera el tipo y grado de bilingüismo de los alumnos, lo que refleja la suposición de que todos son igualmente competentes en español y descarta la posibilidad de que esa no sea su lengua materna.

Las especificaciones para desarrollar reactivos carecen de precisión suficiente para controlar adecuadamente sus características gráficas, textuales y contextuales. No se encontró evidencia de que en el desarrollo de las pruebas hayan participado especialistas en disciplinas como la sociolingüística y la antropología. Tampoco hay evidencia de que se hayan piloteado con muestras representativas de grupos culturales, lingüísticos y socioeconómicos diversos, ni de que se realicen entrevistas para aportar evidencia de validez cognitiva o, de manera más particular, entrevistas cognitivo-culturales que analicen la forma en que las interpretaciones de los ítems que hacen los estudiantes están influidas por factores lingüísticos y culturales.

En el desarrollo de las pruebas parece no hacerse revisiones de aspectos como contenido, estilo, aspectos lingüísticos y posibles fuentes de sesgo cultural. No se encontraron indicadores de que se hayan efectuado los análisis apropiados para examinar el funcionamiento diferencial de los reactivos entre grupos poblacionales definidos por factores lingüísticos, culturales, socioeconómicos o de género. Con base en los microanálisis realizados, resulta evidente la importancia de que en el futuro se desarrollen sistemáticamente análisis de sesgo que incluyan no solo factores culturales y de diversidad lingüística, sino también regionales y socioeconómicos.

No hay evidencias de estudios de generalizabilidad para examinar confiabilidad y validez respecto de factores vinculados a competencias lingüísticas, o para comparar el desempeño de distintos grupos culturales y socio-económicos. Tampoco de que se prevea la necesidad de ajustar tiempos o calendarios de aplicación de las pruebas en función de la geografía o las condiciones climáticas de las diferentes regiones del país.

No parece haber procedimientos para eliminar reactivos con sesgo, ni estrategias y mecanismos de corrección de sesgo por factores como el género, la edad, los antecedentes escolares, el perfil lingüístico del hogar o el perfil laboral del estudiante y su familia, aunque los cuestionarios de contexto recaban alguna información sobre tales factores. Los microanálisis permitieron identificar fuentes potenciales de sesgo lingüístico y cultural en algunos de los reactivos analizados. Esas fuentes podrían identificarse y corregirse estableciendo un mecanismo formal para tal efecto.

Se subraya que la atención adecuada a la diversidad lingüística en pruebas para educación básica o media superior no debe entenderse como la recomendación de que se traduzcan las pruebas a len-

guas indígenas, no solo porque el proceso es largo, costoso y difícil de implementar correctamente, sino porque en casi todo el sistema educativo nacional, la lengua de instrucción es el español.

APLICACIONES

La aplicación es un paso crucial para toda prueba, pues es cuando se hacen llegar los cuadernillos con ítems a los sujetos, se promueve su respuesta, se recoge la información y se analiza el comportamiento de los ítems. De la forma en que se realice dependerá la calidad de los puntajes y su utilidad para los fines de la prueba. Estandarizar las aplicaciones es importante para controlar la mayor cantidad de variables que puedan afectar las respuestas, e implica que se defina e implemente un procedimiento uniforme, para asegurar que los sustentantes tienen la misma oportunidad de demostrar su talento. Las limitaciones de estandarización de la aplicación de una prueba pueden poner en riesgo la generación de puntajes comparables, la medición del constructo y, por tanto, la utilidad e interpretación de resultados. Condiciones laxas, inseguras o no estándar de aplicación pueden invalidar la interpretación de los puntajes para algunos o todos los examinados y estropear el trabajo de las otras etapas del proceso de desarrollo de la prueba.

Los criterios de calidad de las aplicaciones comprenden aspectos que se realizan antes, durante y después de la aplicación.

Antes de la aplicación es importante contar con un listado de escuelas actualizado y confiable, sea para una aplicación censal o como marco muestral; que las muestras estén basadas en diseños sólidos y los estratos hayan sido definidos con base en argumentos defendibles; y tener controles para verificar que los sustentantes sean los que se planificaron. Además es necesario un minucioso proceso de planeación de la aplicación que incluya manuales probados en campo; un cronograma detallado; identificar al personal que participará en la aplicación; precisar procedimientos para garantizar confidencialidad y seguridad de los materiales y las respuestas de los sustentantes; y mecanismos para controlar la calidad de la aplicación. En esta etapa se selecciona y capacita al personal de aplicación, lo cual involucra la definición de criterios para reclutarlo, seleccionarlo y entrenarlo; el establecimiento de procedimientos de capacitación que aseguren el conocimiento de los materiales y el dominio de las funciones a realizar en campo, la documentación de estos procesos, y la definición de procedimientos para monitorear la aplicación.

Durante la aplicación de las pruebas se busca que no haya irregularidades que puedan afectar las respuestas, lo que implica motivar la respuesta de los alumnos; contar con mecanismos para lidiar con la no respuesta y prevenir y enfrentar la copia o cualquier tipo de fraude; implementar controles de calidad que permitan asegurar que las condiciones de aplicación sean estandarizadas, que se realicen conforme a lo planeado y se aseguren los materiales y las respuestas.

Después de la aplicación es importante tener procedimientos sistematizados para preparar el procesamiento de datos y contar con personal calificado para su manejo ya que la conformación y verificación de las bases de datos es crucial en esta etapa, así como la documentación de los procesos y la definición de procedimientos para notificar y documentar irregularidades.

En este apartado se presentan los principales hallazgos de nuestro análisis de las aplicaciones de ENLACE-B, ENLACE-MS y EXCALE. Los hallazgos se presentan de acuerdo con los momen-

tos del proceso: en primer término, en cuanto a criterios de validez previos a la aplicación; en segundo, respecto de criterios durante la administración de la prueba; y en tercero relacionado con criterios posteriores a la aplicación. La mayoría de retos que se enfrentan antes, durante y después de una aplicación son comunes a las pruebas analizadas, pero existen particularidades que ameritan atención por sus implicaciones para los objetivos de cada una.

Criterios de calidad antes de la aplicación

Selección de la muestra

Como punto de partida para una aplicación censal o muestral es indispensable una base de datos de escuelas actualizada y confiable, pues de ella depende la determinación de los recursos humanos, tecnológicos y financieros necesarios para la aplicación, al igual que la solidez de los análisis que se efectúen sobre los resultados y las conclusiones que puedan derivarse de ellos. Para ENLACE-B y MS la base de datos se genera a partir de la información provista por las entidades federativas, no de un sistema central, por lo cual reviste particular importancia la precisión de los reportes estatales, que deben tener procesos de actualización y verificación de la información. El marco muestral de EXCALE se construye a partir de la Estadística Educativa (forma 911) del ciclo escolar anterior. La validación de la muestra busca detectar y subsanar las inconsistencias sobre número de grupos, docentes, alumnos, ubicación, y demás datos requeridos para la aplicación.

Aunque la actualización de las bases de datos de las tres pruebas fue considerada adecuada por la mayoría de las entidades federativas al responder un cuestionario al respecto, es probable que algunas limitaciones en el proceso de actualización hayan incidido en las oscilaciones del censo de escuelas y alumnos de ENLACE-B, y en irregularidades presentadas en la aplicación, como falta de correspondencia de las cajas y hojas de respuesta entregadas en cada escuela, e insuficiencia de cuadernillos y hojas de respuesta. En ENLACE-MS no se aprecian cambios en las tendencias en el censo de escuelas y estudiantes a lo largo de siete aplicaciones. El comportamiento de la población de estas pruebas es bastante predecible, y las previsiones para la aplicación son más confiables, con el ahorro asociado en costos y tiempo. En EXCALE se advierten errores de domicilios y claves de escuelas, y presencia de escuelas que habían sido dadas de baja.

> Convendría valorar la posibilidad de conformar un sistema único de información, que se actualizara en las entidades con orientaciones específicas y sirviera para conformar las bases de datos de las pruebas. Se sugiere revisar los plazos en que éstas se solicitan los que se dan para su validación, a fin de asegurar que sea suficiente; estos plazos debieran considerar el número de escuelas y algunas condiciones de las entidades federativas que podrían hacer la validación más compleja (por ejemplo, su geografía y vías de acceso a las localidades). Además, podrían emplearse proyecciones de cambios derivados de los movimientos de matrícula y escuelas, y con mecanismos para hacerles frente.

En ENLACE-B y ENLACE-MS se utilizan diseños muestrales solo para el pilotaje de reactivos que serán incluidos en las pruebas operativas y el pre-test, y para las muestras controladas. Los manuales técnicos de ambas pruebas están redactados de forma que generan confusión acerca de

las muestras que se emplean, sus propósitos y procedimientos de cálculo. Las descripciones de las muestras no clarifican si se refieren a las del pilotaje de reactivos o a las muestras controladas utilizadas durante la aplicación definitiva; tampoco parecen estar considerando la naturaleza anidada de las unidades de muestreo.

En cuanto al pilotaje de reactivos de ENLACE-B y MS, por razones de costo y accesibilidad, éste se ha realizado en el Estado de México y el Distrito Federal, ya que se considera que cuentan con planteles representativos de todas las modalidades educativas. Es necesario fundamentar la selección de estas entidades y mostrar evidencia de su adecuación para valorar si los constructos medidos a través de los reactivos probados tienen el mismo significado para todos los individuos de la población objetivo. Se sugiere que el pilotaje también incluya la valoración del proceso de aplicación o administración de la prueba; esto implicaría considerar también la diversidad de composición y características de las Áreas Estatales de Evaluación (AEE), como uno de los elementos en la definición de la muestra.

Las muestras controladas de ENLACE-B y MS parecen servir para propósitos que no son completamente equivalentes. En ENLACE-B se lleva a cabo la aplicación del pre-test, conformado por reactivos seleccionados en el piloto. En ENLACE-MS, con la muestra controlada, se pretende: a) una aplicación más rigurosa de la prueba operativa, de manera que se obtenga información confiable sobre los valores psicométricos de los ítems que se utilizarán para la calificación; b) aplicar el pre-test para realizar la equiparación con la operativa y así mantener los puntajes en la misma escala; c) aplicar cuestionarios de contexto para obtener información sobre las características de los alumnos; y d) realizar estudios experimentales. Tener propósitos comunes en las muestras controladas de aplicaciones censales facilitaría la comprensión de este proceso por las AEE y demás participantes en las diferentes aplicaciones.

En EXCALE el muestreo pretende asegurar que los resultados de logro sean representativos a nivel nacional, por entidad y por modalidad educativa, por lo que sigue un esquema probabilístico, estratificado, bietápico y por conglomerados. Para valorar la solidez del diseño muestral se han hecho estudios que permiten comprobar que las nuevas versiones de las pruebas, a excepción de las muestras por entidad, tienen una precisión comparable a estudios internacionales como TIMSS o PISA (10% de desviación estándar en la variable de interés).

La validación de diseños muestrales y la selección de unidades por instancias diferentes a las que realizan el diseño muestral es una práctica que se ha hecho de forma no sistemática en las pruebas analizadas. En la aplicación 2008 de ENLACE-MS, el INEE validó la selección de unidades propuesta por el CENEVAL. Convendría que la validación del diseño muestral y la selección de escuelas y alumnos por especialistas distintos a los responsables del diseño sea una práctica regular en las diferentes evaluaciones, y que los reportes se añadan a los documentos técnicos, para contribuir a sustentar la solidez del diseño muestral.

En los documentos técnicos de las pruebas analizadas no se encontró información sobre los márgenes aceptables de muestra obtenida con respecto a la planificada. Los márgenes previstos, al igual que los reemplazos de escuelas en la muestra, deben especificarse como parte de un sistema de aseguramiento de la calidad, que incluya los procedimientos definidos para la validación de la muestra por las AEE y los registros que hacen los aplicadores durante la administración de las pruebas. Convendría que los reportes de aplicación incluyeran cifras absolutas y relativas de escuelas y alumnos programados y obtenidos, desagregados por dominios y estratos muestrales. Deben aprovecharse los reportes de aplicación generados por el INEE para los EXCALE.

Planeación de las aplicaciones

Para posibilitar la aplicación de las pruebas debe haber una planeación detallada de los recursos humanos, técnicos y financieros necesarios para la recepción segura de originales en medio digital e impreso, el monitoreo de todo el proceso de producción de instrumentos, preprensa, impresión, personalización, empaque, distribución, transporte, aplicación propiamente dicha y procesos posteriores.

Las tres pruebas cuentan con manuales técnicos impresos que hacen referencias generales a los pasos del proceso de aplicación, desde la organización administrativa de los recursos humanos hasta la entrega de los archivos de lectura. Estos manuales podrían complementarse con la inclusión de tiempos aproximados que sirvan como un referente común para todas las entidades federativas. Se recomienda utilizar materiales de apoyo adicionales, aprovechando las nuevas tecnologías, que aseguren el dominio de los procesos por los diferentes actores participantes.

En el documento "Estrategia Operativa" de EXCALE se consigna la información sobre los diversos actores del proceso y las fechas en que deben hacerse las labores preparatorias. Dada la diversidad y extensión del país, hay aspectos de la estrategia operativa que pueden variar entre entidades federativas; sin embargo, tener una estructura unívoca facilita el control y optimización del proceso. Es particularmente detallada la "Información para el ENLACE" de EXCALE, que incorpora las actividades que deben hacer quienes ejerzan dicha función durante las últimas tres semanas antes de la aplicación y hasta dos semanas después, culminando con la retroalimentación.

La calidad de los resultados de una evaluación estandarizada está estrechamente relacionada con la de los cuadernillos y hojas de respuestas; por ello debe contarse con el personal y tiempo apropiados para la revisión y eventual ajuste de los originales, lo que permita mitigar riesgos de exclusión de ítems cuyo funcionamiento pudiera haberse visto afectado por problemas asociados con su producción.

Las tres pruebas disponen de especificaciones para los procesos de impresión y personalización de materiales, pero deben identificarse explícitamente características de seguridad de la imprenta, los medios de transporte y almacenamiento de los materiales de evaluación. Una mejora importante para el proceso de impresión de materiales de evaluación sería el establecimiento de estándares de control y seguridad por parte de un ente externo al impresor. Además es recomendable que los cuadernillos, cuestionarios y hojas de respuestas, se entreguen diagramados y verificados en versión final para impresión, y que la imprenta se circunscriba a la reproducción de las cantidades que se le especifiquen por cada tipo de material, con supervisión externa antes, durante y después de la impresión, asegurando la destrucción de originales, muestras y sobrantes, así como la apropiada disposición de materiales para archivo. Esto implica que la imprenta no ensamble cuadernillos a partir de bloques de ítems. Es probable que para ello se requiera el fortalecimiento del equipo interno de diagramación y armado, al menos durante los picos de trabajo, lo cual redundaría en mayores garantías para el proceso.

En ENLACE-B y MS se hace referencia a normas ISO para el control de la calidad, que especifican los parámetros aceptados para su aseguramiento durante todas las etapas del proceso. Este tipo de monitoreo de procedimientos de control de calidad no es práctica común; no parece haber otro programa internacional o nacional (PISA, TIMSS, PIRLS, SABER, etcétera) que utilice este tipo de normas de aseguramiento de la calidad. Por otra parte, para las aplicaciones realizadas por el INEE se han definido estándares específicos de levantamiento de datos que recuperan tanto la experiencia del Instituto como de aplicaciones de pruebas realizadas con anterioridad en México. Se recomienda que los estándares de control de calidad de las evaluaciones que se manejen se establezcan en apego a los aplicables a evaluaciones educativas estandarizadas.

Los manuales, instructivos y demás formatos de soporte, son herramientas clave para el entrenamiento del personal, contribuyen a la estandarización y, por ende, a tener resultados confiables y comparables. Para las tres pruebas se cuenta con manuales de aplicación probados en campo, que precisan las tareas a desarrollar por cada participante. Estos documentos incluyen elementos generalmente aceptados para la aplicación de pruebas estandarizadas, y podrían enriquecerse haciendo uso de la tecnología, incorporando gráficas, animaciones y sonido, que permitirían contar con herramientas didácticas para sesiones de entrenamiento, buscando mejorar la capacitación del personal de campo, lo que redundaría en una mayor estandarización en el ejercicio de sus responsabilidades.

Selección y capacitación del personal de aplicación

La selección, vinculación, entrenamiento y remuneración del personal requiere de importantes esfuerzos orientados a asegurar la calidad de la aplicación. Por ello es recomendable establecer normas generales para todo el personal involucrado, que sean seguidas homogéneamente por las entidades federativas, en contraste con lo señalado en el manual de ENLACE 2013, en el sentido de que cada estado es responsable de definir las características, requisitos y forma de contratación del personal que eventualmente se contrate para algunas de estas tareas. Establecer roles y perfiles es sencillo y reduce la probabilidad de una implementación dispar de procedimientos que puede provocar riesgos innecesarios.

En cuanto a EXCALE, en el documento "Estrategia General de Capacitación" se establecen roles y perfiles del personal de aplicación, en forma que parece suficiente: contar con escolaridad equivalente a licenciatura y experiencia acorde a la responsabilidad.

En las tres pruebas se ha implementado una estrategia de capacitación en cascada que responde a las necesidades básicas del proyecto; sin embargo, las orientaciones para su implementación varían. En ENLACE-B y MS se proveen esencialmente los materiales a usar con diversos actores, sin incluir orientaciones para la capacitación y su monitoreo. En EXCALE se precisa la estructura y contenido de la capacitación, las actividades a desarrollar y su duración; se implementan también cuestionarios para recuperar la retroalimentación sobre el proceso de capacitación por parte de diferentes actores. Se recomienda que todas las pruebas consideren la provisión de materiales y el diseño de estrategias de capacitación, buscando la efectividad y estandarización del entrenamiento para todo el personal.

Por otra parte, para todas las aplicaciones con control externo a la escuela, incluyendo pre-tests, podrían considerarse operativos de distribución y recolección de materiales independientes, que aseguren su llegada a cada escuela el día de la aplicación, mitigando riesgos operativos. Esto tiene un impacto financiero, pero podría brindar mayor seguridad a los ítems a ser aplicados en un futuro. En general se cuenta con previsiones apropiadas para la exitosa conducción de las aplicaciones y el tratamiento de posibles contingencias, que se reportan a través de formatos impresos. Se hace necesario tipificar las situaciones que se presentan durante la aplicación y determinar su frecuencia, a través de una solución tecnológica que capture los datos directamente, sin reprocesamiento ni transcripción, y permita contar con informes detallados por escuela.

Criterios de calidad durante la aplicación

Minimización de carga, motivación, no respuesta y fraude

En las tres pruebas se cuenta con información suficiente para fijar límites realistas de carga para los estudiantes. Los procedimientos para responderlas son sencillos, se han mantenido estables y no generan carga importante para los estudiantes. En cambio los cuestionarios de contexto para alumnos son muy extensos. Se debe determinar su pertinencia en función del aprovechamiento real que se hace de ellos y explorar qué información se puede recuperar a través de otras fuentes.

Las aplicaciones se llevan a cabo durante la jornada escolar, se anuncian a todos los actores del proceso con antelación suficiente y por diversos medios. Como excepciones se han presentado situaciones especiales que han dificultado que algunos sustentantes potenciales apliquen efectivamente la prueba.

Hay un importante despliegue en medios acerca de los propósitos de las pruebas y el uso de sus resultados, que se refuerza con previsiones explícitas sobre el particular en los manuales, a efecto de motivar a los alumnos para que hagan su mejor esfuerzo y no respondan a la ligera.

La prevención de copia y demás formas de fraude, incluyendo suplantación de identidad, dictado y sustracción de materiales de evaluación, entre otras, es de especial importancia cuando se generan puntajes individuales para los alumnos, que a su vez son agregados en diferentes niveles (aulas, escuelas, zonas educativas, regiones, entidades). La integridad de los puntajes debe ser promovida desde antes de la aplicación; también se deben implementar medidas durante y después de ésta, tales como definir requerimientos de identificación de los estudiantes; asignar a los sustentantes asientos específicos; requerir un espacio adecuado entre ellos; monitorear el proceso en forma permanente; y restringir el acceso de celulares, tabletas y otros aparatos.

En ENLACE-B y ENLACE-MS se precisa al personal de la aplicación la importancia de no permitir la copia. También se solicita que en caso de detectar acciones de copia o dictado de repuestas, se registre lo propio en el reporte de irregularidades. En ENLACE-B las instrucciones a aplicadores (docentes) se dan el mismo día de la aplicación, por parte de los directores. En esta prueba se identifica una sanción indirecta si se detecta alguna irregularidad, pues en la conformación del puntaje de "Aprovechamiento Escolar" de los docentes que participan en Carrera Magisterial no se toman en cuenta los puntajes de alumnos donde se detectó algún tipo de copia. Es necesario que todas las aplicaciones, incluyendo la de EXCALE, incorporen medidas preventivas de comportamientos de copia o fraude durante la aplicación. También debe proporcionarse información a los aplicadores y coordinadores de aplicación acerca de comportamientos de los estudiantes que pueden indicar que se está dando la copia o algún tipo de fraude y las medidas que deben tomar ante el incumplimiento de la normatividad establecida.

Procedimientos para el control de calidad de las aplicaciones

El control de la calidad de las aplicaciones es fundamental para el éxito de toda evaluación estandarizada ya que posibilita la valoración objetiva del cumplimiento de lo previsto, la identificación de oportunidades de mejora y la cualificación de los diversos procesos y procedimientos asociados con la operación de campo. Es necesario que el control de calidad se planee y ejecute de forma estandarizada y que los hallazgos queden debidamente registrados para aprovechar la información en favor de futuros ciclos del respectivo proyecto.

Los procedimientos de control de las aplicaciones están previamente establecidos y se instruye a la mayoría de los actores involucrados en este proceso para su apropiada implementación. Debe contarse con mayores facilidades tecnológicas para la captura de la información en la fuente, que permitan disponer de reportes integrales en el menor tiempo posible acerca de la aplicación, las incidencias que se presentaron y la forma en que se trataron, para optimizar la estandarización de los procesos e incrementar la calidad de las aplicaciones en cada nuevo operativo.

En ENLACE-B se hacen esfuerzos de aseguramiento de calidad, con apoyo de padres de familia, miembros de Consejos Escolares de Participación Social y observadores externos. La capacitación de estos actores se realiza el día de la aplicación, lo que podría ser una desventaja para el dominio de sus funciones.

En EXCALE los enlaces ejercen la función de monitores externos, y el control de calidad de la aplicación se apoya en reportes escritos llenados por el personal involucrado. Aunque se llevan a cabo capacitaciones intensivas para el personal de aplicación, en cerca de la mitad de las entidades federativas, el apego al protocolo de aplicación es inferior a 70%.

En ENLACE-MS se cuenta con participación de observadores externos, a quienes se invita a verificar el cumplimiento de las normas durante la aplicación. No hay un proceso previo de entrenamiento, aunque en los manuales se sugiere que los directores se reúnan con ellos y con los padres de familia una semana antes de la aplicación. Para esta prueba, de acuerdo con la información recolectada, solo en cinco entidades federativas (de las 25 que respondieron el cuestionario) participaron observadores externos. Dos de las entidades señalaron que hubo observadores externos en la mayoría de las escuelas (90% o más); en las otras tres no contaron con registro. Es importante señalar que solo en una de las cinco entidades se implementó un proceso de capacitación para estas figuras.

En las tres pruebas se encontró que algunas entidades federativas diseñan estrategias de monitoreo de las aplicaciones, pero no cuentan con orientaciones y recursos provistos por la DGEP o el INEE que promuevan el diseño, implementación y capitalización de la información recuperada.

Criterios de calidad posteriores a la aplicación

Preparación del procesamiento de los datos

El manejo de las bases datos es de gran importancia, pues de él depende la integridad de los puntajes y otra información generada a partir de datos primarios.

Para las tres evaluaciones se encontró documentación técnica que detalla el proceso de lectura óptica de respuestas y preparación de los datos para el proceso de calificación, incluyendo algunos elementos de verificación y limpieza. En cambio, no se encontró información descriptiva y normativa sobre la creación, estructura, formato y cuidado de los archivos, la forma de introducir datos y el procedimiento para asignar identificadores a sustentantes. Aunque de manera general en la documentación técnica de las tres pruebas se encuentra información similar, hay variaciones sobre el detalle de la descripción y la actualización de la información. Es necesario definir la información común que debe tener la documentación técnica sobre verificación y preparación de datos para el análisis, y que se actualice y complemente con normas que faciliten cumplir con este criterio.

La capacitación, entrenamiento y calificaciones del personal que maneja las bases de datos debe precisarse en la documentación técnica, permitiendo valorar si esos elementos son adecuados para la comprensión y realización de los procesos técnicos establecidos. Esta información también sirve para la selección de nuevo personal y asegurar que cuente con entrenamiento y capacitación para realizar los procedimientos tal como se hayan estipulado.

En la revisión de la documentación técnica de las pruebas se encontró información sobre las áreas responsables del manejo de las bases de datos. Sin embargo, la información sobre la formación, experiencia y entrenamiento del personal no está documentada; solo de manera general se señala que el personal debe contar con perfiles técnicos y experiencia acordes con los requerimientos. Es necesario explicitar el perfil profesional y la capacitación necesaria para el personal de estas áreas.

El aseguramiento de la calidad de las bases de datos comprende verificar que la estructura de los datos se apegue a la de los instrumentos; que los datos tengan suficientes redundancias; que las bases tengan identificadores únicos de los informantes, que sean consistentes y permitan, si es necesario, relacionar la información de alumnos y escuelas; que se hagan verificaciones aleatorias de submuestras de las bases de datos; y que se documenten los procedimientos de preparación de dichas bases.

En este punto la información de las tres pruebas varió considerablemente. En EXCALE se encontró la más completa, incluyendo todos los procedimientos antes descritos para asegurar la calidad de las bases de datos. En ENLACE-B la mayor parte de los procedimientos están documentados, con excepción de las redundancias de datos para control de calidad, y las actividades de preparación. En ENLACE-MS no se encontró documentación oficial que describa los procedimientos para maximizar la calidad de las bases de datos. La diferencia entre la información encontrada a este respecto en las tres pruebas evidencia la necesidad de sistematizar y aprovechar los procedimientos definidos para cada una de ellas.

Procesamiento y verificación de los datos

El procesamiento y verificación de datos tiene como propósito producir archivos libres de errores. Los procedimientos involucrados son de gran importancia para contar con datos confiables para los análisis previstos.

Los procedimientos para asegurar que la lectura de respuestas y el procesamiento y verificación de datos son confiables, comprenden hacer dobles verificaciones en forma sistemática; en caso de que la lectura de datos se haga en forma descentralizada, asegurar que se cumplan los estándares en todos los sitios; revisar que la estructura de bases se apegue a la acordada, las variables estén en rangos válidos y los identificadores sean únicos e íntegros; contrastar archivos de datos con instrumentos y cuestionarios; calcular estadísticas analíticas para cada ítem y estadísticas descriptivas para todas las variables con el fin de verificar que no haya valores extremos o faltantes; y documentar todos los pasos del proceso.

Se encontraron variaciones en la documentación técnica de las tres pruebas. En EXCALE la documentación da cuenta de todos los procedimientos descritos para asegurar la confiabilidad de la información y se incluyen reportes derivados de los procedimientos realizados. Convendría incluir sistemáticamente reportes que muestren su aplicación, el informe de resultados y las decisiones tomadas.

Los manuales técnicos de ENLACE-B incluyen los procedimientos descritos en este apartado, aunque no se dispone de reportes o evidencias de que hayan sido realizados conforme a lo establecido. Una limitación importante se refiere a la verificación de los estándares de lectura de datos, cuando se hace en forma descentralizada (parcial o total), pues no se encontraron procedimientos específicos al respecto. Tampoco se identificaron los que corresponden al contraste de archivos de datos con instrumentos y cuestionarios. En ENLACE-MS la documentación no incluye procedimientos sobre el procesamiento de datos y el aseguramiento de su calidad. De acuerdo con información adicional, se identificó la existencia de procedimientos de análisis de reactivos que se implementan en las pruebas piloto, pre-test y operativa para verificar su calidad psicométrica, así como el cumplimiento de los lineamientos técnicos institucionales y de contenido. Convendría que estos procedimientos fueran incluidos en la documentación técnica de la prueba y que se añadieran evidencias de su realización conforme a lo establecido.

Después de la aplicación, en ENLACE-B se aplica un algoritmo de detección de copia que se menciona en los manuales. No se encontró información descriptiva de este proceso ni evidencias concretas de los estudios realizados. En EXCALE se hace una revisión de trenes de respuesta para detectar casos de comportamiento irregular que deben ser analizados por personal especializado del INEE. En ENLACE-MS no se encontraron evidencias de aplicación de algoritmos ni del uso de herramientas tecnológicas para detección de copia. Sería deseable que los procesos empleados por ENLACE-B y EXCALE se documenten y analicen, para que se sistematice la experiencia y se aproveche en otras evaluaciones.

Notificación de irregularidades

La notificación de irregularidades permite identificar las desviaciones entre lo previsto y lo ejecutado en campo, como insumo para la toma de decisiones durante el procesamiento de datos. Es fundamental disponer de un reporte agregado que documente las eventualidades y facilite su resolución.

En las tres pruebas los actores de la operación de campo tienen entre sus responsabilidades documentar las irregularidades que encuentren en el ejercicio de sus funciones, mediante el llenado de formatos impresos que son compilados posteriormente. No se encontraron reportes de irregularidades para las pruebas ENLACE-B y MS. En EXCALE se elabora un reporte técnico de la aplicación, que consolida la información obtenida en las diferentes etapas del proceso.

Es necesario contar con una solución tecnológica que permita la captura de los datos directamente en la fuente y su sistematización en tiempo real, evitando transcripciones y demás procesos manuales, a efectos de contar con información, que además de propiciar la mejora del proceso, ahorre tiempo y facilite la toma de decisiones por parte de las instancias de dirección de los proyectos.

Conclusiones

La aplicación de pruebas de logro educativo es compleja y tiene gran importancia para los propósitos de mejora de la calidad de la educación. Su complejidad radica en los múltiples aspectos y procesos técnicos que tienen que considerarse antes, durante y después del proceso de aplicación, mismos que deben ser consonantes con los propósitos para los que ha sido diseñada la prueba. Téngase presente que en México la complejidad de los procesos de aplicación también tiene que ver con la manera en que está organizado el sistema educativo nacional y los sistemas educativos de las entidades federativas, así como sus características geográficas y condiciones sociales y políticas.

Tanto las pruebas censales como las muestrales requieren de una gran coordinación en el nivel federal, estatal, escolar y con otros actores involucrados, como pueden ser los observadores externos. Evidentemente, las aplicaciones censales conllevan un desafío mucho mayor; la periodicidad y dimensión de las evaluaciones hace que los tiempos asociados a cada etapa sean reducidos, y esta condición puede limitar la generación de lineamientos estandarizados y probados que guíen la aplicación en campo, su control de calidad, la generación y sistematización de reportes de los procesos involucrados antes, durante y después de la aplicación, y la retroalimentación para aplicaciones futuras. Es necesario que estas consideraciones sean tomadas en cuenta como insumos para la definición de futuros planes de evaluación.

Contar con una base actualizada de datos de alumnos, docentes y escuelas así como verificarla mediante auditoría externa e identificar sistemáticamente oportunidades para la mejora de los procesos, es fundamental para contar con la información requerida para llevar a cabo cualquier evaluación estandarizada. En tal sentido, se hace necesario disponer de los recursos humanos, tecnológicos y financieros que aseguren la disponibilidad oportuna de la información requerida para la aplicación de las pruebas. Las evaluaciones de logro educativo, independientemente de su carácter muestral o censal, deben contar con documentación sistematizada que muestre no solo los procedimientos establecidos, sino que además permita establecer su cumplimiento. Igualmente, es fundamental implantar controles de calidad estrictos en cada una de las fases de la aplicación; que se involucre personal externo al proyecto para asegurar su calidad; y se diseñe un proceso sistemático de mejora continua, que retome la experiencia de cada una de las aplicaciones como base para la toma de decisiones futuras.

La variación entre procedimientos de aplicación encontrada en las tres pruebas analizadas evidencia la necesidad de una definición común que facilite el trabajo de los diversos actores involucrados en la aplicación y el seguimiento de la misma, a la vez que se aprovechen experiencias positivas y se busque su adopción general. Esto puede ser de gran ayuda para las entidades federativas, dado el papel que desempeñan en los procesos de aplicación muestral y censal.

En las tres pruebas el control y la toma de decisiones acerca de la aplicación, podrían mejorar sustancialmente mediante la adopción de tecnologías que permitan tener información procedente directamente de la fuente, en línea, antes, durante y después de la aplicación, manteniendo soportes impresos o correos electrónicos como mecanismos alternos para casos de carencia de infraestructura.

Se sugiere también que para futuras aplicaciones, en especial las muestrales, se analice la viabilidad financiera y logística de contar con un operador para distribuir los materiales directamente de la imprenta a la escuela, y su posterior recolección y retorno, para obviar instancias interme-

dias y asegurar la cadena de custodia de los cuadernillos y hojas de respuestas antes, durante y después de la aplicación.

Usos y consecuencias

El interés que han generado en México los resultados de las pruebas de gran escala, y los múltiples esfuerzos e iniciativas relacionados con ellas en todos los niveles, son sin duda una señal positiva que refleja el alto valor que la sociedad asigna a la calidad de la educación en el país. Un sistema nacional de evaluación educativa basado en pruebas estandarizadas representa un esfuerzo muy importante de diagnóstico, inherentemente enfocado a la atención de necesidades y mejora del sistema educativo. Sin embargo, los objetivos específicos que se persiguen mediante el desarrollo de una prueba concreta pueden diferir según las particularidades del sistema educativo y las filosofías, prioridades, intereses, y perspectivas de los actores interesados. Esta expresión incluye en principio a alumnos y padres de familia; maestros, directivos y otros agentes escolares y organismos que los representan; políticos y autoridades federales, estatales y locales; expertos en educación, medición y estadística, política pública, y economía entre otros; y organismos diversos de la sociedad civil.

Las mejores prácticas internacionales indican que los usos previstos o deseados de una prueba no solo deben informar su diseño, sino que establecen los parámetros más relevantes a considerar al evaluar las características técnicas de cada uno de sus componentes. Consecuentemente, en cierta medida el presente apartado representa la base conceptual para todo el esfuerzo de validación que realizó el comité de especialistas. La validación de una prueba implica un esfuerzo permanente de colección y análisis de evidencias que documentan y sostienen las interpretaciones y usos propuestos, pero también implica tener conciencia de posibles usos y consecuencias no previstas o indeseables, y darles seguimiento.

Este apartado sintetiza los resultados de nuestro análisis de usos y consecuencias de las pruebas ENLACE-B, ENLACE-MS y EXCALE que se detallan en los reportes individuales de cada prueba. La noción de validez relativa a usos y consecuencias de las pruebas combina consideraciones teóricas, psicométricas y operativas; su aplicación a casos particulares se discute activamente en la literatura especializada. Nuestro análisis entiende la validez consecuencial de forma amplia como la forma en que se difunden los resultados de las pruebas, los usos que se hacen de dichos resultados, así como las consecuencias que ha traído consigo su utilización en el sistema educativo mexicano.

Desde la perspectiva de la política educativa, las consecuencias de la prueba son relevantes sin importar si corresponden a definiciones particulares de validez. Por otro lado, nuestro análisis se limita al uso de instrumentos en contextos específicos de política educativa, y no pretende ni puede ofrecer una evaluación o juicio cualitativo general sobre el impacto social de estas políticas, o su idoneidad en comparación con otras alternativas que pudieran implementarse.

En este capítulo se describen primero temas generales comunes que se derivan del análisis de usos y consecuencias de ENLACE-B, ENLACE-MS, y EXCALE, y que consideramos se aplican a las tres pruebas. Aunque algunos detalles específicos pueden diferir ligeramente entre pruebas, nuestro análisis sintético busca extraer temas y lecciones generales que juzgamos importante

informen una revisión de los procesos de desarrollo y uso de pruebas de gran escala en México. Como corolario a cada apartado, se hacen recomendaciones para el desarrollo de la siguiente generación de pruebas nacionales en México.

Conviene señalar que nuestro análisis no pretende ni puede establecer si ciertos usos de pruebas de gran escala son apropiados o deseables desde una perspectiva general de política pública. En cambio, sí se concibe desde la perspectiva técnica como una evaluación del grado en que los usos previstos en los manuales técnicos de las pruebas (y otros ampliamente documentados) se justifican con base en evidencias sólidas, tal como requieren las mejores prácticas internacionales en medición educativa.

Temas y consideraciones generales

Modelos lógicos de uso y fundamentación teórica y empírica

Un primer hallazgo que se aplica por igual a ENLACE-B, ENLACE-MS, y EXCALE, es que la documentación disponible para las tres pruebas no ofrece un marco conceptual sólido que sustente su desarrollo y características técnicas y asiente los parámetros para su posterior validación. Estas pruebas no se construyen a partir de modelos lógicos de uso que permitan operacionalizar con una especificidad adecuada los objetivos generales de uso que se prevén para cada una. Lo anterior conduce a objetivos y usos indeterminados que no se definen con precisión más allá de aserciones generales y abstractas, como detectar áreas de oportunidad y orientar la práctica pedagógica del docente; informar juicios de valor contextualizados para apoyar la toma de decisiones documentada; proveer información útil para el plantel y los profesores; o generar información diagnóstica para cada alumno.

La documentación de las tres pruebas no distingue claramente consecuencias, objetivos generales y específicos, mecanismos de operación y metas de distinto plazo para actores y niveles. Aunque están relacionados, estos conceptos son diferentes y corresponden a aspectos diversos del desarrollo y validación de una prueba. Es notorio que las tres pruebas buscan objetivos ambiciosos y prevén gran variedad de usos y consecuencias: para ENLACE-B se enlistan 18 usos o consecuencias que se espera se deriven de la prueba; para ENLACE-MS 12; y para EXCALE 10 . Estos usos cubren todos los niveles y actores educativos, desde estudiantes hasta maestros, directores, tomadores de decisiones y sociedad. Esta gama de usos no se concentra en un apartado del manual técnico respectivo, de manera que las tablas de usos y consecuencias contenidos en los reportes individuales de cada prueba fueron compilados por nosotros a partir de aseveraciones de distinto grado de especificidad que se ofrecen en distintas secciones de los manuales y, en algunos casos, incluso en documentos distintos al manual.

Desde luego la experiencia internacional ofrece ejemplos de pruebas que se usan para más de un propósito. Sobre esto se debe notar primero que, aunque en algunos casos el diseño de una prueba puede justificar usos múltiples, en muchos otros esto no es así, lo cual con frecuencia da lugar a problemas de validez estudiados extensamente en la literatura. Y segundo, la cantidad, variedad y alcance de usos que se prevé dar a las pruebas mexicanas son notables incluso en este contexto. En particular, los autores no conocemos casos de pruebas de gran escala que busquen informar a la vez la autoevaluación de los estudiantes, el diagnóstico y práctica pedagógica en el aula, la evaluación de programas y políticas, el monitoreo de resultados del sistema, y la rendición de cuentas a nivel de docentes, escuelas y subsistemas (en el caso de las

pruebas ENLACE). Lo anterior contraviene no solo las mejores prácticas internacionales, sino preceptos fundamentales de medición y evaluación en gran escala. Por otro lado, es importante recalcar que la alternativa no es simplemente administrar una multiplicidad de pruebas para que cada una cumpla un propósito distinto: la experiencia internacional reciente advierte sobre efectos negativos serios por el uso excesivo de pruebas, tanto para el aprendizaje y bienestar emocional de los alumnos, como para la práctica docente en el aula. El balance entre estos dos escenarios es complejo y requiere de un análisis bien fundamentado de las necesidades y prioridades del sistema educativo.

Aunque todos los escenarios abstractos y los objetivos generales de uso señalados para cada prueba son en principio deseables, si se consideran individualmente, en el marco de un sistema nacional de evaluaciones de aprendizaje de gran escala, se esperaría que cada uso previsto se asociara con inferencias, interpretaciones y acciones específicas por parte de actores particulares. Consistentemente notamos la poca fundamentación y argumentación lógica que describa cómo ocurrirá cada uno de los usos previstos; qué tipos de inferencias específicas pueden derivarse en cada nivel de información; qué usuarios, unidades o acciones están involucrados en cada caso; por cuál(es) mecanismo(s) específicos se producirán las consecuencias y efectos positivos esperados; y, por último, la evidencia teórica y empírica que respalda todos estos supuestos y predicciones.

En síntesis, se encontró un gran número de usos previstos de las pruebas, que involucran distintos tipos de información y de usuarios. Los usos propuestos se describen con grados de especificidad muy heterogéneos, unos claramente asociados a datos y usuarios particulares, otros solo de forma amplia y abstracta. No se establece una jerarquía de usos propuestos, ni se priorizan los objetivos que se siguen en relación con la información que se produce. El diseño de las pruebas no está claramente alineado con los usos que se proponen de éstas. En general, se evidencia la ausencia de modelos lógicos que permitan alinear los usos previstos de las pruebas a aspectos técnicos de su diseño y a las fuentes de evidencia teórica y empírica necesarias para justificar tales usos.

De lo anterior se desprende la recomendación de que la siguiente generación de pruebas de rendimiento para el sistema educativo mexicano debe partir de un modelo lógico conceptual y empíricamente fundamentado, que detalle: a) usos previstos basados en necesidades y prioridades de distintos grupos de usuarios; b) contexto y criterios claros en relación con los recursos necesarios y disponibles, el proceso de desarrollo y aplicación de pruebas, la diseminación de resultados, y la capacitación de usuarios, entre otros; c) efectos y consecuencias esperadas a corto, mediano y largo plazos; y, d) mecanismos y criterios de seguimiento para determinar el grado en que ocurren estos usos previstos en la práctica.

Un ejemplo: usos diagnósticos que informen la práctica pedagógica

Un caso en que se concreta con nitidez la falta de un modelo conceptual es el del uso diagnóstico que se espera informe y mejore la práctica docente en aula.

Aunque difieren en diseño, estructura y contenido, las tres pruebas destacan beneficios que se derivarían del aprovechamiento de los resultados por parte de maestros y directivos. La documentación incluye declaraciones y objetivos ambiciosos basados en usos diagnósticos que daría el docente, pero se ofrece poca o nula justificación teórica, empírica, e incluso lógica, que los respalde. Esto revela desconexión importante entre usos propuestos y características técnicas de la prueba.

Por ejemplo, aunque se hacen repetidas llamadas a que los docentes hagan uso pedagógico de la prueba, no se ofrece evidencia de cómo podrían derivarse acciones de mejora de tal uso diagnóstico, que asume: 1) precisión suficiente de los puntajes para justificar inferencias y acciones con estudiantes, grupos o escuelas específicas; 2) cobertura similar y suficiente en todos los temas del currículo para permitir diagnósticos relevantes; y 3) sensibilidad de la prueba a diferencias en la calidad (o incluso cantidad) de la instrucción.

Las pruebas en gran escala, particularmente censales y de formas fijas, presentan serias limitaciones en este sentido, dado que solo pueden aspirar a cubrir un pequeño número de temas con poca profundidad, con precisión insuficiente a nivel individual, y tienden a ser insensibles al currículo y la práctica docente.

El diagnóstico correcto, preciso y oportuno es necesario pero no suficiente para la mejora. Los docentes no reciben resultados hasta el inicio del ciclo escolar siguiente, y no para sus grupos sino agregados por escuela y grado. La documentación de las pruebas presta poca o nula atención al tipo de intervenciones o mecanismos mediante los que se derivarían las mejoras esperadas, y no ofrecen lineamientos de uso que los docentes deberán seguir cuando los resultados sugieren áreas de debilidad en sus estudiantes.

Como era predecible, los maestros no están utilizando las pruebas para el trabajo en aula. Paradójicamente esto ocurre en un contexto donde se multiplican esfuerzos en todos los niveles del sistema educativo para promover la mejora de la práctica docente (desde materiales impresos, a programas amplios de apoyo y desarrollo profesional). Aunque desde luego revisten crucial importancia para la mejora, es importante notar que estos esfuerzos podrían existir con total independencia de las tres pruebas nacionales en cuestión.

La siguiente generación de pruebas debe hacer una revisión cuidadosa de los usos pedagógicos que se proponen de la información que generan. Debe asegurarse que se cumplen los requerimientos técnicos necesarios si se busca diagnosticar fortalezas y debilidades a nivel individual y grupal, describir los mecanismos de uso de resultados que se espera de los docentes (y ofrecer evidencia de cómo se traducen en resultados específicos en la práctica), y desarrollar una variedad de mecanismos y materiales de apoyo a los docentes para promover estos usos.

Acceso a reportes e información que facilite usos correctos

La SEP y el INEE llevan a cabo actividades para difundir y promover el acceso a los resultados de las pruebas que desarrollan. Los resultados de ENLACE se difunden principalmente por medio de la página web de la SEP y, en el caso de ENLACE-B, con materiales para padres, reuniones con autoridades, documentos, folletos y carteles informativos, difusión a medios y grupos de opinión, etcétera. EXCALE ofrece informes escritos, presentaciones públicas, materiales informativos, talleres, redes sociales y herramientas interactivas en una página (Explorador y Corpus EXCALE). Sin embargo se encontraron carencias técnicas e inconsistencias entre la información que se ofrece y los objetivos que se siguen, lo que explica que estos esfuerzos no estén teniendo el impacto esperado.

En el caso de ENLACE es indudable que los resultados despiertan fuerte interés entre muchos actores. Por ello es de destacarse la poca atención que han recibido los informes de resultados individuales en el portal web por parte de los usuarios principales: los estudiantes y sus familias. Esto refleja en parte la falta de alineación entre usos propuestos, formatos y tiempos de acceso a la información. Persiste la duda sobre si la mayoría de los padres, e incluso maestros, maneja adecuadamente la web para acceder a la información e interpretarla bien. Las estadísticas de acceso son muy gruesas y no permiten un diagnóstico de quiénes están usando la información y cómo. Además, el uso auto-formativo se dificulta porque los resultados de ENLACE-B se reciben al final del ciclo escolar, y los de ENLACE-MS después de que los estudiantes se gradúan de bachillerato. El formato de los reportes es claro y el lenguaje accesible, pero no dan apoyo para interpretar los resultados; no existen guías de texto, videos, animaciones o elementos gráficos atractivos y eficientes para ejemplificar el uso correcto de los puntajes, ni se detalla la interpretación y uso de los resultados por ítem. Es notorio que el análisis diagnóstico de respuestas incorrectas se limita a repetir textualmente el estándar o contenido que el estudiante, en teoría, no domina.

En cuanto a resultados agregados por estado y subsistema, la SEP presenta cada año los resultados de ENLACE-B en una conferencia de prensa y publica bases de datos con puntajes crudos agregados por escuela en su página. Los resultados no se acompañan de un reporte analítico que profundice el análisis y ayude a contextualizar los resultados, o al menos informe sobre características técnicas de las pruebas y la precisión de los datos. Liberar tal volumen de datos crudos sin contexto o atadura conceptual, genera un vacío de información en el que medios de comunicación y organizaciones de la sociedad civil han pasado a jugar un papel preponderante, con la publicación de listas o rankings de escuelas o entidades federativas.

En contraste con lo anterior, los reportes de resultados e informes temáticos que publica el INEE se establecieron inicialmente como recursos de referencia para autoridades, académicos y medios de comunicación. Los informes presentan los resultados de forma rica y contextualizada, lo que minimiza el riesgo de inferencias o usos simplistas. Los reportes se ajustan a las mejores prácticas internacionales al ofrecer estimados de error estándar que reflejan la precisión de indicadores y comparaciones. No obstante lo anterior, el volumen y detalle de información puede ser excesivo para un usuario sin conocimientos estadísticos.

Un mensaje de la Presidenta de la Junta de Gobierno del INEE confirma que las autoridades estatales tienen dificultad para entender la información de informes de EXCALE. Es notorio también el retraso de los informes en años recientes, con lapsos de hasta tres años entre aplicación y publicación. Por su parte, las herramientas interactivas desarrolladas tienen limitaciones.

El Explorador de EXCALE pretende apoyar usos pedagógicos de la prueba por los docentes, pero la información que ofrece (porcentajes agregados de aciertos en docenas de contenidos temáticos por grado y materia) no es suficiente para una reflexión pedagógica genuina y útil.

Nuestras entrevistas con entidades estatales indican que tratan de complementar el acceso a los resultados de las pruebas con sistemas propios de difusión y consulta de resultados, pero sus esfuerzos se enfocan principalmente a los resultados de ENLACE-B. En la mayoría de las entidades los esfuerzos de difusión y uso de EXCALE son limitados o inexistentes.Llama la atención la ausencia en todas las pruebas de un esfuerzo de monitoreo y diagnóstico del acceso, comprensión y uso de los reportes que informe su adaptación y mejora para apoyar a los usuarios objeto. Las estadísticas de uso son pocas y muy gruesas, sin detalle suficiente para crear un perfil del usuario que utiliza los resultados.

Se encontraron limitaciones importantes en el acceso a la información de las pruebas ENLACE y EXCALE. Por distintos motivos la información que se desprende de ellas no se hace llegar de manera consistente, oportuna, y efectiva a los usuarios. Los alumnos reciben los resultados cuando el ciclo escolar ha terminado, y los docentes al inicio del siguiente y agregados al nivel de escuela y curso. Aunque hay diferencias entre pruebas, e incluso entre estados, en general no se ofrece información de apoyo que condense clara y detalladamente el contexto técnico y sustantivo necesario para la apropiada interpretación de los resultados en cada nivel de agregación. Tampoco se describen suficientemente las limitaciones técnicas de los estimadores en relación con la precisión o cobertura de puntajes, o interpretaciones incorrectas. Más allá de estadísticas gruesas que reflejan bajas tasas de acceso a los datos, no existen esfuerzos sistemáticos de monitoreo que permitan un diagnóstico puntual de los tipos de usuarios que acceden a los reportes individuales y agregados, y los usos que dan a la información.

La siguiente generación de pruebas deberá diseñar mecanismos y formatos de comunicación (textual, gráfico, electrónico) para asegurar que los usuarios reciban resultados en tiempos y formas que posibiliten los usos propuestos. En general se requiere que los mecanismos de comunicación estén directamente alineados a modelos lógicos de uso de los resultados, y que consideren las necesidades y capacidad de cada grupo de usuarios. Los reportes deben ofrecer información técnica suficiente que permita interpretar y contextualizar adecuadamente los resultados. En particular deben explicitarse las limitaciones en las inferencias de alto impacto en función de la precisión de los puntajes (i.e. el error estándar). Finalmente, debe monitorearse el acceso y estudiar en profundidad la comprensión y utilidad de los reportes por los usuarios intencionales, para informar el mejoramiento continuo de cada herramienta comunicacional.

Fomento de la capacidad de interpretación de los resultados

Para ENLACE y EXCALE hay evidencia de que se ha intentado apoyar el desarrollo de capacidades de procesamiento e interpretación de los resultados. En ambos casos se realizan talleres de difusión y uso destinados a autoridades educativas, equipos de supervisión escolar, docentes y directores. El INEE ha hecho talleres con periodistas, así como para investigadores y estudiantes, sobre requerimientos técnicos y análisis de las bases de datos de PISA y EXCALE.

Aunque estos esfuerzos son deseables y parte del modelo lógico de uso e impacto de las pruebas, su profundidad y efectividad ha sido variable. La literatura sugiere claramente que la efectividad de esfuerzos de desarrollo profesional basados en intervenciones breves o sugerencias genéricas tiende a ser limitada cuando no se acompañan de supervisión y asistencia más duradera y próxima a la práctica cotidiana. Además, es claro que la profundidad, alcance y eventual efectividad de estos esfuerzos depende en gran medida de su interacción con la capacidad instalada en cada entidad federativa u organismo. En al menos una de las entidades visitadas se encontró que se ha desarrollado capacidad regional, e incluso en muchas escuelas; en otras dos se detectó preocupación y esfuerzo por desarrollar mayor capacidad para asistir a usuarios de ENLACE-B en el análisis e interpretación de la información, con propósitos formativos. Otras tres entidades no reportaron que se implementen actividades sistemáticas para fomentar capacidad de interpretación y uso de los resultados por parte de maestros, directivos o autoridades.

Es deseable revisar en profundidad la efectividad de los esfuerzos de apoyo y capacitación para uso de resultados de las pruebas que realizan tanto la SEP y el INEE, como los estados. Se requiere un esfuerzo sostenido y coordinado por parte de estas instancias para desarrollar la capacidad técnica en evaluación educativa en México. La capacidad evaluativa que se requiere es muy considerable en términos de la amplitud de las áreas que muestran desarrollo incipiente en el país; debe buscarse el desarrollo de una cultura de la evaluación y medición concebida más allá de las pruebas en gran escala, que eduque a actores clave en ideas como validez inferencial, precisión y evidencias de validez, entre otros. El esfuerzo es amplio también en términos del volumen de organismos, instituciones y personas que en teoría pretendería alcanzarse. A pesar de las dificultades que representa, se requiere un esfuerzo de este tipo para producir avances en el grado de alfabetización en la evaluación y, por lo tanto, en las expectativas sociales de los niveles de calidad requeridos de los sistemas de evaluación.

Se encontró una variedad de esfuerzos a nivel nacional y estatal que buscan desarrollar la capacidad técnica instalada para promover usos apropiados y análisis sofisticados de la información que producen las pruebas. Sin embargo, estos esfuerzos no evidencian coordinación entre organismos y entidades federativas o al interior de éstas. Los esfuerzos más consistentes alrededor de la prueba EXCALE perdieron protagonismo ante la atención desproporcionada que generó ENLACE-B en años recientes. Existe una gran variabilidad entre estados en términos de la capacidad técnica instalada, y la existencia y alcance de esfuerzos dirigidos a desarrollarla.

La siguiente generación de pruebas debe incluir en su diseño y plan de operación, el desarrollo de procesos de capacitación sistemáticos y sostenidos en relación con la interpretación y el uso adecuado de sus resultados. De manera más general, estos esfuerzos deben ser comprensivos e incluir la promoción de la capacidad evaluativa en todos los niveles del sistema educativo, de maestros, comunicadores y tomadores de decisión a nivel estatal y federal.

Uso de los datos para investigación y evaluación de programas

Los manuales técnicos de las tres pruebas hacen mención de los investigadores como usuarios importantes que realizan análisis profundos de los resultados para la generación y evaluación

de conocimientos e hipótesis científicas y, en su caso, para el seguimiento y evaluación de programas y esfuerzos de política educativa. En la práctica, sin embargo, los investigadores tienen acceso inconsistente y limitado a los resultados de las pruebas. No se evidencia una estrategia de fomento y apoyo a este tipo de usos por parte de SEP y, en el caso del INEE, la estrategia es incipiente y se ha implementado de forma inconsistente.

Los resultados de ENLACE-B solo están disponibles en agregado por escuela y entidad, lo que limita al investigador a análisis rudimentarios y poco informativos. Existen solo ejemplos aislados de investigaciones que usaron las bases de datos a nivel alumno para análisis de tendencias y factores asociados, y otras que abordan temas de calidad y equidad educativa, con herramientas estadísticas sofisticadas. Aunque la revisión de la literatura que se realizó podría haber omitido algunos otros estudios, es claro que el volumen disponible refleja un grado de involucramiento de la comunidad de investigación mucho menor al deseable en el caso de un programa nacional de la envergadura y relevancia de ENLACE-B.

La literatura, reportes de organismos especializados, y la información obtenida de los estados, ofrecen ejemplos aislados donde ENLACE-B se utiliza como indicador de impacto principal o único en evaluaciones de políticas y programas a nivel estatal o federal. Sin embargo, es interesante notar que de 25 evaluaciones de impacto de programas de la SEP reportadas en 2012 por el Consejo Nacional de Evaluación (CONEVAL), solo cuatro usaron resultados de ENLACE-B como indicador (Programa Escuelas de Calidad, Escuelas de Tiempo Completo, y Asesor Técnico Pedagógico). Esto refleja limitaciones en la disponibilidad de ENLACE-B en ciertos grados o materias, pero también dificultad de acceso a resultados desagregados que son necesarios para una evaluación de impacto en el aprendizaje. Finalmente, una prueba nacional censal tiene limitaciones serias como variable dependiente en evaluaciones de impacto, relacionadas con la precisión de los puntajes, la cobertura de temas, y la sensibilidad a la instrucción y a otras intervenciones.

El INEE ha buscado fomentar entre los investigadores el uso de los datos de EXCALE de varias formas. El Banco de Indicadores Educativos es una herramienta robusta que permite consultar y almacenar los archivos de donde se derivan los reportes que publica el Instituto. También se ha buscado promover el uso de las bases a nivel del alumno, para realizar análisis sofisticados que enriquezcan los reportes estadísticos nacionales que regularmente ofrece el Instituto.

Este esfuerzo se ha traducido en una serie de trabajos que pueden clasificarse en tres categorías: 1) proyectos realizados al interior del INEE por investigadores y personal propio; 2) una veintena de estudios especiales comisionados por el INEE a especialistas externos nacionales o internacionales, o derivados de colaboraciones entre éstos e investigadores del Instituto (algunos disponibles en la serie Cuadernos de Investigación del propio INEE, otros internamente como documentos de trabajo, y otros en revistas y libros especializados); y 3) menos de veinte estudios publicados por investigadores o instituciones externos al INEE.

La evidencia, por tanto, no apunta a un uso extendido de las bases de EXCALE por parte de especialistas. Es notoria también la ausencia de trabajos de tesis por estudiantes de postgrado, un área de oportunidad importante que un organismo de este tipo típicamente trataría de promover. Finalmente, las herramientas Corpus y Explorador EXCALE reflejan el espíritu de generación y uso de información que se espera de un organismo como el INEE, pero por el momento presentan problemas importantes de concepción e implementación que limitan su utilidad para los usos y usuarios previstos. La herramienta Corpus EXCALE busca apoyar el diagnóstico de habilidades de escritura de los alumnos permitiendo consultar muestras de textos

por grado, estado, modalidad y sexo. Sin embargo, el formato de acceso tiene limitaciones serias (pueden consultarse imágenes de texto una a la vez, lo que limita su utilidad para análisis extenso o detallado) y la evidencia anecdótica existente sugiere que el uso de esta herramienta ha sido mínimo.

Todo lo anterior refleja la necesidad de establecer mecanismos claros y eficientes que no solo permitan, sino que promuevan, apoyen, e incluso incentiven el acceso y uso de los datos por parte de investigadores calificados. El tipo de análisis que se necesita es de alta complejidad, lo que hace deseable un esfuerzo adicional para desarrollar materiales que informen consistentemente a los investigadores sobre características psicométricas y estadísticas de la prueba y requerimientos técnicos necesarios para analizar los datos. Sería deseable ofrecer talleres especializados para investigadores que propicien el uso de los datos de la prueba.

No hay mecanismos claros y eficientes para facilitar el acceso a bases de datos por parte de investigadores, ni una estrategia robusta y coordinada de promoción, incentivo y apoyo técnico para usuarios de esas bases. Tampoco hay programas para desarrollar la capacidad de uso entre estudiantes de postgrado. No se han desarrollado bases de datos longitudinales especializadas, fundamentales para el apoyo de la investigación. Como resultado, el volumen de investigación existente que emplea las bases de datos de ENLACE y EXCALE está lejos de los niveles deseables. Se encontraron solo ejemplos aislados de uso de las bases de datos a nivel individual para investigación y evaluación de programas.

La siguiente generación de pruebas deberá asignar alta prioridad al desarrollo de mecanismos robustos y claros para promover, incentivar y apoyar el uso de las bases de datos para investigaciones que amplíen el conocimiento sobre el sistema educativo mexicano y los factores que afectan el desempeño de alumnos, docentes y escuelas. Estos esfuerzos deberán incluir mecanismos que protejan la privacidad de estudiantes, docentes y escuelas individuales. Deberá asignarse alta prioridad a planes y programas de desarrollo de capacidad entre investigadores y estudiantes de postgrado por medio de programas de becas y apoyo técnico. Es importante considerar explícitamente el papel que puede jugar el nuevo sistema nacional de evaluación en la evaluación del impacto de programas en todos los niveles del sistema educativo. Deben destinarse recursos y esfuerzos específicos para desarrollar un programa de apoyo técnico que permita acceso a bancos de reactivos liberados y asistencia a evaluadores, a fin de construir pruebas apropiadas para analizar el impacto de programas.

Seguimiento de usos y consecuencias previstas e imprevistas

Nuestro análisis revela esfuerzos muy limitados por recolectar y dar seguimiento a los usos de las tres pruebas. Esto resulta en un vacío significativo de información que afecta directamente a los esfuerzos por validarlas y mejorarlas, puesto que no puede saberse si los resultados esperados se dieron en la realidad, cómo y por qué se lograron o no, y si se produjeron resultados distintos a los esperados. Un sistema de estas dimensiones no puede limitar sus esfuerzos de seguimiento a impresiones casuísticas, estudios aislados, o reportes de prensa; se requiere

un esfuerzo sistemático y sostenido de seguimiento de usos y consecuencias, sobre todo en un contexto de consecuencias de alto impacto.

Una vez que se documenta evidencia sobre usos y consecuencias previstos, deben establecerse mecanismos para recolectar información sobre usos no previstos y consecuencias negativas no anticipadas. La información recabada sugiere que no se están dando algunos de los usos previstos y sí algunos imprevistos. Entre los primeros pueden mencionarse los de carácter pedagógico por parte de los docentes, o una variedad de usos propuestos que involucran a estudiantes, familias y escuelas. Entre las consecuencias no previstas a monitorear, pueden mencionarse la inflación de puntajes a través del tiempo; el estrechamiento curricular; las prácticas de entrenamiento para la prueba; y los reconocimientos individuales y grupales con base en los puntajes obtenidos.

La relevancia ética de estas consideraciones es mayúscula; aunque los desarrolladores no pueden impedir todos los usos injustificados de una prueba, sí deben implementar mecanismos robustos que faciliten los usos previstos y monitorear y minimizar los imprevistos y negativos que se consideren de mayor riesgo o seriedad.

Se encontraron esfuerzos muy limitados o nulos por hacer seguimiento sistemático y sostenido de los usos y consecuencias que tienen las pruebas en la práctica entre los distintos grupos de usuarios. Hay poca evidencia de esfuerzos de seguimiento enfocados a determinar si se han producido las consecuencias previstas; aún en menor medida se da seguimiento para detectar evidencias de consecuencias imprevistas, positivas o negativas. Este vacío de información limita significativamente los esfuerzos de mejora y validación continua, necesarios en un sistema de pruebas nacionales de alto impacto y/o visibilidad.

La siguiente generación de pruebas deberá establecer mecanismos robustos de monitoreo para determinar el grado en que se producen usos y consecuencias previstas e imprevistas, y apoyar la mejora de instrumentos y procedimientos. Es importante tener mecanismos de colecta de información sobre usos e iniciativas relacionadas con la prueba por parte de estados y subsistemas. Igualmente se debe obtener información sobre usuarios que acceden a los portales web. Es muy importante hacer seguimiento cercano de consecuencias para la práctica docente en aula, tanto previstas como imprevistas, en particular, estrechamiento curricular y enseñanza dirigida a la prueba. Es también importante desarrollar colaboraciones permanentes con centros de investigación para la validación continua de las pruebas, y crear bases de datos sobre trabajos y artículos elaborados en este ámbito. Finalmente, deberán existir lineamientos y mecanismos de información y reacción cuando se detectan usos injustificados y perniciosos de las pruebas.

Consideraciones particulares sobre cada prueba

Sobre ENLACE de educación básica

Es claro el interés que existe alrededor de la prueba ENLACE-B, que se manifiesta en la gran variedad de usos, esfuerzos, programas e intervenciones en todos los niveles, desde autoridades federales y estatales, hasta escuelas, maestros y padres, así como organizaciones de la sociedad

(en palabras de un entrevistado "uso generalizado pero no sistematizado"). Los resultados de la encuesta aplicada a autoridades educativas estatales que se realizó para el presente estudio, confirman la gran variedad de programas que buscan utilizar los resultados de la prueba para informar esfuerzos de mejora de docentes o escuelas. Tres de cada cuatro estados reportan como prioridad el uso diagnóstico de ENLACE-B para informar esfuerzos de autoevaluación en las escuelas; más de la mitad mencionó como objetivo prioritario informar la práctica docente, mientras que la mitad reportó orientar la capacitación de profesores. Existe interés en usar las pruebas para informar la investigación y la evaluación de impacto de programas, pero hasta la fecha el número de estudios de este tipo dista mucho del esperable en sistemas de estas dimensiones. Como se mencionó anteriormente, esto refleja el poco interés que se ha prestado a este tipo de usos, lo que es lamentable dado que en principio esto no requiere de gran inversión de tiempo o recursos.

Dado que ENLACE-B se autodefine como prueba formativa, no debería utilizarse para evaluar directamente a alumnos o docentes. Esto contrasta con el uso generalizado y explícito de los puntajes para entregar premios a alumnos y escuelas de altos puntajes, y como parte de la evaluación de maestros dentro del programa federal Carrera Magisterial. Es particularmente preocupante esta discordancia entre las características de la prueba y el uso que se hace de ella porque muchas veces son las propias autoridades federales y estatales las que utilizan los resultados de forma injustificada (por ejemplo, los ordenamientos de escuelas por puntaje bruto en el sistema de reporte de resultados de la SEP).

Si se contrasta con la estabilidad de usos propuestos que reflejan los manuales, parece evidenciarse una tendencia de corrupción o inflación de funcionalidad, donde los usos de un instrumento se extienden sin que ello refleje un cambio de la misión y el diseño de la prueba. Este tipo de inercia inflacionaria de uso tiende a corromper el indicador (lo que se conoce como Ley de Campbell), limita el uso diagnóstico que inicialmente se buscaba, y puede traer efectos no deseables si se combinan con incentivos, sanciones, y consecuencias de alto impacto.

La evidencia recogida y un informe reciente de la OCDE al respecto, sugieren como consecuencia importante el uso significativo del tiempo del aula destinado a la instrucción enfocada a la prueba, así como incentivos para que alumnos y maestros busquen obtener altos puntajes, lo que ha generado prácticas y dinámicas que ponen en peligro la integridad de los resultados. El reporte de la OCDE concluye que "los efectos no intencionales de ENLACE-B parecen ser importantes. A pesar de la gran cantidad de datos recolectados, hasta qué punto se utilizan para un propósito formativo no está claro". Al evaluar las consecuencias del uso de ENLACE-B es importante considerar la dimensión ética de una prueba de alto impacto.

Una máxima de la medición en educación es que solo deberán tomarse decisiones o acciones que afectan a personas o grupos cuando existe evidencia sólida de la validez y confiabilidad de las medidas que lo justifique. Por tanto, la necesidad de gran claridad en la definición de lo que mide o no una prueba, o los usos que se consideran justificados o injustificados (y ciertamente lo que se considera evidencia técnica sólida) no se deriva únicamente de un concepto de rigor técnico, sino también de uno de probidad ética. La promoción del uso de los datos de las pruebas conlleva el riesgo de usos injustificados, pero debe buscarse minimizarlos, sobre todo por parte de la propia autoridad.

Sobre ENLACE de educación media superior

Nuestro análisis encontró un sistema robusto de desarrollo de pruebas basado en la experiencia del organismo desarrollador. Sin embargo, como en ENLACE-B, la atención que se presta a la calidad técnica de los ítems no se extiende a otros aspectos de las pruebas. El vacío lógico en cuanto a usos formativos por parte del estudiante y su familia es particularmente pronunciado en esta prueba, dado que se aplica al final de la educación media superior, cuando los alumnos están por graduarse, y no se ofrece apoyo para informar reflexiones y esfuerzos de mejora por parte de los alumnos. Por tanto, no sorprenden las estadísticas de acceso y otras evidencias empíricas que reflejan un bajísimo interés (alrededor de 5%) de quienes en teoría son los usuarios principales de la prueba. De manera similar los mecanismos de comunicación de resultados a docentes impiden los usos formativos, ya que estos reciben la información al inicio del año escolar siguiente y agregada al nivel de la escuela, con lo que no es posible conocer el desempeño último de los alumnos que estuvieron bajo su supervisión.

Encuestas y entrevistas con autoridades estatales realizadas para este estudio ofrecen además evidencia amplia de usos no previstos para los que no hay justificación técnica o que expresamente se identifican como incorrectos en el manual técnico, como comparaciones de alto impacto, competencia, y preparación de alumnos dirigida a aumentar los puntajes de la prueba. En algunos casos las autoridades federales y estatales están promoviendo medidas y programas que contravienen directamente el espíritu y la letra del manual técnico, como programas de preparación de los alumnos para la prueba, o la publicación de rankings de escuelas y la asignación de reconocimientos e incentivos a estas. Previsiblemente estos procesos se están traduciendo en patrones de inflación de puntajes.

Al igual que en ENLACE-B, en ENLACE-MS no se promueve el uso de las bases de datos por la comunidad académica. Más aún, la prueba se desarrolla y opera con cierto hermetismo por parte del desarrollador, que incluso se mostró inicialmente reticente a compartir información para este reporte. Relacionado con lo anterior, llama la atención la diferencia de estructura entre ENLACE-B y MS, ya que uno es desarrollado al interior de la SEP y el otro se subcontrata a un organismo especializado. Aunque esto puede explicarse en términos prácticos, esta separación puede afectar la coherencia del sistema y dificultar la propagación de procesos de mejora necesarios para ambas pruebas.

Sobre EXCALE

La encuesta y entrevistas realizadas arrojaron poca evidencia de uso sistemático a nivel estatal para informar políticas educativas y procesos de mejora. Se constató difusión poco activa de resultados a supervisores, maestros y padres de familia en la mayoría de entidades federativas. En consecuencia muchos maestros y familias simplemente no están familiarizados con la prueba, no conocen sus resultados, y por supuesto no los usan para ningún propósito discernible. Los usos pedagógicos que se pretende den los docentes a los resultados de EXCALE en principio no parecen corresponder con los de una prueba de diseño matricial, que ofrece solo resultados agregados a nivel regional. Los materiales de apoyo pedagógico que ha desarrollado el INEE con ayuda de expertos, son de alta calidad y pueden ser valiosos para los maestros, pero no representan un uso directo alineado al diseño de la prueba, ni constituyen un uso pedagógico de ésta por parte de los docentes en el sentido tradicional, y podrían haberse desarrollado de manera independiente.

Aún en el caso de las autoridades estatales, las respuestas con frecuencia reflejan cierta confusión sobre los objetivos de EXCALE y la diferencia entre esta prueba y ENLACE-B. A nivel federal se encontró también un uso limitado y decreciente; aunque las áreas de la SEP a cargo del currículo estuvieron involucradas en el desarrollo de las especificaciones de EXCALE, e inicialmente usaron los resultados para esfuerzos de difusión y mejora, en años recientes la tendencia ha sido clara y sugiere que la prueba es cada vez menos utilizada. Es evidente que EXCALE ha perdido visibilidad y prominencia debido a percepciones, justificadas o no, entre diversos actores, sobre el valor que agrega a la evaluación tomando en cuenta la existencia de ENLACE y PISA.

Varios factores del diseño y operación de EXCALE han contribuido a reducir su papel en el debate educativo en años recientes. Por una parte, su diseño matricial no ofrece resultados por escuela, lo que impide uno de los usos más mediáticos de la información. Por otra, el retraso de los informes finales ya mencionado (lapsos de tres años entre aplicación y publicación) ha afectado la percepción sobre la relevancia y utilidad de la prueba para los usos informativos globales que se prevén. Algunos actores reportaron que asignan mayor relevancia a la prueba PISA para un diagnóstico amplio del sistema educativo mexicano, por la posibilidad de hacer comparativos con otros países y por su énfasis en medir competencias para la vida no atadas a un currículo específico. Finalmente, no se ha promovido y facilitado de forma extensa y sistemática el uso de las bases de datos de EXCALE por parte de investigadores, lo que es un eje fundamental de aprovechamiento de una prueba de este tipo.

A pesar de las limitaciones de EXCALE, nuestro estudio también es claro en señalar el gran impacto del trabajo del INEE en general. Sus esfuerzos han tenido a todas luces un impacto significativo y positivo en la cantidad y, sobre todo, calidad de los trabajos de evaluación educativa que se realizan en las entidades federativas. En este sentido, tanto los informes de resultados que publica el INEE, como el proceso mismo de desarrollo de las pruebas EXCALE, están teniendo impacto y beneficio importante al modelar y promover el trabajo riguroso de evaluación a nivel local.

Conclusiones

Los sistemas de evaluación en gran escala pueden buscar objetivos diferentes y dar visiones distintas del sistema educativo desde perspectivas pedagógicas, filosóficas e incluso políticas. Sin embargo, la concreción de una visión del sistema educativo en uno de evaluación de características específicas, puede y debe evaluarse contrastando los objetivos que se buscan con los que es factible conseguir desde un punto de vista técnico y la evidencia empírica que representa su funcionamiento en la práctica. La información que se presenta en los apartados anteriores de este reporte genera dudas sobre aspectos importantes de la calidad técnica de ENLACE-B, ENLACE-MS y EXCALE, y tiene implicaciones importantes para el diseño de la próxima generación de pruebas nacionales.

En principio es de esperarque el uso correcto y generalizado de una prueba y, por tanto, su efectividad para los propósitos que persiga, dependerá en gran medida de: 1) la legitimidad que pueda adquirir y mantener a los ojos de los usuarios principales; 2) la capacidad de éstos para usar de manera adecuada la información que ofrece; y 3) los apoyos disponibles para respaldar ese uso y transformarlo en cambios sostenibles. Aunque no es realista esperar que una prueba nacional tenga credibilidad unánime y desde el inicio se use eficazmente en cada ámbito, el órgano desarrollador sí puede y debe aspirar a producir pruebas cuya solidez y

transparencia técnica le permitan fomentar actitudes positivas y de confianza por parte de los usuarios, extender los usos efectivos previstos de los resultados, y minimizar o evitar propiciar usos injustificados.

Esto es importante porque, en la práctica, diseñar un sistema de pruebas que cumpla más de un objetivo de forma apropiada representa un incremento significativo en términos de complejidad técnica y costo. La experiencia internacional y el sentido común sugerirían, por tanto, guardar un escepticismo saludable ante un sistema de evaluación que ofrezca cumplir objetivos de distinto tipo mediante un único instrumento o procedimiento de evaluación. Es esperable que sea difícil para un desarrollador aportar elementos probatorios que muestren que el sistema, además de un monitoreo descriptivo de tendencias nacionales o regionales, ofrece información pedagógica relevante para el docente en aula, bases sólidas para la rendición de cuentas de maestros, escuelas, y sistemas, y referencias precisas para evaluar la efectividad de programas y políticas a nivel local y nacional.

El desarrollo de pruebas de alta calidad técnica, basadas en las mejores prácticas internacionales y respaldadas por comités técnicos especializados, es por tanto condición necesaria pero insuficiente para asegurar la efectividad de un sistema de evaluación, especialmente en contextos de usos y consecuencias de alto impacto. Igualmente importante es el rigor y la transparencia con que se documentan y comunican tanto las características técnicas, como los usos y consecuencias de las pruebas. En este sentido, además de capacidad técnica, es importante la participación y compromiso de los implicados en el proceso de desarrollo, aplicación, y utilización de las pruebas, y el desarrollo de capacidades por parte de usuarios para que puedan tomar un rol de consumidor crítico, consciente de los alcances y limitaciones de la información que se deriva de la evaluación.

Este tipo de tensiones y complejidades conceptuales, técnicas, y logísticas inherentes a los sistemas de pruebas en gran escala, no son siempre evidentes y, por tanto, es importante explicitarlas desde el diseño, y transmitirlas a los usuarios, ofreciendo un análisis realista de objetivos y prioridades respecto de los usos propuestos de la prueba y los efectos positivos que pueden derivarse de ella. De no hacerlo así, el peligro que se corre es claro: el sistema resultante podría no cumplir adecuadamente en la práctica ninguno de los objetivos que en teoría se buscan, y multiplicar el riesgo de consecuencias negativas en menoscabo del esfuerzo inicial de mejora.