Documentos técnicos

Manual técnico del Plan Nacional para la Evaluación de los Aprendizajes PLANEA 2015.

sexto de primaria y tercero de secundaria

Documentos rectores

Instituto Nacional para la Evaluación de la Educación
México

Manual técnico del Plan Nacional para la Evaluación de los Aprendizajes
PLANEA 2015. Sexto de primaria y tercero de secundaria
Primera edición, 2018

D. R. © Instituto Nacional para la Evaluación de la Educación
Barranca del Muerto 341, Col. San José Insurgentes,
Del. Benito Juárez, C. P. 03900, Ciudad de México

La elaboración de este documento estuvo a cargo de la Unidad de Evaluación del Sistema Educativo Nacional. El contenido, la presentación, así como la disposición en conjunto y de cada página de esta obra son propiedad del INEE. Se autoriza su reproducción por cualquier sistema mecánico o electrónico, para fines no comerciales. Cítese de la siguiente manera:

INEE (2018). *Manual técnico del Plan Nacional para la Evaluación de los Aprendizajes PLANEA 2015. Sexto de primaria y tercero de secundaria*. México: autor.

Índice

Índice de tablas, figuras y cuadros

Tablas

Figuras

Cuadros

Presentación

Por mandato constitucional, y en su calidad de organismo público autónomo, el Instituto Nacional para la Evaluación de la Educación (INEE) se ocupa de realizar diversas acciones para evaluar la calidad de la educación obligatoria en México y para difundir sus resultados. Con este propósito se diseñan y aplican evaluaciones de los componentes, de los procesos y de los resultados en la educación básica y media superior, tanto pública como privada, con el objetivo de ofrecer información que permita tomar decisiones dirigidas a garantizar el derecho a una educación de calidad.

Una de las fuentes principales de información la constituyen las evaluaciones de los aprendizajes clave de los estudiantes que, a partir del ciclo escolar 2014-2015 y en coordinación con la Secretaría de Educación Pública (SEP), se llevan a cabo mediante el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA).

PLANEA está integrado por un conjunto de pruebas diseñadas para conocer la medida en que los estudiantes logran el dominio de una serie de aprendizajes clave del currículo en diferentes momentos de su educación obligatoria (INEE, 2016a). En 2015, las pruebas se organizaron en tres modalidades: la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN), la Evaluación del Logro referida a los Centros Escolares (ELCE) y la Evaluación Diagnóstica Censal (EDC).[1] Estos instrumentos se distribuyen con distinta periodicidad a alumnos de diferentes grados educativos; en ocasiones, su aplicación es responsabilidad del Instituto, y otras, de la SEP o de los directivos y docentes de las propias escuelas (INEE, 2015a).

El presente manual técnico documenta los referentes conceptuales y técnicos de las evaluaciones de logro que se realizan mediante instrumentos PLANEA de aplicación externa, a saber, ELSEN y ELCE. La población objetivo la conforman los estudiantes que se encontraban finalizando sus estudios de sexto de primaria y tercero de secundaria en 2015. La aplicación de ambos instrumentos coincidió en ese año y permitió evaluar aprendizajes clave ligados a los campos disciplinares de Lenguaje y Comunicación y de Matemáticas, además de algunas habilidades socioemocionales para la convivencia escolar con los compañeros.

[1] A partir del 2017, la EDC ya no es coordinada por el INEE.

Introducción

Los manuales técnicos que acompañan a los instrumentos de evaluación como PLANEA permiten dar a conocer los procedimientos y los resultados de cada una de sus fases de construcción y aplicación. Lo anterior es de suma importancia para contar con evidencias que sustenten la confiabilidad y la validez de las evaluaciones en tanto son utilizadas como herramientas para la toma de decisiones en materia de mejora educativa (INEE, 2014). Con este manual, el Instituto Nacional para la Evaluación (INEE) cumple además con el mandato de transparencia y rendición de cuentas que le confiere su estatuto, y su responsabilidad con la sociedad civil.[2]

A fin de exponer las actividades y referentes que sirvieron para la construcción, la aplicación y la calificación de PLANEA en 2015, en el manual se retoman los criterios técnicos que el INEE considera necesarios para avalar la calidad de un instrumento de evaluación (DOF, 2017, 28 de abril).

Cada uno de los criterios técnicos retoma estándares característicos de los instrumentos de evaluación de alta calidad, propuestos por parte de los organismos de evaluación educativa internacionales y nacionales, entre ellos, la European Commission (EC); el Grupo de Evaluación de las Naciones Unidas (UNEG, por sus siglas en inglés); el Comité para el Desarrollo de Estándares de Pruebas Educativas y Psicológicas de la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Measurement in Education (NCME); el Joint Committee on Testing Practices (JCTP), el Educational Testing Service (ETS), el Centro Nacional de Evaluación para la Educación Superior (Ceneval) y el propio Instituto (INEE, 2014).

Estos criterios o principios refieren a un proceso de carácter secuencial en el que los productos de una fase se convierten en los insumos de la siguiente, por lo que resulta de suma importancia que todas se implementen de manera cuidadosa y dirigida al cumplimiento de los propósitos de las pruebas.

Las fases críticas del proceso de desarrollo de un instrumento de evaluación, con sus respectivos pasos o etapas, son las siguientes: conceptualización del instrumento de evaluación; elaboración o desarrollo; administración y resguardo del instrumento; procesamiento y análisis de la información; difusión y uso de los resultados, y mantenimiento (DOF, 2017, 28 de abril). En esta última se propone la publicación del informe técnico del instrumento.

Además de estas fases y sus respectivas tareas, en este manual se incluyen otros tópicos de interés con el fin de aprovechar la publicación para que sea un medio que fortalezca la cultura respecto a la evaluación. Valenzuela, Ramírez y Alfaro (2010) conciben esta cultura

[2] Adicionalmente, el Instituto ha publicado otros documentos y fascículos informativos para dar a conocer con oportunidad las características de las pruebas, sus calendarios de aplicación, y sus resultados (INEE, 2015b; 2015c; 2016a; 2017a, entre otros).

como un cúmulo complejo de creencias, valores y prácticas que caracterizan a una comunidad, en este caso, dedicada a la educación y la evaluación como instrumentos de transformación y con vocación formativa. Uno de los resultados de poseer y fomentar dicha cultura es mostrar una disposición positiva hacia las acciones conjuntas de diagnóstico y retroalimentación oportuna que determinan el éxito de PLANEA. La intención es involucrar y compartir con el lector experto, con el no experto y con los diferentes usuarios, los propósitos de las pruebas, sus usos adecuados y no adecuados, los referentes técnicos, otras actividades de mantenimiento y algunos indicadores de validez.

Antecedentes de los instrumentos de evaluación

Alrededor de 2006 comenzaron a cobrar relevancia dos instrumentos que se utilizaron para evaluar el aprendizaje alcanzado por los alumnos de educación básica en México: ENLACE (Evaluación Nacional del Logro Académico en Centros Escolares) y EXCALE (Exámenes de la Calidad y el Logro Educativos), administrados por la Secretaría de Educación Pública (SEP) y el Instituto Nacional para la Evaluación de la Educación (INEE), respectivamente.

ENLACE se aplicó anualmente de 2006 a 2014 a estudiantes de distintos grados de primaria, secundaria y el último ciclo de bachillerato, con la intención de conocer el nivel de dominio que habían alcanzado en el desarrollo de competencias para la lectura, las matemáticas y en ciencias.[3] Por su parte, los EXCALE se aplicaron de 2005 a 2014 a diferentes muestras de alumnos y con distinta periodicidad, para informar respecto al conocimiento de los estudiantes en relación con el currículo nacional de tercero de preescolar, tercero y sexto de primaria, tercero de secundaria y el último grado de educación media superior. Las pruebas permitieron recabar información de los aprendizajes de Español, Matemáticas, Ciencias Sociales, Formación Cívica y Ética, y Ciencias Naturales.[4]

En 2013 el INEE solicitó a un comité de expertos la elaboración de un estudio para analizar la validez y la confiabilidad de dichas pruebas. Los hallazgos se organizaron en cinco ámbitos: alineación con sus referentes, aspectos psicométricos, atención a la diversidad, aplicación, y usos y consecuencias (Martínez, 2015).

Entre los aspectos positivos, el estudio enfatizó la fortaleza conceptual de los exámenes, así como el interés que suscitaron los ejercicios de evaluación en las autoridades educativas y en grupos de la sociedad civil. Algunos problemas tuvieron que ver con la falta de controles durante los procesos de administración, y el uso insuficiente o inadecuado de los resultados. Por ejemplo, en el caso de EXCALE los datos no fueron recuperados para la toma de decisiones en política educativa, mientras que las consecuencias socialmente asociadas a ENLACE, entre ellas el uso de sus resultados para generar *rankings* en escuelas o para otorgar incentivos a docentes, disminuyeron la confianza en las evaluaciones, además de que propiciaron un fenómeno de inflación de resultados (Contreras y Backhoff, 2014) y otro de preparación de los estudiantes para la situación de prueba.

[3] En la página oficial de las pruebas se pueden consultar antecedentes, procedimientos técnicos, resultados y otra información relevante http://www.enlace.sep.gob.mx/

[4] El INEE ofrece un explorador en línea que permite consultar los contenidos y resultados de los EXCALE http://www.inee.edu.mx/explorador/

A partir de esa revisión crítica, el Instituto diseñó en coordinación con la SEP un nuevo plan de evaluación para conocer el nivel de dominio de los aprendizajes clave de los estudiantes, que conserva las principales fortalezas de ENLACE y EXCALE, y supera sus debilidades (INEE, 2016a).

Lo anterior implicó el esfuerzo y la coordinación entre diferentes autoridades a niveles nacional y estatal para mejorar las condiciones de aplicación de las pruebas y generar confianza en los resultados; para integrar información proveniente de distintas fuentes de evidencia y contar con datos contextualizados acerca de los estudiantes, las escuelas y el sistema educativo; para evitar prácticas y efectos no deseados como el otorgamiento de premios a los profesores, y para apoyar una cultura de la evaluación adecuada que invite a la participación y el compromiso de toda la comunidad escolar a fin de incentivar mejores aprendizajes.

Condiciones de factibilidad de las pruebas

En casi todos los frentes se ha logrado desarrollar estrategias alternativas para mejorar la calidad de la evaluación nacional;probarán su eficacia conforme los resultados del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) se vayan convirtiendo en referentes para planificar acciones conjuntas que aseguren el derecho a una educación de calidad.

En este contexto, las condiciones de factibilidad de un proyecto se refieren a la disponibilidad de los recursos necesarios para llevar a cabo los objetivos propuestos. En el caso de PLANEA, y mediante los nuevos esquemas de la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) y la Evaluación del Logro referida a los Centros Escolares (ELCE), se han aprovechado las fortalezas que la Organización para la Cooperación y el Desarrollo Económicos (OCDE) ya anotaba en 2012 respecto a la comunidad escolar mexicana, a saber:

- El compromiso docente con los aprendizajes de los alumnos.
- El desarrollo de un marco integral para la evaluación dentro del aula.
- La alineación de las calificaciones con los resultados de los aprendizajes esperados.
- El fomento de la participación de los padres de familia en el aprendizaje de sus hijos.
- Las iniciativas para el desarrollo de competencias docentes en materia de evaluación del alumnado (Santiago, McGregor, Nusche, Ravela y Toledo, 2012).

Adicionalmente, y para atender el propósito de publicar información contextualizada sobre los resultados del aprendizaje, se contaba ya con cuestionarios de contexto que se aplicaron a los alumnos participantes en las pruebas ENLACE y EXCALE. Estos cuestionarios fueron revisados para fortalecer la recopilación de datos del contexto económico, social y familiar de los estudiantes, y complementar con información de las habilidades socioemocionales relacionadas con la convivencia escolar y los Recursos Familiares Asociados al Bienestar (RFAB).

Algunas de las áreas de oportunidad que subsisten en PLANEA tienen que ver con aspectos de logística de aplicación o con la validez cultural del instrumento. Por ejemplo,

sería deseable contar con instrumentos que, además de reactivos de opción múltiple, incorporaran otro formato de preguntas y estímulos evaluativos. Para elaborar este tipo de exámenes y poder calificarlos en tiempo y forma, se requeriría una mayor infraestructura tecnológica que permitiera realizar las diversas aplicaciones de manera computarizada.

En lo que concierne a la validez cultural y la atención a la diversidad, y sin dejar de señalar los esfuerzos que se han hecho para indagar al respecto (Backhoff, Solano, Contreras, Vázquez y Sánchez, 2015), también será deseable contar con estándares de aprendizaje y versiones de las pruebas adaptadas a los diversos contextos de la república mexicana y las necesidades educativas especiales de los estudiantes.

Marco legal o normativo

El Estado tiene la responsabilidad de informar a la ciudadanía sobre los mecanismos y la calidad de la educación que ofrece a la población. Conocer los resultados de los aprendizajes es útil para que la sociedad pueda exigir a las autoridades el cabal cumplimiento del derecho a recibir educación de calidad (INEE, 2016a).

El marco legal que sustenta estas obligaciones puede encontrarse, por supuesto, en el artículo 3º de la Constitución Política de los Estados Unidos Mexicanos (CPEUM), además del artículo 29 de la Ley General de la Educación (LGE), y los artículos 25 y 27 de la Ley del Instituto Nacional para la Evaluación de la Educación (LINEE).

El artículo 3° constitucional detalla, en su fracción IX, que al INEE le corresponde evaluar la calidad, el desempeño y los resultados del Sistema Educativo Nacional (SEN) en la educación preescolar, primaria, secundaria y media superior. Para ello posee la atribución de diseñar y realizar las mediciones que correspondan a componentes, procesos o resultados del sistema. Por su parte, el artículo 29 de la LGE establece, en su fracción I, que la evaluación deberá llevarse a cabo sin perjuicio de la participación que las autoridades educativas federal y locales tengan, de conformidad con la LINEE y los lineamientos que para ello expida el Instituto.

En dicha ley, el artículo 25 recuerda el mandato constitucional por el que el Instituto debe diseñar y realizar mediciones y evaluaciones que correspondan a componentes, procesos o resultados del Sistema Educativo Nacional (SEN) relacionados con la educación obligatoria. A su vez, la fracción IX del artículo 27 define que esta institución tiene entre sus atribuciones la de "diseñar e implementar evaluaciones que contribuyan a mejorar la calidad de los aprendizajes de los educandos, con especial atención a los diversos grupos regionales, a minorías culturales y lingüísticas y a quienes tienen algún tipo de discapacidad".

Por lo tanto, la nueva generación de pruebas PLANEA tiene entre sus propósitos la atención de las atribuciones legales y constitucionales que se le han conferido al Instituto.

Planeación de los instrumentos de evaluación

De acuerdo con los criterios técnicos del Instituto Nacional para la Evaluación de la Educación (INEE), la primera fase de desarrollo de un examen tiene que ver con la *conceptualización del instrumento de evaluación*, esta fase implica una planeación general y el diseño del instrumento.

En este apartado de planeación se desagregan las actividades y productos relacionados con las decisiones estratégicas que se tomaron para dar origen a la nueva generación de pruebas del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) y para definir la naturaleza y el alcance de los instrumentos. En particular, a fin de comprender los productos de la planeación de PLANEA, es importante recordar que la iniciativa intentó retomar las fortalezas y superar las limitaciones de EXCALE (Exámenes de la Calidad y el Logro Educativos) y ENLACE (Evaluación Nacional del Logro Académico en Centros Escolares), que se aplicaron en años anteriores.

Las decisiones y acciones de planeación están encaminadas a contar con el marco referencial que sustentará la forma y el fondo de las actividades de construcción de las pruebas, de su ensamble, de su calificación y, en particular, de sus usos y consecuencias (DOF, 2017, 28 de abril). Lo anterior se apoya mediante la elaboración de una ficha técnica cuyos rubros se detallan enseguida.

Propósito y modalidades de PLANEA

PLANEA tiene como propósito general conocer la medida en que los estudiantes logran el dominio de un conjunto de aprendizajes clave en diferentes momentos de la educación obligatoria. Se pretende que sus resultados puedan aprovecharse para la mejora educativa a partir de:

- Informar a la sociedad sobre el estado que guarda la educación en términos del logro de aprendizaje de los estudiantes y de la equidad (o inequidad) que existe en los resultados educativos.
- Aportar a las autoridades educativas información relevante para el monitoreo, la planeación, la programación y la operación del sistema educativo y sus centros escolares.
- Ofrecer información pertinente, oportuna y contextualizada a las escuelas y a los docentes, que ayude a mejorar sus prácticas de enseñanza y el aprendizaje de todos sus estudiantes.
- Contribuir al desarrollo de directrices para la mejora educativa con información relevante acerca de los resultados educativos y los contextos en que se dan.

Para lograr estas metas, el Plan integra la información proveniente de tres modalidades de evaluación. La primera, la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN), es responsabilidad del INEE y busca obtener información del Sistema Educativo Nacional (SEN) mediante la evaluación de un gran número de aprendizajes y un diseño de aplicación que permite realizar inferencias a nivel nacionales, estatal y por estrato escolar. Por su parte, la Evaluación del Logro referida a los Centros Escolares (ELCE) ofrece información a las escuelas acerca de un conjunto representativo de los aprendizajes que se consideran clave al término de la educación primaria, secundaria y media superior.

La Secretaría de Educación Pública (SEP) aplica las pruebas ELCE en coordinación con las autoridades educativas estatales, y realiza el análisis de resultados en conjunto con el Instituto. Adicionalmente, las autoridades educativas estatales distribuyen los resultados por escuela y zona escolar, mientras que el INEE reporta los resultados nacionales y a diferentes niveles de desagregación. En todos los casos, la información puede complementarse con los datos de los cuestionarios de contexto, que permiten vincular los resultados de logro con características del entorno personal, familiar, escolar y social de los alumnos.

Se considera que ELSEN y ELCE son pruebas de aplicación externa por oposición a la Evaluación Diagnóstica Censal (EDC), cuya aplicación y calificación son responsabilidad de los docentes frente a grupo y permiten conocer el resultado de los aprendizajes de los estudiantes al iniciar el cuarto grado de primaria.[5]

Objeto de evaluación

Las pruebas están dirigidas a conocer el nivel de logro de los estudiantes en los campos disciplinares de Lenguaje y Comunicación, y Matemáticas, respecto a una serie de aprendizajes clave que cuentan con las siguientes características:

- Son relevantes para el dominio de los conocimientos y habilidades del campo formativo correspondiente.
- Son relativamente estables en el tiempo, con independencia de los cambios curriculares.
- Facilitan la adquisición de nuevos aprendizajes.

Los instrumentos utilizados en PLANEA se diseñan a partir de la identificación de aprendizajes clave establecidos en los planes y programas de estudio y otros referentes curriculares, y de la integración del conocimiento disponible sobre las tendencias en la enseñanza y el aprendizaje de los campos disciplinares evaluados (INEE, 2017b).

En 2015 también se evaluaron habilidades socioemocionales relacionadas con la convivencia escolar en sexto de primaria y tercero de secundaria mediante un cuestionario adicional al de contexto. Estas habilidades están previstas de manera transversal en el currículo y se evalúan porque son necesarias para la interacción social exitosa, dentro y fuera de la

[5] A partir de 2017, la EDC ya no es coordinada por el INEE y dejó de formar parte de las modalidades y el esquema de PLANEA.

escuela, y porque son fundamentales para el desarrollo personal, escolar y social de las personas. Para ello se aplica a los estudiantes un cuestionario que explora aspectos asociados con la solidaridad y el respeto hacia los compañeros, el manejo de conflictos en la escuela y la percepción de bienestar y colaboración en la escuela.

Población objetivo

La población objetivo de las evaluaciones con instrumentos de aplicación externa (ELSEN y ELCE) son los alumnos que terminan sexto de primaria, tercero de secundaria, y el último grado de educación media superior. En este manual se reporta el marco de referencia exclusivo de las pruebas aplicadas a muestras de alumnos representativas de los grados terminales de primaria y secundaria en 2015.

Evaluar a los estudiantes al finalizar un nivel escolar ofrece un buen indicador de la eficacia del proceso educativo en su conjunto, reconociendo los logros de los alumnos a lo largo de varios años de trabajo en los que conforman una red compleja de conocimientos, habilidades y competencias.

Para la aplicación de 2015 no se consideraron adaptaciones de las pruebas para población en lengua indígena o con necesidades educativas especiales. En estos casos los estudiantes recibieron ayuda de sus docentes o instructores particulares durante la aplicación de las pruebas.

Características generales

Para sistematizar otros de los rasgos distintivos que se determinaron desde la conceptualización de PLANEA, se incluye la tabla 1 con algunas características importantes de los instrumentos abocados a la evaluación de aprendizajes clave.

Tabla 1. Características generales de las pruebas PLANEA

Tipo de instrumento	Selección de respuesta
Tipo de reactivos	Opción múltiple con cuatro opciones de respuesta
Modalidad de administración	Lápiz y papel
Longitud del instrumento	150 por campo disciplinar Cada estudiante responde 50 de cada campo
Procedimiento de calificación	Criterial

Estos instrumentos comparten con el cuestionario de contexto la modalidad de aplicación, puesto que son autoadministrables mediante un cuadernillo de preguntas y una hoja de respuestas que los alumnos llenan a lápiz. Por su parte, el cuestionario de alumnos cuenta con 42 reactivos que miden habilidades socioemocionales relacionadas con la convivencia escolar, y 82 reactivos con los que se obtiene información del perfil del alumno, y de su entorno familiar y escolar.

Hay otras características de las pruebas que dependen de la modalidad de evaluación de PLANEA, es decir, ELSEN y ELCE que fueron explicadas en el apartado "Propósito y modalidades de PLANEA". Considerando estas diferencias, en la tabla 2 se pueden contrastar sus rasgos distintivos.

Tabla 2. Características distintivas ELSEN-ELCE

	ELSEN	ELCE
Diseño de aplicación	Muestra de escuelas y alumnos	Muestra de alumnos en todas las escuelas
Ensamble	Matricial	Versión única derivada de ELSEN
Resultados	Agrupados a nivel nacional, estatal y por estrato escolar	Para cada escuela
Responsable	INEE	SEP y escuelas

Como se puede observar en el rubro de resultados, la unidad de análisis para la prueba ELCE es la escuela, mientras que para ELSEN es el SEN desagregado en las entidades federativas, las modalidades, los tipos de servicio u otros estratos.

En las escuelas donde se aplican las pruebas de ELSEN, no se aplican las de ELCE, ya que están alineadas entre sí y comparten contenidos. Esta característica de diseño y la implementación de una estrategia de renovación y mantenimiento de las versiones permite hacer inferencias válidas y confiables a partir de los resultados tanto de ELSEN como de ELCE.

Además de vigilar la aplicación, para asegurar la complementariedad de ambas pruebas, el INEE realiza una verificación estadística a fin de comparar sus resultados y asegurar la congruencia entre ellos.

La aplicación de ELSEN y ELCE se acompaña de materiales complementarios como los cuestionarios de contexto. Además de los que se aplican a los alumnos, se generan también cuestionarios para docentes y directores. La información permite contextualizar los resultados de logro considerando que los estudiantes y sus escuelas trabajan en diferentes entornos socioeconómicos donde prevalecen distintos patrones culturales y conviven diversos actores sociales y educativos. De esta manera, se cuenta con elementos de reflexión sobre los distintos puntos de partida de los estudiantes en sus procesos de aprendizaje, además de que es posible tomar en cuenta los factores internos y externos que dificultan o facilitan la adquisición de los aprendizajes (INEE, 2014).

Usos previstos y no previstos

En congruencia con los propósitos y la ficha técnica de las pruebas PLANEA, es posible determinar los usos que pueden tener sus resultados y que están planteados desde su diseño. Sólo aquellas acciones que se fundamenten en las características y alcances de PLANEA se consideran válidas.

Entre los usos previstos se encuentran el monitoreo del SEN y de los resultados de los centros escolares para tomar decisiones dirigidas a mejorar las condiciones y estrategias relacionadas con los procesos de enseñanza y de aprendizaje.

Darle contexto a los resultados también puede apoyar las determinaciones por parte de las autoridades educativas respecto al equipamiento de las escuelas, la capacitación docente, la actualización de programas en diversos niveles de concreción curricular, etcétera.

En el interior de los planteles, los resultados permiten identificar las fortalezas y debilidades de los alumnos relacionadas con los aprendizajes esperados y los contenidos curriculares. Estos datos pueden ayudar a modificar o complementar las prácticas dentro del aula, acompañadas de otras fuentes de información como la evaluación que realiza de manera cotidiana el propio docente y la autoevaluación de los estudiantes.

Se espera también que, en línea con los modelos más actuales de participación y gestión escolar (SEP, 2016, 21 de julio), los resultados de PLANEA también sean de utilidad para los padres de familia y la comunidad educativa en general. Dado que PLANEA permite comprender los factores que favorecen y obstaculizan los aprendizajes, además de ser consecuentes con las obligaciones de rendición de cuentas, los resultados pueden apoyar acciones de acompañamiento entre padres y maestros, y entre padres e hijos.

Ahora bien, también es necesario considerar que los resultados de los instrumentos de evaluación pueden derivar en usos no previstos por quienes diseñan y construyen PLANEA. Siempre que estos usos sean congruentes con el propósito de las pruebas y sus referentes conceptuales, no se corren riesgos. Sin embargo, es importante establecer cuáles serían los posibles usos inadecuados de la evaluación para prevenir que ocurran (Martínez, 2015), por ejemplo, utilizar los resultados para calificar a los estudiantes o para hacer *rankings* de escuelas. A continuación se señalan algunos de ellos en términos de las limitaciones de PLANEA:

- Las pruebas asociadas a los campos disciplinares básicos están conformadas únicamente por reactivos de opción múltiple que evalúan una muestra representativa de aprendizajes clave, por lo que es responsabilidad del profesor y del propio estudiante explorar otros conocimientos y habilidades mediante herramientas alternativas de evaluación.
- Los resultados no se deben usar para identificar a los mejores alumnos, puesto que las pruebas sólo evalúan aprendizajes clave asociados a dos campos disciplinares: Lenguaje y Comunicación, y Matemáticas. Un estudiante modelo debe mostrar su competencia en todas las asignaturas, además de exhibir habilidades socioemocionales que le permitan aprovechar su conocimiento en escenarios académicos, sociales y personales.
- Tanto ELSEN como ELCE se aplican a muestras de alumnos, por lo que no es posible generar reportes de resultados para cada estudiante matriculado.
- El esquema de aplicación de ELSEN no permite generar comparativos históricos anuales de los resultados nacionales en términos de logro de aprendizajes. Esta decisión se fundamenta en experiencias previas de evaluación que mostraron que no es posible notar cambios significativos a determinados niveles de desagregación cuando

las pruebas se aplican con tanta frecuencia, además de que se puede "desgastar" a los estudiantes.

- Las pruebas no pueden utilizarse para elegir a los mejores docentes, puesto que sólo evalúan aprendizajes clave asociados a dos campos disciplinares y relacionados con lo que solicita el currículo. La responsabilidad principal de un buen docente va mucho más allá del dominio de contenidos o de habilidades escolares; se asocia sobre todo con la capacidad de adaptar su enseñanza a la diversidad de sus alumnos y de acompañarlos en sus procesos de aprendizaje en los ámbitos académicos, personales, sociales, emocionales, entre otros.
- Las pruebas ELSEN y ELCE no fueron diseñadas para evaluar escuelas, por lo que no se deben emitir juicios de valor relacionados con la calidad de los planteles o subsistemas. Una valoración al respecto exige tomar en cuenta muchos otros factores, como la infraestructura de una escuela, los programas de mantenimiento y seguridad, el trabajo colegiado dentro de los cuerpos académicos, las estrategias de seguimiento y apoyo a los estudiantes, etcétera.

Es posible que, en el transcurrir de las diferentes aplicaciones y de acuerdo con las prácticas sociales que se asocien a PLANEA, surjan otros usos para los resultados que deben analizarse y reportarse en documentos de divulgación y en los planes de mejora y actualización que se tienen proyectados para las pruebas.

Perfiles de los cuerpos colegiados

Dado que la construcción de instrumentos de la naturaleza de PLANEA exige de la participación de múltiples grupos de especialistas, la planificación de los instrumentos de evaluación también requirió de la definición de las características que tendrían cada uno de los cuerpos colegiados que participan en el desarrollo de las pruebas (AERA, APA y NCME, 2014).

Para determinar el diseño de la evaluación y construir los exámenes se determinó convocar a especialistas en diseño curricular, autoridades educativas, investigadores de los campos disciplinares que evaluaría la prueba, autores de libros y docentes en ejercicio. En la medida de lo posible se estableció como objetivo contar con especialistas de las diferentes regiones y tipos de escuelas de educación básica del país (general, general multigrado, indígena, privada, comunitaria y telesecundaria).

Las reuniones estuvieron a cargo del personal técnico de la Dirección de Evaluaciones Nacionales de Resultados Educativos del INEE, cuyas actividades principales fueron:

- Elaborar los lineamientos y procedimientos a seguir en cada comité.
- Capacitar a los participantes de los comités.
- Moderar las reuniones de los comités.
- Conformar los productos terminales y recabar evidencia documental.
- Elaborar reportes acerca de los alcances de cada reunión.

Los cuerpos colegiados se dedicaron a la definición de las características generales de la prueba, la selección de contenidos y la elaboración de especificaciones, así como a la elaboración y la validación de reactivos, y al establecimiento de niveles de logro y puntos de corte.

En el *Informe de resultados PLANEA 2015* (INEE, 2017a) se pueden consultar los nombres de los expertos que conformaron los comités de especialistas de las pruebas de Lenguaje y Comunicación, y Matemáticas. Cabe señalar que para los cuestionarios de contexto fue posible retomar el trabajo de los especialistas que diseñaron los cuestionarios de EXCALE y ENLACE, el cual se fortaleció con revisiones internas realizadas por el equipo del INEE.

Diseño de los instrumentos de evaluación

El diseño es el segundo paso de la fase de *conceptualización del instrumento* de evaluación. Implica diversas tareas para operacionalizar el constructo que se pretende medir (objeto de medida), entre ellas, la determinación de los aprendizajes clave, y de la estructura de la prueba, así como la elaboración de las especificaciones de los reactivos o de las tareas evaluativas, que en su conjunto constituyen la tabla de contenidos y especificaciones.

En el diseño de la evaluación se toman como referencia la ficha técnica del examen y un marco conceptual que va a guiar la definición del objeto de medida del instrumento (Downing y Haladyna, 2006). La participación de los cuerpos colegiados es de suma importancia en esta etapa, ya que la claridad, la suficiencia y la pertinencia con que se definan el qué y el cómo se va a medir tienen impacto directo en las futuras acciones de construcción y validación de los reactivos y las pruebas.

En este apartado se describen de manera general el marco de referencia y los procedimientos que sirvieron para el diseño de los instrumentos del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) que se abocan a la medición de aprendizajes clave para los campos disciplinares de Lenguaje y Comunicación, y Matemáticas.

Marco de referencia para el desarrollo de los instrumentos

En línea con los objetivos de PLANEA, el marco de referencia inmediato para los instrumentos fueron los planes y programas de estudio. En México estos planes son de observancia nacional, además de que se cuenta con el Acuerdo 592 por el que se articula la educación básica (SEP, 2011a). En dicho Acuerdo la autoridad educativa indica lo que deben aprender los alumnos de todo el país.

Los instrumentos de PLANEA dirigidos a la evaluación de campos disciplinares se diseñaron a partir de los aprendizajes esperados que aparecen en los planes de estudios vigentes durante el diseño del objeto de medición, a saber, los planes y programas derivados de la Reforma Integral de la Educación Básica de 2011 (SEP, 2011b). Dicha reforma hace énfasis en la promoción de *competencias*, entendidas como la capacidad de responder a diferentes situaciones de la vida cotidiana, académica o profesional, lo que implica un saber hacer, es decir, la aplicación de habilidades, en combinación con el conocimiento que resulte pertinente (saber conocer), y con la posibilidad de valorar adecuadamente las consecuencias de la acción gracias a un saber ser relacionado con valores y actitudes.

Otros elementos que guiaron la implementación de los programas de estudios 2011 fueron los *Estándares Curriculares*, que comprenden el conjunto de aprendizajes esperados que se quiere lograr en todos los alumnos durante los periodos escolares (SEP, 2011a).

Los *aprendizajes esperados* son un indicador de logro y definen lo que cada alumno debe alcanzar en términos de saber, saber hacer y saber ser. Se encuentran graduados de tal manera que se va accediendo progresivamente a *contenidos* cada vez más complejos.

Los *contenidos* son el conjunto de conocimientos, habilidades, destrezas, actitudes y valores que deben aprender los educandos y que los docentes deben estimular para incorporarlos en la estructura cognitiva del alumno. La definición de contenidos concretos en el texto de los programas permite trazar relaciones temporales a lo largo de los planes de estudio, y vincularlas con acciones específicas en el aula que se determinan para cada aprendizaje esperado y el estándar curricular (SEP, 2011a).

Todos estos elementos se utilizan como señales para la elaboración de las pruebas de logro, dado que son referentes acerca de lo que se espera de cada alumno en términos de *aprendizajes clave*.

Como se dijo anteriormente, los aprendizajes clave "son relativamente estables en el tiempo, relevantes para el dominio de los conocimientos y habilidades del campo formativo correspondiente, y facilitadores en la adquisición de nuevos aprendizajes" (INEE, 2015b, p. 15).

Con esta red de elementos como antecedente, y para el diseño de los instrumentos de PLANEA, en cada uno de los planes de estudio se identificaron los aprendizajes clave de los campos de formación de Lenguaje y Comunicación, y de Matemáticas.

En cuanto a referentes técnicos y teóricos, se pueden mencionar los marcos conceptuales y los manuales de los Exámenes de la Calidad y el Logro Educativos (EXCALE), todos disponibles en la sección de proyectos y documentos técnicos de la página electrónica del Instituto Nacional para la Evaluación de la Educación (INEE); los trabajos de Bronzina, Chemello y Agrasar (2009) y su enfoque para la evaluación y la enseñanza de las Matemáticas en los estudios del Segundo Estudio Regional Comparativo y Explicativo (SERCE), y los diversos materiales que publica la Organización para la Cooperación y el Desarrollo Económicos (OCDE) para fundamentar las actividades que realiza a fin de revisar y evaluar los esfuerzos que implementan diferentes países para mejorar los resultados educativos.[6]

Delimitación conceptual del objeto de medida y contenido de las pruebas

El objeto de medida es el conjunto de características o atributos que se miden en el instrumento de evaluación (INEE, 2014, p. 7). Con PLANEA, el propósito era evaluar el logro de aprendizajes clave en dos campos disciplinares, Lenguaje y Comunicación, y Matemáticas. Se tomó en cuenta que ambos son campos relacionados entre sí, que son herramientas esenciales para el desarrollo del aprendizaje de otras áreas del conocimiento, y que resultan buenos indicadores de los resultados educativos en general.

[6] Los documentos del Instituto se encuentran alojados en http://www.inee.edu.mx, y los de la OCDE (OECD, por sus siglas en inglés) en www.oecd.org/edu/evaluationpolicy

Para delimitar el objeto de medida y establecer los aprendizajes clave que se evalúan en PLANEA, se realizaron las siguientes actividades que más adelante se explican detalladamente:

- **Revisión de los materiales curriculares editados por la Secretaría de Educación Pública (SEP).** El personal técnico del INEE compiló los documentos en los que se establecen los planes y programas de las asignaturas elegidas para cada grado escolar al que correspondía la evaluación.
- **Elaboración de tabla de contenidos curriculares.** Un Comité Académico elaboró un documento en el que se presentan los estándares curriculares y aprendizajes clave suceptibles de ser evaluados por PLANEA. Adicionalmente, en la tabla de contenidos se especifica lo que se pretende evaluar con cada contenido curricular y el contexto en el que se debe dar la evaluación. La tarea de correlacionar contenidos con aprendizajes esperados y estándares curriculares se hizo para cada campo disciplinar y por cada grado escolar. Este insumo permitió definir enunciados cortos o especificaciones que se utilizan para la construcción de los reactivos.
- **Definición conceptual y operacional del objeto de medida.** Este conjunto de elementos, de carácter analítico, permite establecer una definición integral de lo que van a medir los instrumentos.

Después del análisis reticular, la responsabilidad de determinar los elementos y contenidos mínimos a considerar para obtener información relevante del objeto de medida recae en un Comité Académico. Los comités de PLANEA estuvieron conformados por autoridades educativas relacionadas con el currículo y con materiales educ ativos, docentes en ejercicio, autores de libros de texto, investigadores y especialistas en didáctica.

Cada uno los cuerpos colegiados operacionalizó los objetos de medida mediante componentes congruentes con los planes de estudio: ámbitos y unidades de evaluación, en el caso de Lenguaje y Comunicación (anexo A), o ejes temáticos, temas y dominios cognitivos, para las pruebas de Matemáticas (anexo B). Del mismo modo, las tablas de contenido de primaria se definieron a partir de los aprendizajes esperados ubicados en los programas, mientras que para secundaria la definición requirió de la recuperación de contenidos.

En el siguiente apartado se describe el modo en que se determinó el número de especificaciones y reactivos en tablas de contenidos, para obtener evidencias claras y suficientes del correspondiente campo disciplinar.

Determinación de la longitud del instrumento y de los pesos específicos de los contenidos

La distribución de las especificaciones y reactivos por cada ámbito, eje temático y nivel educativo es resultado de la carga curricular que tienen los aprendizajes considerados clave para la evaluación, además de una tarea de jerarquización de la importancia de los aprendizajes.

Asimismo, la longitud del instrumento se determina considerando la modalidad de aplicación y que se debe contar con un número mínimo de reactivos para obtener información suficiente acerca de cada campo disciplinar y los ámbitos o temas para los que se quieran generar calificaciones.

El resultado de este ejercicio se plasma en las tablas de contenidos curriculares. La generación de estas tablas fue responsabilidad de un Comité Académico para cada campo disciplinar y grado escolar, que realizó las siguientes tareas:

- Análisis de la retícula.
- Determinación del nivel de relevancia o peso específico.
- Determinación del número de reactivos y longitud del instrumento.
- Conformación final de la tabla de contenidos a evaluar.

Las tablas de contenidos de PLANEA se conformaron progresivamente mediante la selección y la distribución de estándares curriculares, de aprendizajes esperados, de contenidos, y de los ejes temáticos o ámbitos extraídos de los planes y programas. Posteriormente, se incluyeron otros elementos que definen y apoyan la operacionalización de la evaluación, como las unidades de evaluación, los dominios cognitivos, la importancia de cada aprendizaje clave para el objeto de medida, la especificación y la sugerencia de evaluación.

Para definir los aprendizajes clave se consideró el conjunto de contenidos que dan cuenta de un aprendizaje esperado en primaria y en secundaria. Cabe señalar que aunque algunos contenidos temáticos incluidos en las pruebas PLANEA no se reflejan como aprendizajes esperados en los planes y programas, se creyó conveniente integrarlos en la evaluación para atender el marco referencial que establecen la definición global del objeto de medida y los propósitos de cada instrumento.

Con este mapeo inicial y durante la definición de los aprendizajes clave, se hizo una codificación. A cada aprendizaje esperado o contenido del programa se le asignó un nivel de importancia en una escala del 1 al 3:

- Se les asignó un valor de 1 o 2, a los contenidos o aprendizajes esperados que son considerados "esenciales" o "muy importantes" y que cumplen con las características definidas anteriormente como aprendizajes clave. De ellos se derivaron una o más especificaciones que permiten evaluar los aprendizajes de los alumnos.
- Se asignó un valor de 3 a aquellos que no se consideran para ser evaluados, ya sea porque poseen una baja carga curricular o porque requieren de herramientas de evaluación distintas a las que ofrece un examen de opción múltiple. Por ejemplo, en el caso de Matemáticas se dejó fuera la evaluación de las construcciones geométricas con regla y compás.

Posteriormente, se elaboraron especificaciones y reactivos, por lo menos uno por especificación, con la posibilidad de construir más de uno en los casos necesarios para conformar pruebas con una longitud de 150 reactivos. De acuerdo con el diseño matricial de PLANEA, cada sustentante responde 50 reactivos que se distribuyen en seis impresos con dos bloques de 25.

Elaboración y validación de las especificaciones

La definición del objeto de medida requiere, después de la delimitación curricular y conceptual, su concreción en especificaciones, que son enunciados cortos donde se describen, entre otras cosas, las características de los aprendizajes clave y las estrategias para evaluarlos. El propósito principal es brindar los elementos y las instrucciones necesarios para la posterior construcción de los reactivos.

Para elaborar las especificaciones de PLANEA, se reunieron comités académicos por asignatura y grado escolar a fin de precisar y operacionalizar los aprendizajes clave mediante definiciones y enunciados lo suficientemente claros para ser comprendidos por diversos docentes que posteriormente apoyarían en la elaboración de los reactivos.

Los comités académicos se conformaron con expertos en didáctica de los campos disciplinares evaluados, y especialistas en contenidos escolares. En su capacitación se incluyeron elementos para considerar la diversidad cultural del país durante el desarrollo de especificaciones. Además, entre los revisores de las especificaciones hubo especialistas en educación multicultural o en diversidad. En la tabla 3 se presenta una descripción de los participantes de estos comités.

Tabla 3. Participantes en los comités de elaboración de especificaciones, por grado evaluado y campo disciplinar

	6º primaria		3º secundaria	
	LyC	MAT	LyC	MAT
Total	9	14	5	19

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Para cada especificación se identificó el aprendizaje clave con el cual se ancló el reactivo; se detalló la importancia del contenido, y se identificó la función que tiene éste dentro del currículo; se especificaron los conocimientos y habilidades requeridos para responder el reactivo y se sugirieron las habilidades cognitivas involucradas cuando se le contesta correctamente. Los especialistas determinaron diseñar más de una especificación por aprendizaje clave, cuando se requería considerar aspectos diferentes a los aprendizajes esperados o a los contenidos.

A continuación se muestran extractos de las tablas de contenidos de sexto de primaria y tercero de secundaria, con la finalidad de mostrar algunos de los elementos que se incorporaron en el diseño de las especificaciones de PLANEA (figuras 1 y 2).

Figura 1. Extracto de la tabla de contenidos de 6º de primaria

Consecutivo	Nivel	Grado	Asignatura	Clave ID	Eje	Aprendizaje esperado	Tema	Especificación
1	Primaria	6	Matemáticas	M0615AAA001	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.1 Leer y escribir números naturales sin ceros intermedios.
2	Primaria	6	Matemáticas	M0615AAA002	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.2 Leer y escribir números naturales con ceros intermedios.
3	Primaria	6	Matemáticas	M0615AAA003	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.3 Comparar números naturales sin ceros intermedios.
4	Primaria	6	Matemáticas	M0615AAA004	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.4 Comparar números naturales con ceros intermedios.
5	Primaria	6	Matemáticas	M0615AAA005	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.5 Identificar la expresión numérica de una fracción dada una representación gráfica en un modelo continuo.
6	Primaria	6	Matemáticas	M0615AAA006	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.6 Identificar la representación gráfica de una fracción en un modelo continuo dada su expresión numérica
7	Primaria	6	Matemáticas	M0615AAA007	Sentido numérico y pensamiento algebraico	1. Resuelve problemas que impliquen leer, escribir y comparar números naturales, fraccionarios y decimales, explicitando los criterios de comparación.	Números y sistemas de numeración	1.7 Identificar la expresión numérica de una fracción dada una representación gráfica en un modelo discreto.

Figura 2. Extracto de la tabla de contenidos de 3º de secundaria

Consecutivo	Nivel	Grado	Asignatura	Clave ID	Eje	Aprendizaje esperado	Tema	Contenido	Especificación
1	Secundaria	9	Matemáticas	M0915AAA001	Sentido numérico y pensamiento algebraico	Convierte números fraccionarios a decimales y viceversa.	Números y sistemas de numeración	7.1.1 Conversión de fracciones decimales y no decimales a su escritura decimal y viceversa.	1. Identificar la escritura decimal de una fracción decimal o no decimal o viceversa.
2	Secundaria	9	Matemáticas	M0915AAB002	Sentido numérico y pensamiento algebraico	Conoce y utiliza las convenciones para representar números fraccionarios y decimales en la recta numérica.	Números y sistemas de numeración	7.1.2 Representación de números fraccionarios y decimales en la recta numérica a partir de distintas informaciones, analizando las convenciones de esta representación.	2. Ubicar en la recta numérica números fraccionarios dados dos puntos cualesquiera.
3	Secundaria	9	Matemáticas	M0915AAB003	Sentido numérico y pensamiento algebraico	Conoce y utiliza las convenciones para representar números fraccionarios y decimales en la recta numérica.	Números y sistemas de numeración	7.1.2 Representación de números fraccionarios y decimales en la recta numérica a partir de distintas informaciones, analizando las convenciones de esta representación.	3. Ubicar en la recta numérica números decimales dados dos puntos cualesquiera.
4	Secundaria	9	Matemáticas	M0915AAC004	Sentido numérico y pensamiento algebraico	Resuelve problemas utilizando el máximo común divisor y el mínimo común múltiplo.	Números y sistemas de numeración	7.2.2 Resolución de problemas que impliquen el cálculo del máximo común divisor y el mínimo común múltiplo.	4. Resolver problemas que impliquen el cálculo del máximo común divisor.
5	Secundaria	9	Matemáticas	M0915AAC005	Sentido numérico y pensamiento algebraico	Resuelve problemas utilizando el máximo común divisor y el mínimo común múltiplo.	Números y sistemas de numeración	7.2.2 Resolución de problemas que impliquen el cálculo del máximo común divisor y el mínimo común múltiplo.	5. Resolver problemas que impliquen el cálculo del mínimo común múltiplo.
6	Secundaria	9	Matemáticas	M0915ABA006	Sentido numérico y pensamiento algebraico	Resuelve problemas aditivos que implican el uso de números enteros, fraccionarios o decimales positivos y negativos.	Problemas aditivos	7.1.3. Resolución y planteamiento de problemas que impliquen más de una operación de suma y resta de fracciones.	6. Resolver problemas aditivos con números fraccionarios con distinto denominador.
7	Secundaria	9	Matemáticas	M0915ABB007	Sentido numérico y pensamiento algebraico	Resuelve problemas aditivos que implican el uso de números enteros, fraccionarios o decimales positivos y negativos.	Problemas aditivos	7.2.3. Resolución de problemas aditivos en los que se combina números fraccionarios y decimales en distintos contextos, empleando los algoritmos convencionales.	7. Resolver problemas aditivos con números decimales.

Después de completar la tabla de contenidos con todas estas precisiones, se procedió a la elaboración de las *especificaciones largas*,[7] que son documentos técnicos donde se encuentran de manera exhaustiva las características de los reactivos y de los contenidos a evaluar. Se elaboran para describir y delimitar los contenidos seleccionados por los comités académicos, y orientan la elaboración de los reactivos que servirán para la evaluación de los aprendizajes (INEE, 2014).

Una vez concluida la elaboración de especificaciones, los comités académicos revisaron el trabajo hecho por cada uno de sus miembros mediante un protocolo elaborado expresamente para tal fin. La revisión permitió verificar que las especificaciones se apegaran al plan de evaluación asentado en la tabla de contenidos, al enfoque programático y a los procedimientos técnicos marcados por el INEE.

Cada especificación fue revisada por cuatro personas diferentes, dos de ellas integrantes del mismo grupo colegiado, pero esta vez acompañadas por personal técnico del INEE. Se contó también con la participación de dos docentes de educación indígena, quienes, además de revisar el contenido de la especificación, se aseguraron de evitar que se presentaran errores ligados al sesgo cultural, étnico y lingüístico.

En aquellos casos en que alguno de los revisores detectó algún error o imprecisión, la especificación se turnó a la persona que la elaboró para realizar los cambios pertinentes.

En total, se elaboraron y validaron 122 especificaciones para la prueba de Lenguaje y Comunicación de sexto de primaria, y 100 para la de tercero de secundaria. En el caso de Matemáticas, se validaron 93 especificaciones para la prueba de sexto de primaria, y 100 para la de tercero de secundaria. Para algunas de ellas, cada instrumento presenta más de un reactivo, y se alcanzó un total de 150 por prueba.

[7] Una especificación corta desglosa los alcances de la definición del objeto de medida, mientras que la especificación larga profundiza en lo que se requiere evaluar, así como en su delimitación en cuanto a características y ejemplos de reactivos.

Desarrollo de los instrumentos de evaluación

Los criterios técnicos para el desarrollo, el uso y el mantenimiento de instrumentos de evaluación agrupan como segunda fase las actividades de elaboración y piloteo de las tareas evaluativas o los reactivos, así como el ensamble de las pruebas. Por ello y porque estas acciones permiten "materializar" o concretar la evaluación, es que la fase se denomina *desarrollo de los instrumentos* de evaluación.

Lo más importante para asegurar el éxito en esta fase es implementar diversos filtros de calidad técnica y de contenido para asegurar que los reactivos estén alineados con las especificaciones, sean correctos y estén libres de varianza irrelevante para el constructo (DOF, 2017, 28 de abril). Estos elementos impactan en la validez de la prueba, ya que permiten asegurar que los reactivos miden lo que se ha establecido y que los resultados que arrojan varían en función de los conocimientos y habilidades de los participantes, y no por otros factores internos o externos que no se consideran en el objeto de medida.

Elaboración de las tareas evaluativas o reactivos

A partir de la formulación de especificaciones realizada por los comités descritos en el apartado anterior, se reunieron nuevos grupos de especialistas en los programas de estudio para desarrollar reactivos técnicamente correctos. Éstos debían obedecer al formato de opción múltiple con cuatro posibles respuestas (sólo una correcta) y contar con las características técnicas que promueve el Instituto (INEE, 2005). Para ello se les proporcionó la capacitación necesaria, además de apoyo constante por parte de personal especializado.

Los miembros de este grupo colegiado construyeron tres reactivos por cada una de las especificaciones de las pruebas de primaria y secundaria, y para cada campo disciplinar.

El Comité elaborador de reactivos del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) estuvo conformado prioritariamente por investigadores y docentes en servicio, quienes tuvieron el encargo de construir reactivos apegados a cada una de las especificaciones de Lenguaje y Comunicación, y de Matemáticas. En la tabla 4 se muestra el perfil profesional de los participantes de este comité.

Tabla 4. Participantes en los comités de elaboración de reactivos

	6º primaria		3º secundaria	
	LyC	MAT	LyC	MAT
Total	18	16	8	24

Nota. El acrónimo LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Los reactivos contaron con los siguientes componentes básicos:

- **Instrucciones**
  Indicaciones adicionales a la base del reactivo, dirigidas a la acción de lectura de los textos.
- **Texto**
  Elemento que complementa (en el caso particular de Lenguaje y Comunicación) a la base del reactivo y a las alternativas de respuesta.
- **Base del reactivo**
  Estímulo en forma de pregunta o enunciado incompleto al cual debe responder el estudiante.
- **Alternativas**
  Cuatro opciones plausibles, de las cuales se debe seleccionar la respuesta correcta.
- **Distractores**
  Opciones incorrectas pero plausibles. Se llaman así porque distraen a los estudiantes que no conocen la respuesta correcta.

Para la generación de reactivos se establecieron algunas reglas básicas, entre las que destacan las siguientes: la base de la pregunta debe ser autocontenida y debe incluir el máximo de contenido posible del reactivo; se deben evitar los enunciados negativos; cada alternativa debe tener correspondencia gramatical con la base del reactivo; todas las alternativas deben ser plausibles; sólo una alternativa tiene que ser la respuesta correcta; es necesario evitar las alternativas que den pistas inadvertidas o alternativas del modo: "todas las anteriores" o "ninguna de las anteriores", y la respuesta correcta se debe ubicar aleatoriamente entre los distractores.

Asimismo, se solicitó a los elaboradores de reactivos que tuvieran especial cuidado en apegar su trabajo a los lineamientos señalados en los documentos normativos editados por el Instituto Nacional para la Evaluación de la Educación (INEE), en donde se describe una serie de criterios esenciales a considerar para cada reactivo (INEE, 2005). Dichos criterios se pueden clasificar en los siguientes rubros generales:

a) Congruencia con la especificación.
b) Contenido.
c) Uso del lenguaje.
d) Redacción.
e) Materiales de apoyo (ilustraciones, gráficos, tablas, etcétera).
f) Formato.
g) Reactivos de opción múltiple.
h) Normas específicas para los reactivos de Lenguaje y Comunicación.

Una vez elaborados los reactivos, los miembros de este comité también verificaron, mediante un protocolo elaborado para tal efecto, que cada uno de los reactivos hubiera seguido fielmente las indicaciones asentadas en la especificación correspondiente. Además de que no existieran fallas en la redacción y que, de las cuatro opciones de respuesta, tres fueran plausibles y sólo una correcta.

Adicionalmente, cada miembro del comité tuvo que llevar a cabo una prueba empírica a pequeña escala para:

1. Observar las posibles dificultades del alumno al responder los reactivos.
2. Estimar el nivel de dificultad de los reactivos.
3. Conocer el comportamiento de los distractores.
4. Conocer el tiempo que se tardan los estudiantes en contestar el reactivo.

Cada reactivo fue revisado por otros tres expertos, uno de los cuales es especialista en diversidad cultural en educación. Los otros revisores son parte del personal técnico del INEE y se han especializado en didáctica del campo disciplinar a evaluar.

Todos estos filtros son previos a la validación de los reactivos. Para realizar esta tarea se requirió la participación de un nuevo cuerpo colegiado al que se le denominó Comité de validez y sesgo. Esta vez se contó con la participación de docentes en ejercicio de diversos estratos escolares, provenientes de las entidades federativas que se detallan en la tabla 5.

Tabla 5. Entidades de procedencia de los participantes del Comité de validez y sesgo, por grado evaluado

6º primaria	3º secundaria
Aguascalientes	Baja California
Chihuahua	Ciudad de México
Colima	Durango
Ciudad de México	Guadalajara
Guanajuato	Guerrero
Hidalgo	Morelos
México	Oaxaca
Puebla	Puebla
Querétaro	San Luis Potosí
Quintana Roo	Tabasco
Sinaloa	Tamaulipas
Sonora	Veracruz
Tamaulipas	Yucatán
Zacatecas	Zacatecas

La tarea principal de este comité fue la revisión técnica y cultural de ítems, dando especial atención a la identificación de posibles sesgos socio-culturales, regionales o de género. Su trabajo principal fue de carácter técnico pedagógico y se vinculó estrechamente a la realidad educativa de las distintas entidades federativas del país.

La labor de este comité es considerada crucial para la validez técnica y cultural de la prueba, ya que cada reactivo es expuesto a una evaluación sobre los siguientes aspectos:

- Problemas de contenido curricular
  - Falta de alineamiento curricular.
  - Evalúa conocimientos ajenos al currículo.
  - No posee correspondencia con la especificación.
  - Evalúa conocimientos ajenos a los contenidos de la especificación.
  - Reactivo demasiado fácil.
  - Dificultad muy pobre para el grado escolar.
  - Reactivo demasiado difícil.
  - Dificultad excesiva para el grado escolar.
  - Cobertura o sensibilidad a la instrucción.
  - Conocimientos que no se enseñan en el aula.

- Problemas de sesgo
  - Vocabulario o redacción.
  - Uso de palabras cuyo significado tiene variaciones entre los grupos sociales evaluados.
  - Situación.
  - Condición poco cercana a la cotidianeidad de los alumnos de algún grupo social.
  - Estereotipos.
  - Concepciones sociales asociadas a algunos grupos sociales, culturales, lingüísticos, étnicos, de género o socioeconómicos, las cuales son potencialmente ofensivas o denigrantes.

- Problemas técnicos de construcción
  - Errores conceptuales.
  - Problemas o deficiencias respecto a los principios teóricos o conceptuales de la disciplina en cuestión.
  - Contexto inapropiado.
  - No se considera apropiado para los alumnos del grado escolar evaluado.
  - Texto inapropiado.
  - La lectura no se considera apropiada para los alumnos del grado escolar evaluado.
  - Complejidad o difícil comprensión de la base del reactivo.
  - Dificultad innecesaria para entender el problema que plantea el reactivo debido a su redacción y/o sintaxis.
  - Palabras poco comunes y tecnicismos innecesarios.
  - Uso de palabras que no son familiares para los alumnos evaluados o términos técnicos innecesarios que dificultan entender el problema planteado.
  - Información innecesaria
  - Información que, al no ser indispensable, dificulta comprender el problema planteado (se exceptuaron los casos en que la especificación del reactivo requería discriminar información).
  - Redacción, puntuación y ortografía.

- Presenta errores de redacción, puntuación, acentuación, uso de mayúsculas, etcétera.
- Reactivo sin respuesta.
- Ninguna de las opciones del reactivo resuelve completamente, de forma correcta, el problema planteado en el reactivo.
- Más de una opción de respuesta.
- Presenta más de una respuesta correcta o alguno de los distractores es parcialmente correcto.
- Opciones poco plausibles o sin coherencia.
- Las opciones presentadas como distractores son absurdas o fácilmente descartables.
- Pistas de solución.
- La respuesta correcta presenta pistas en su redacción, construcción o estructura; condiciones que la convierten en la única elegible. O bien, una o varias de las opciones presentan pistas que las hacen fácilmente descartables.
- Diseño gráfico.
- Ilustraciones inapropiadas, confusas, mal distribuidas o diseño gráfico poco atractivo para el estudiante.
- Otros problemas.

Cada uno de los reactivos de PLANEA fue revisado por dos docentes mediante una guía elaborada para ese propósito. La asignación de los reactivos que revisarían los docentes se determinó mediante procedimientos aleatorios.

Las observaciones de los profesores determinaron que los reactivos fueran aceptados, descartados o corregidos. A continuación, se explica cada dictamen:

✓ **Aceptar**
- Los reactivos se validan sin sufrir modificaciones.

× **Descartar**
- Los reactivos se dan de baja del banco; suele suceder con los ítems valorados con sesgo o como no apegados al contenido curricular.

– **Corregir**
- Los reactivos se someten a un proceso de revisión y mejora basado en las observaciones de los docentes que pueden señalar, por ejemplo, un uso de palabras o términos no apropiados al grado escolar, uso de situaciones no cotidianas para todos los alumnos, así como fallas en la redacción, en la ortografía o en las opciones de respuesta, entre otras.

En la tabla 6 se contabilizan las acciones que se realizaron como respuesta a las observaciones planteadas por los miembros del Comité de validez y sesgo, para las pruebas de cada grado evaluado.

Tabla 6. Acciones derivadas de las observaciones del Comité de validez y sesgo, por grado evaluado

	6º primaria	3º secundaria
Aceptar	208	119
Descartar	0	0
Corregir	128	30
Total	**336**	**149**

Piloteo de las tareas evaluativas o reactivos

Los reactivos aceptados y los que fueron modificados se probaron empíricamente utilizando una muestra similar a la población objetivo. Esta prueba permitió conocer el comportamiento estadístico de cada uno de los reactivos y seleccionar aquellos que se incluyeron en el examen definitivo. Además, el estudio piloto se diseñó con el objetivo adicional de poner a prueba la logística y los mecanismos de aplicación, con la finalidad de realizar un levantamiento operativo eficaz y efectivo.

El piloteo se llevó a cabo el 25 y el 26 de febrero de 2015 en 233 escuelas ubicadas en seis entidades de la república mexicana: Aguascalientes, Baja California, Chiapas, Querétaro, Tamaulipas y Yucatán. La Dirección de Operación en Campo del INEE estuvo a cargo de la aplicación, para lo cual capacitó a distintas figuras con diversas responsabilidades, entre ellas los responsables operativos estatales y sus respectivos apoyos; los coordinadores de zona, también con figuras de apoyo, además de los aplicadores, los responsables de digitalización, los digitalizadores y los clasificadores necesarios para cada entidad.

Para el levantamiento de datos de la prueba piloto de PLANEA Básica se empleó una muestra de conveniencia, es decir, la muestra se tomó de acuerdo con las necesidades del Instituto, el cual estableció las escuelas participantes, y dentro de ellas se seleccionó a los alumnos de forma aleatoria.

Las escuelas de la muestra de aplicación fueron divididas en cuatro modalidades diferentes que se describen a continuación:

- **Estrato Indígena** (EI). Escuela que atiende a la población indígena.
- **Rural Pública** (RP). Escuela ubicada en una zona rural o de campo, y que cuenta con el apoyo del Estado.
- **Urbana Pública** (UP). Escuela que se establece dentro de la ciudad, y que cuenta con el apoyo del Estado.
- **Urbana Privada** (UPV). Escuela sostenida por particulares, ubicada dentro de la ciudad.

En total, se requirió la participación de 233 escuelas y 7 549 alumnos distribuidos de manera proporcional en cada una de las modalidades.

Los parámetros e indicadores estadísticos de los reactivos que se calcularon en esta fase son:

- Porcentaje de alumnos que respondió correctamente el reactivo.
- Porcentaje de estudiantes que respondió cada una de las opciones de respuesta.
- Comparación de los porcentajes de alumnos que respondieron correctamente el reactivo por estrato escolar (escuelas urbanas públicas, escuelas rurales públicas, cursos comunitarios, escuelas indígenas y escuelas privadas).
- Correlaciones de punto biserial.
- Índice de discriminación de los reactivos.
- Indicadores de ajuste al modelo de Rasch.

Para la prueba de Lenguaje y Comunicación se analizó también el comportamiento de todos los reactivos ligados a un texto "fuente", a fin de seleccionar aquellos que obtuvieran los mejores resultados estadísticos. Los procedimientos de análisis se realizaron mediante el *software R*, en específico el paquete CTT (*Classical Test Theory Functions*) y el *software* JMetrik, además de ConQuest V2.

De manera general, los criterios de selección de los ítems evitan la incorporación a la prueba de reactivos con algunos de los siguientes problemas:

1. **Baja discriminación (B).** La correlación punto biserial corregida es menor a 0.15. En el INEE, los reactivos con una correlación *p bis* menor a 0.15 han sido excluidos de los análisis. Este criterio es extremadamente laxo en comparación con el utilizado para las pruebas del Programa para la Evaluación Internacional de los Estudiantes (PISA, por sus siglas en inglés) 2015 y la Evaluación Nacional del Progreso Educativo (NAEP, por sus siglas en inglés) que son de 0.30 y 0.20, respectivamente.
2. **Distractores con correlación punto biserial muy alta (K).** En los reactivos de opción múltiple calificados de forma dicotómica se considera que un distractor tiene una correlación punto biserial alta si es mayor a 0.05. Esto puede deberse a que la respuesta fue mal especificada, o a que el reactivo tiene más de una respuesta correcta, etcétera.
3. **Respuesta correcta con correlación punto biserial negativa (N).** En los reactivos de opción múltiple calificados de forma dicotómica se reporta si la respuesta correcta tiene una correlación punto biserial menor a -0.05.
4. **Sub-ajuste extremo del modelo de Rasch (X).** Una vez que se ajustó el modelo, si un reactivo presenta una diferencia muy grande entre los valores observados y los esperados, se dice que existe sub-ajuste extremo, y ocurre cuando el *infit* es mayor a 2, por lo que se distorsiona o degrada el modelo de medición (Linacre, 2002).
5. **Dificultad por debajo de lo esperado (E).** Ocurre cuando el reactivo es más fácil de lo esperado y la dificultad promedio del reactivo (p+) es mayor a 90%.
6. **Dificultad por arriba de lo esperado (D).** Ocurre cuando el reactivo es más difícil de lo esperado y la dificultad promedio del reactivo (p+) es menor a 20%.
7. **Sobreajuste del modelo de Rasch (F).** Una vez que se ajustó el modelo, si un reactivo presenta una mínima diferencia entre los valores observados y los esperados, se dice que existe sobre-ajuste, y ocurre cuando el *infit* es menor a 0.8.

8. **Sub ajuste moderado del modelo de Rasch (f).** Una vez que se ajustó el modelo, si un reactivo presenta una ligera diferencia entre los valores observados y los esperados, se dice que existe sub ajuste, y ocurre cuando el *infit* es mayor a 1.2 y menor que 2, lo cual genera reactivos con bajo poder de discriminación.
9. **Alta no respuesta (M).** Ocurre cuando el porcentaje de valores perdidos es superior a 10%.
10. **Tamaño de muestra insuficiente para cada categoría de respuesta (s).** Ocurre cuando el número mínimo de individuos que responden en cada una de las categorías es menor a 15.
11. **Funcionamiento diferencial de reactivos (H/M, B/A).** El funcionamiento diferencial de reactivos ocurre cuando diferentes grupos de estudiantes que responden un reactivo y tienen el mismo nivel de habilidad, no poseen la misma probabilidad de contestar correctamente, es decir, dicho reactivo está funcionando de manera distinta según el contexto y el grupo al que se aplique, después de fijar la habilidad. Es común que estos grupos sean determinados por variables como el género, o las características étnicas, culturales y geográficas, etcétera.

Se construyen dos grupos de comparación, el grupo de referencia, con aquellos individuos para los cuales se espera una ventaja al responder la prueba, y el grupo focal, constituido por los individuos para los cuales se espera una desventaja. El funcionamiento diferencial se calcula con el modelo de Rasch como la interacción que existe entre un reactivo y la subpoblación con la que se quiere hacer la comparación (Paek y Wilson, 2011):

$$\text{logit}(p) = \theta n - \delta i + \tau ig$$

Con el término de *interacción* se identifican los reactivos con funcionamiento diferencial y se procede a clasificarlos en diferentes categorías, de acuerdo con los criterios que se enuncian en la tabla 7 (Wilson, 2005, p. 167).

Tabla 7. Criterios de interacción DIF* (estudio piloto)

Categoría	Criterio	Codificación
Insignificante que favorece al grupo de referencia	$\tau ig < 0 y 2 \lvert \tau ig \rvert < 0.426$	**A-**
Insignificante que favorece al grupo focal	$\tau ig > 0 y 2 \lvert \tau ig \rvert < 0.426$	**A+**
Intermedio que favorece al grupo de referencia	$\tau ig < 0 y 0.426 \leq \lvert \tau ig \rvert < 0.638$	**B-**
Intermedio que favorece al grupo focal	$\tau ig > 0 y 0.426 \leq \lvert \tau ig \rvert < 0.638$	**B+**
Grande que favorece al grupo de referencia	$\tau ig < 0 y 2 \lvert \tau ig \rvert \geq 0.638$	**C-**
Grande que favorece al grupo focal	$\tau ig > 0 y 2 \lvert \tau ig \rvert \geq 0.638$	**C+**

* Differential Item Functioning (DIF) o funcionamiento diferencial del ítem o reactivo.

Procedimiento para el ensamble de los instrumentos

Además de los datos estadísticos, para el ensamble de las pruebas PLANEA se tomaron en cuenta los propósitos y diseños de aplicación que se consideran diferentes para la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) y para la Evaluación del Logro referida a los Centros Escolares (ELCE).

Debido a que PLANEA ELSEN evalúa un conjunto amplio de contenidos del currículo, el personal técnico del INEE implementó un diseño denominado matricial en el que cada alumno evaluado solamente contesta un subconjunto del total de reactivos que integran la evaluación. A partir del universo de reactivos, se construyen formas de examen, relacionadas entre sí, que, al ser aplicadas a los alumnos, permiten conocer lo que éstos saben de ese conjunto amplio de contenidos.

Por su parte, las pruebas ELCE constan de un subconjunto de reactivos que son representativos de los implementados en las ELSEN. Lo anterior permitió utilizar la misma escala de medida para reportar los resultados de ambas evaluaciones, debido a que comparten el mismo constructo. Para que esto pueda ser así, se deben cumplir ciertas restricciones respecto a la manera en que se construyen los subconjuntos de reactivos, mismas que se explican en el subapartado de Ensamble ELCE.

Ensamble ELSEN

Las ELSEN de Lenguaje y Comunicación, y de Matemáticas se ensamblaron mediante un diseño matricial de seis bloques de reactivos por cada uno de los campos disciplinares, los cuales tienen las siguientes características:

- Dificultad promedio similar.
- Distribución de dificultades similar.
- Incluyen reactivos de cada uno de los niveles de desempeño de cada categoría de la evaluación (dimensiones o ejes temáticos).
- Tienen la misma cantidad de reactivos (de 25 a 30).

Para su administración a los alumnos, los seis bloques de reactivos de una evaluación se agruparon en seis formas de examen para cada campo disciplinar. En la tabla 8 se incluye un ejemplo de la conformación de bloques y la dificultad promedio de cada una de las formas de PLANEA 2015, con base en los resultados de la aplicación nacional.

Tabla 8. Bloques de reactivos por forma ELSEN 2015

Campo disciplinar	Forma	Bloques		Dificultad promedio	Total de reactivos
LyC	1	A	B	589.70	50
	2	B	C	597.41	51
	3	C	D	624.89	51
	4	D	E	600.76	50
	5	E	F	607.88	50
	6	F	A	621.13	50
MAT	1	A	B	692.79	50
	2	B	C	690.31	50
	3	C	D	682.23	50
	4	D	E	678.50	50
	5	E	F	670.49	50
	6	F	A	677.14	50

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Cada uno de los alumnos evaluados resuelve una forma de examen. La forma 1 está constituida por los bloques A y B; la forma 2, por los bloques B y C, y así sucesivamente. Cada forma de examen tiene entre 50 y 60 reactivos.

Se dice que el diseño es balanceado porque cada uno de los bloques se administra al inicio y al final de alguna de las formas de examen. Asimismo, se dice que el diseño es incompleto porque cada alumno resuelve solamente una parte de la evaluación.

El diseño matricial permite obtener las puntuaciones de los alumnos en una escala en común debido a la manera en que se eslabonan los bloques de reactivos. Por ejemplo, el bloque A se incluye en la forma 1 y en la forma 6, estableciendo un vínculo con el bloque B y el bloque F.

Por la manera en que se construyen, cada forma de examen es una muestra representativa de la evaluación completa, por lo que es factible hacer inferencias del desempeño de los alumnos respecto a toda la evaluación, es decir, el alumno no necesita responder a todos los reactivos de la evaluación para poder valorar su nivel de desempeño en la escala completa, la cual se construye con base en todos los reactivos que la conforman.

Ensamble ELCE

Para ensamblar las pruebas ELCE se utiliza una forma de las ELSEN. En principio, la forma de examen que se selecciona es indistinta debido a la manera en que fueron construidas, pero después de su primera aplicación en 2015 no puede elegirse la misma forma equivalente. Esto se debe a que, una vez aplicadas, las ELCE son públicas y la sociedad en general pueden consultarlas[8].

[8] En la siguiente liga se puede resolver la prueba ELCE de 2015 y obtener de manera inmediata los resultados http://planea.sep.gob.mx/ba/prueba_en_linea/

La figura 3 ejemplifica la conformación de las ELCE para Lenguaje y Comunicación, y para Matemáticas, donde se han seleccionado la forma 1 y la forma 7 para integrarla. Este esquema se seguirá en las siguientes aplicaciones procurando que las formas sean equivalentes.

Figura 3. Diseño de ensamble de las pruebas ELSEN y ELCE

ELSEN

Asignatura	Forma	Bloques	
Lenguaje y Comunicación	1	A	B
	2	B	C
	3	C	D
	4	D	E
	5	E	F
	6	F	A

Asignatura	Forma	Bloques	
Matemáticas	7	A	B
	8	B	C
	9	C	D
	10	D	E
	11	E	F
	12	F	A

ELCE

	Forma	Bloques	
Lenguaje y Comunicación	1	A	B

	Forma	Bloques	
Matemáticas	7	A	B

Nota: se han utilizado las mismas letras para los bloques de la evaluación de Lenguaje y Comunicación y la de Matemáticas solamente para efectos del ejemplo y no significa que sean iguales para ambas evaluaciones.

Administración o aplicación de los instrumentos de evaluación

El 10 y el 11 de junio de 2015 se llevó a cabo la primera aplicación del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) a alumnos de sexto grado de primaria. En las pruebas de Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) participaron 104 204 estudiantes de una muestra de 3 446 escuelas en 28 entidades federativas.[9] Por su parte, la meta fue aplicar la Evaluación del Logro referida a los Centros Escolares (ELCE) en las 73 306 primarias restantes (INEE, 2015d).

Una semana después, el 17 y el 18 de junio, se aplicaron las pruebas ELSEN a 144 517 estudiantes de tercero de secundaria, inscritos en 3 529 escuelas en 29 entidades federativas.[10] Las ELCE se distribuyeron en las 27 920 secundarias restantes (INEE, 2015d).

La administración o la aplicación de los instrumentos de evaluación constituye una de las fases más importantes para dotar de validez al proceso (AERA, APA y NCME, 2014). En el caso de PLANEA, esto se relaciona directamente con las medidas que se implementan para garantizar la seguridad, la confidencialidad y las condiciones estandarizadas de aplicación de las dos modalidades de la prueba, ELSEN y ELCE.

En el primer caso, es el Instituto el responsable de la implementación del diseño y los protocolos de administración. En el segundo, la Secretaría de Educación Pública (SEP) y las autoridades estatales se encargan de cumplir con las condiciones de aplicación.

El Instituto Nacional para la Evaluación de la Educación (INEE) cuenta con tres mecanismos de control en la aplicación de la prueba ELCE:

1. **Establece** lineamientos generales, entre los cuales está que los aplicadores no deben ser parte de la comunidad escolar.
2. **Supervisa** que el trabajo de campo asegure la confiabilidad de los resultados.
3. **Realiza** una verificación estadística para comparar los resultados de esta prueba y los de ELSEN, a fin de asegurar la congruencia de ambas.

A partir de estas medidas, el INEE puede dictaminar si los resultados de determinada entidad, modalidad educativa o escuela no son confiables y, por lo tanto, recomendar su uso limitado.

[9] No participó Oaxaca, y los datos de Guerrero, Michoacán y Chiapas no se consideran por no atender los criterios de representatividad a nivel estatal.

[10] Al igual que en el caso de primaria, no participó Oaxaca y no se consideran los datos de Michoacán y Chiapas.

En cualquier caso, debe reiterarse que los datos no están diseñados para juzgar el desempeño de los docentes, jerarquizar escuelas, o justificar decisiones de alto impacto para los estudiantes, los docentes o los planteles (INEE, 2017b).

Diseño de aplicación

Las evaluaciones de PLANEA se aplican en dos días, con sesiones de dos horas, y atienden un calendario genérico de aplicación que se ilustra en la tabla 9.

Tabla 9. Calendario de aplicación para PLANEA básica

Primer día	Segundo día	Duración
Organización de la aplicación	Organización de la aplicación	1 hora
Prueba de LyC	Prueba de MAT	2 horas
Receso		20 minutos
Cuestionario para alumnos	Cuestionario para alumnos	50 minutos

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Al finalizar la segunda jornada, se distribuyen los cuestionarios que apoyan la evaluación de habilidades socioemocionales y la obtención de datos respecto al contexto escolar y familiar de los estudiantes. Algunas escuelas y alumnos participaron en un tercer día de aplicación para responder otros cuadernillos de evaluación que sirvieron para la equiparación de puntuaciones (ver apartado de "Modelos para la comparabilidad de resultados en el tiempo").

Para la aplicación de ELSEN, el INEE seleccionó una muestra representativa de alumnos y de escuelas. Por su parte, la SEP administró la prueba ELCE en los demás planteles del país. De este modo, cada alumno seleccionado solamente fue evaluado ya sea por ELSEN o por ELCE, y, una vez que la aplicación se llevó a cabo, la SEP contó con información de todas las escuelas.

Las muestras de cada población objeto de estudio, a saber, los alumnos de sexto de primaria y tercero de secundaria que cursaban el ciclo escolar 2014-2015, se determinaron mediante un diseño muestral que se puede consultar en el anexo C.

Para aplicar los instrumentos, el Instituto seleccionó una muestra de alumnos de determinadas escuelas, y con ellos se formó un máximo de tres grupos de aplicación, de acuerdo con la matrícula total en la escuela. Todos los alumnos seleccionados resolvieron las evaluaciones de Lenguaje y Comunicación, y de Matemáticas.

Se determinó que, en las aulas de aplicación de ELSEN, las personas encargadas de la administración de los instrumentos repartieran las formas de examen en orden espiral y de manera cíclica, lo que permitió asumir que cada forma de examen se aplicó a una población equivalente de alumnos. Por ejemplo, para la evaluación de Matemáticas (formas 7, 8, 9, 10, 11 y 12), un aplicador repartió las formas de examen en el aula de aplicación en la siguiente secuencia: 9, 10, 11, 12, 7, 8, 9, 10, 11, 12, 7, 8, 9, 10, y así sucesivamente.

Por su parte, la SEP y las autoridades educativas administraron una forma única de las pruebas ELCE.

Modalidades y protocolos de administración

Las pruebas ELSEN y ELCE se aplican en papel y lápiz, con un cuadernillo con instrucciones y los respectivos ejercicios, además de una hoja de respuestas por campo disciplinar personalizada con los datos de los sustentantes. No se permite utilizar materiales de consulta o de apoyo como calculadoras, celulares, tabletas o cualquier otro aparato electrónico. Los estudiantes entregan sus materiales al terminar cada uno de los días de aplicación, y sólo conservan un talón desprendible de la hoja de respuestas que tiene impreso el número de folio para la consulta posterior de sus resultados (DGEP, 2015).

El INEE se encarga de la aplicación de ELSEN y la SEP se coordina con las Áreas Estatales de Evaluación (AEE) de cada entidad federativa para las pruebas ELCE. La evaluación busca minimizar cualquier modificación a las actividades normales de la escuela, y dado que sólo se aplica a muestras de alumnos de sexto de primaria o de tercero de secundaria, no se suspenden clases ni las actividades escolares para el resto de los estudiantes.

En el caso de las pruebas ELCE, la aplicación en el plantel es organizada por un coordinador-aplicador, quien recibe apoyo del director de cada escuela. Por medio de un Sistema de Monitoreo Digital, la SEP da seguimiento a las principales actividades programadas para antes, durante y después de la aplicación.

La Secretaría cuenta con documentos normativos que apoyan las actividades de la aplicación, entre ellos, un manual de lineamientos generales, otro para el responsable operativo, y uno para el coordinador-aplicador. Además, en la página electrónica de la SEP se encuentran también las guías para el aplicador, para el director y para el observador externo.[11]

A continuación se detallan algunas condiciones e instrucciones que se atienden antes, durante y después de la aplicación de PLANEA (ELSEN y ELCE), con el propósito de garantizar que se realice en condiciones homogéneas en todo el país, y se genere confianza en los resultados.

Antes…

- Se programa una fase de reclutamiento en la que los aplicadores y otras figuras operativas se registran en el sitio electrónico del INEE, antes de ser seleccionados como candidatos.
- Se capacita a las personas que participan en las diferentes etapas de la aplicación, mediante una presentación con los aspectos normativos y operativos más importantes y los documentos correspondientes, que se ponen a disposición de las entidades federativas en la página electrónica de la Dirección General de Evaluación de Políticas

[11] Material recuperado el 11 de agosto de 2015 de: http://planea.sep.gob.mx/ba/aplicacion/documentos_normativos

(DGEP) de la SEP. En la tabla 10 se puede revisar la cantidad de participantes capacitados y las fechas de 2015 en que se llevaron a cabo las reuniones.

Tabla 10. Programa de capacitación de las figuras de PLANEA 2015

Figura	Capacitados	Fechas
Enlaces administrativos estatales	43	6 y 7 de mayo
Responsable operativo estatal	36	20, 21 y 22 de mayo
Instructor estatal	65	25, 26 y 27 de mayo
Coordinador de zona	756	1 y 2 de junio
Aplicadores en primaria	4 775	9 de junio
Aplicadores en secundaria	5 776	16 de junio

Se integran expedientes de las personas que participan como coordinadores-aplicadores y aplicadores, con el propósito de verificar que cumplan con el perfil requerido y dar seguimiento al desempeño mostrado en su participación, para su consideración en futuras aplicaciones.

- Se notifica a los directivos de los planteles con aproximadamente ocho días de anticipación para que se implementen estrategias que aseguren la participación de los alumnos.
- Se empaquetan los materiales en las oficinas estatales, escuelas, oficinas o bodegas que se equipan para coordinar la distribución y la recuperación de materiales. En 2015 PLANEA empleó diversos instrumentos: cuestionarios dirigidos a 6 937 directores y 16 004 docentes, así como cuadernillos y cuestionarios para 284 515 alumnos. Estos materiales se organizaron en 21 617 cajas que se distribuyeron entre los 10 551 grupos que conformaron la muestra. Adicionalmente, se dispuso de diversos materiales informativos, para la capacitación y el control de la operación (INEE, 2015d).
- Participa un coordinador-aplicador por escuela, quien, junto con el aplicador externo (si es el caso), se presenta con el director del plantel para explicar detalladamente la logística de aplicación.
- El coordinador-aplicador traslada los materiales de evaluación a la escuela en cajas selladas, para garantizar la confidencialidad de la prueba, y sólo son abiertas en presencia del director y observadores externos (padres de familia o líderes de la comunidad, empresarios, entre otros).

Durante…

- Se elabora un reporte de arranque nacional que permite informar, de manera inmediata, las incidencias para iniciar la aplicación. Algunas causas para postergar alguna jornada de aplicación, o suspender ambas, fueron el paro de labores, condiciones climatológicas adversas, la negativa para admitir la prueba por parte de la comunidad escolar, condiciones de inseguridad y la ausencia del personal operativo.

Aun así se alcanzaron tasas muy altas de aplicación que se muestran en la tabla 11.

Tabla 11. Porcentaje de aplicación nacional con respecto a las escuelas programadas, por campo disciplinar y grado evaluado

6º primaria		3º secundaria	
LyC	MAT	LyC	MAT
91.9	90.7	92.6	92.7

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

- El día de la aplicación asisten los observadores externos para supervisar que la aplicación se lleve a cabo conforme a la normatividad establecida. Dichas personas no intervienen en el proceso de aplicación.
- Se establece un control para la identificación de los grupos de aplicación y los alumnos que integran cada uno de dichos grupos, con el propósito de realizar estudios para detectar la copia durante la aplicación.
- Los directores de todas las escuelas contestan un cuestionario de contexto que tiene la finalidad de obtener información sobre las características del plantel.
- El docente asesor o tutor de cada grupo contesta un cuestionario de contexto para recabar información referente al o los grupos que atiende.

Después…

- Los paquetes de hojas de respuestas y los materiales correspondientes a cada día de aplicación se colocan dentro de una caja que debe sellarse con la etiqueta firmada por el aplicador, el director y, en su caso, un observador externo. Esta acción se realiza dentro de la escuela.
- El coordinador-aplicador es el encargado de trasladar y entregar las cajas al responsable operativo o coordinador regional, según sea el caso.
- Por medio del Informe de aplicación, cuyo diseño permite su lectura óptica, se obtiene información rápida acerca de las condiciones en que se realizó la aplicación en cada uno de los planteles participantes y de las incidencias más frecuentes, lo cual sirve como retroalimentación para las aplicaciones futuras.

Protocolos de resguardo de la información

Para garantizar la confidencialidad de los instrumentos y de los resultados, se cuenta con normas a seguir durante el traslado, el resguardo y el regreso de los materiales de aplicación a las instalaciones dispuestas para tal fin.

La operación de PLANEA-ELSEN requirió un total de 36 oficinas estatales, entre ellas, dos oficinas para el Distrito Federal[12] (oriente y poniente), dos para México (área metropolitana y Toluca), tres para Veracruz (Minatitlán, Poza Rica y Xalapa) y una en Oaxaca para el levantamiento de escuelas administradas por el Consejo Nacional de Fomento Educativo

[12] Fue hasta 2016 cuando el Distrito Federal adquirió el nombre de Ciudad de México.

(CONAFE). Además, para coordinar la distribución y la recuperación de materiales y realizar el resto de las tareas operativas, se instalaron 103 oficinas regionales.

Para la operación de las coordinaciones de zona se utilizaron 653 espacios, de los cuales 50% se concertaron con instancias de los tres niveles de gobierno (escuelas, oficinas, bodegas, salas de juntas); 31.5%, con particulares (locales e incluso viviendas), y 18.5% se arrendaron (en hoteles). Además, se dispuso de 78 vehículos rentados ex profeso para este operativo (INEE, 2015d).

En el caso de las pruebas ELCE, todos los procesos son responsabilidad de un responsable operativo o coordinador regional, además del coordinador-aplicador, quien cuenta con el apoyo del director de plantel y los observadores externos para asegurar que los materiales se reciban y distribuyan después de abrir las etiquetas de seguridad de las cajas. De la misma manera, estas figuras controlan los espacios de distribución y resguardo de materiales y, después de la aplicación, sellan nuevamente los paquetes en presencia de observadores externos.

Procedimientos para el análisis de resultados de los instrumentos de evaluación

La cuarta fase para el desarrollo de un instrumento de evaluación engloba los análisis que se llevan a cabo para conocer las propiedades métricas de los reactivos y establecer el modelo de puntuación que se va a utilizar para expresar los resultados. En una iniciativa de evaluación como el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) es necesario tomar en cuenta el error de muestreo, los valores de confiabilidad y el error de medida para dictaminar si los resultados cumplen con criterios estadísticos de calidad.

Los análisis de PLANEA-ELSEN (Evaluación del Logro referida al Sistema Educativo Nacional) son responsabilidad del personal técnico del Instituto Nacional para la Evaluación de la Educación (INEE), con la participación de la Dirección General de Medición y Tratamiento de Datos y las direcciones académicas de éste. El análisis de las pruebas de Evaluación del Logro referida a los Centros Escolares (ELCE) lo realiza la Secretaría de Educación Pública (SEP) con la colaboración del Instituto.

Evaluación de la métrica de los reactivos y de los instrumentos

Una vez que se recibieron los productos de la lectura de hojas de respuesta de PLANEA, se realizaron diversos procedimientos de depuración de bases de datos para detectar y corregir anomalías como datos incompletos, repeticiones y codificaciones dobles o no permitidas. Este tratamiento previo al análisis de los resultados permite evitar inconsistencias en fases posteriores.

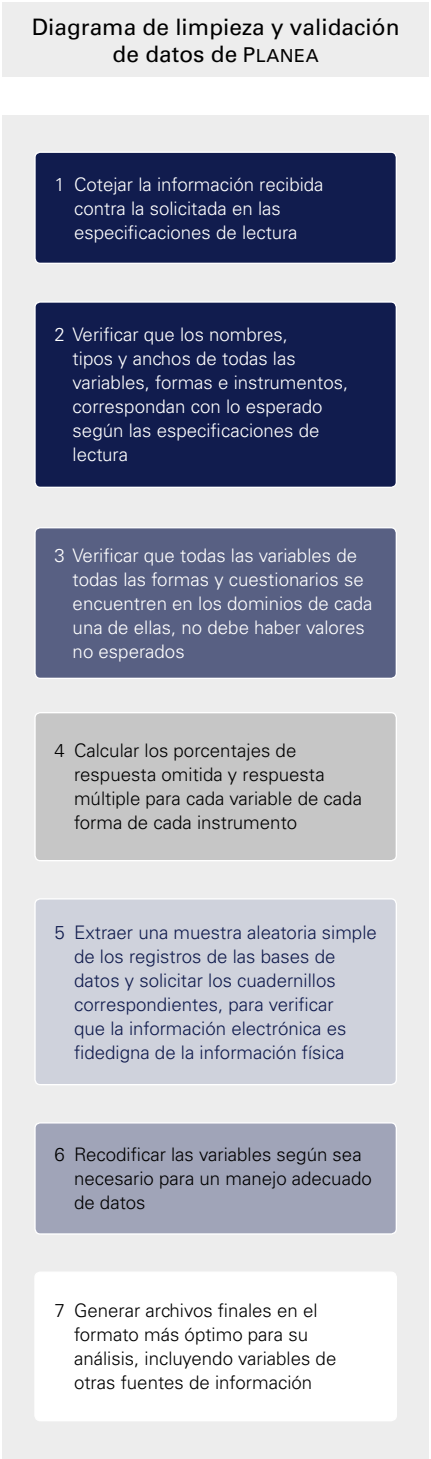
Cuando no fue posible corregir las anomalías se siguió un sistema de códigos que permiten su identificación y que se presenta en la tabla 12.

Tabla 12. Sistema de códigos para limpieza de bases de datos

Código	Descripción
0	Fallo
1	Acierto
5	Reactivo no alcanzado (se considera como no administrado)
6	Reactivo no administrado por error de impresión
7	Reactivo no administrado (no se incluyó en su cuadernillo de examen)

El tratamiento de limpieza y validación de las bases de datos de PLANEA puede sistematizarse mediante los pasos que ilustra la figura 4. El producto fue una base de datos completa por nivel escolar y por campo disciplinar.

Figura 4. Proceso de limpieza y validación de datos



Posteriormente y en congruencia con el modelo de medición establecido durante la fase de planeación del instrumento, se ejecutaron análisis estadísticos con base en el modelo logístico simple de Rasch, el cual permite conocer los valores de los parámetros e indicadores de cada reactivo de PLANEA, el valor de habilidad de cada sustentante, y otros índices de los instrumentos en su totalidad.

En primera instancia, estos datos sirven para clasificar los reactivos de acuerdo con su comportamiento estadístico que, si bien ya fue probado en el piloteo, vuelve a considerarse como indicador de calidad en esta fase de calibración. Un segundo filtro lo constituye el análisis de funcionamiento diferencial que se reporta en el apartado de *Indicadores de validez*.

El primer análisis de calidad implica revisar tres indicadores:

- Correlación punto biserial (r_{pb}).
- Ajuste al modelo de Rasch ($infit$).
- Curva característica del reactivo.

Para el caso de la correlación punto biserial, se considera que sólo los reactivos con $r_{pb} \geq 0.15$ discriminan de forma aceptable, por lo que pueden tomarse en cuenta para la calificación. Como ejemplo de las tareas que realiza la Dirección de Tratamiento de Datos del INEE, enseguida se presentan figuras con las gráficas de la distribución de la correlación biserial para las pruebas de Lenguaje y Comunicación, y de Matemáticas que se aplicaron en sexto de primaria.

Figura 5. Distribución de la correlación r_{pb} de los reactivos de Lenguaje y Comunicación, 6º de primaria

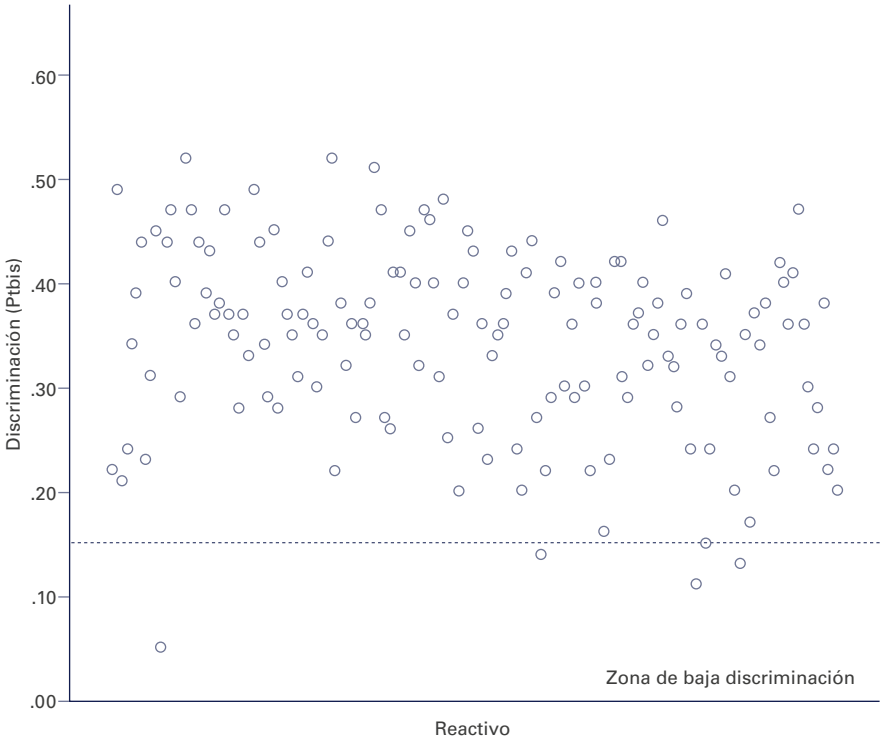
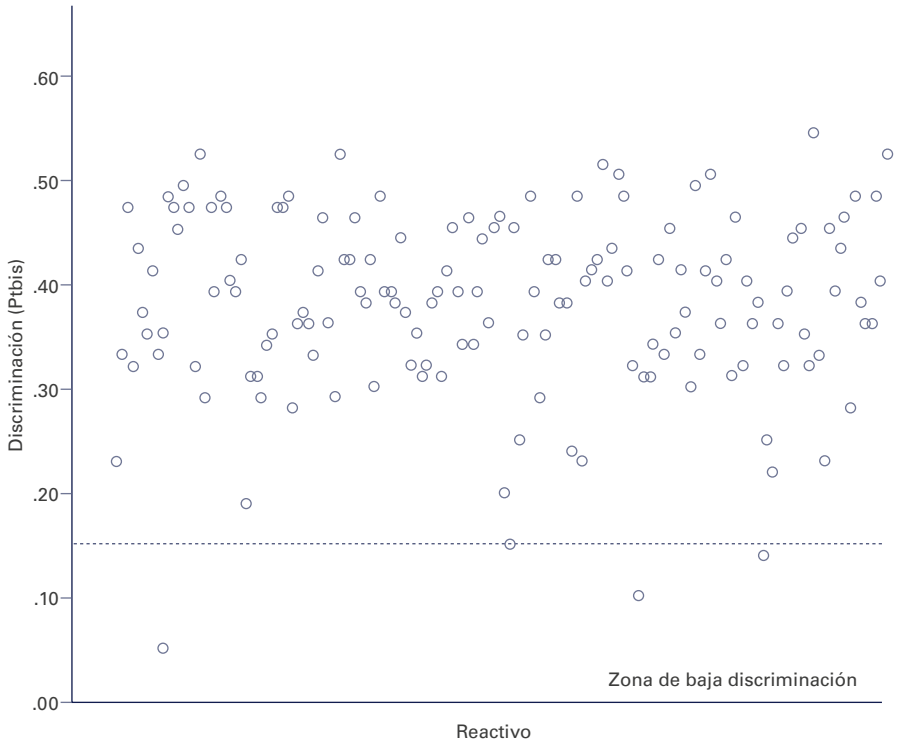


Figura 6. Distribución de la correlación r_{pb} de los reactivos de Matemáticas, 6º de primaria



Con respecto al ajuste de los reactivos al modelo de Rasch, en la tabla 13 se incluyen los valores y la curva característica que se considera aceptable para cada uno.

Tabla 13. Clasificación de reactivos para calificación

Ajuste modelo Rasch	Clasificación	Curva característica
$0.8 \leq infit \leq 1.2$	El reactivo ajusta de forma aceptable	La curva característica formada con los datos observados está cerca de la curva característica esperada
$infit < 0.8$ o $1.2 < infit < 2$	El reactivo desajusta moderadamente	-
$infit \geq 2$	El reactivo debe ser eliminado de la prueba	La curva característica formada con los datos observados no se parece a la curva característica esperada

Las siguientes son las gráficas de distribución del ajuste del modelo para las pruebas de Lenguaje y Comunicación, y Matemáticas que se aplicaron en tercero de secundaria.

Figura 7. Distribución del ajuste al modelo de Rasch (infit) de los reactivos de Lenguaje y Comunicación, 3º de secundaria

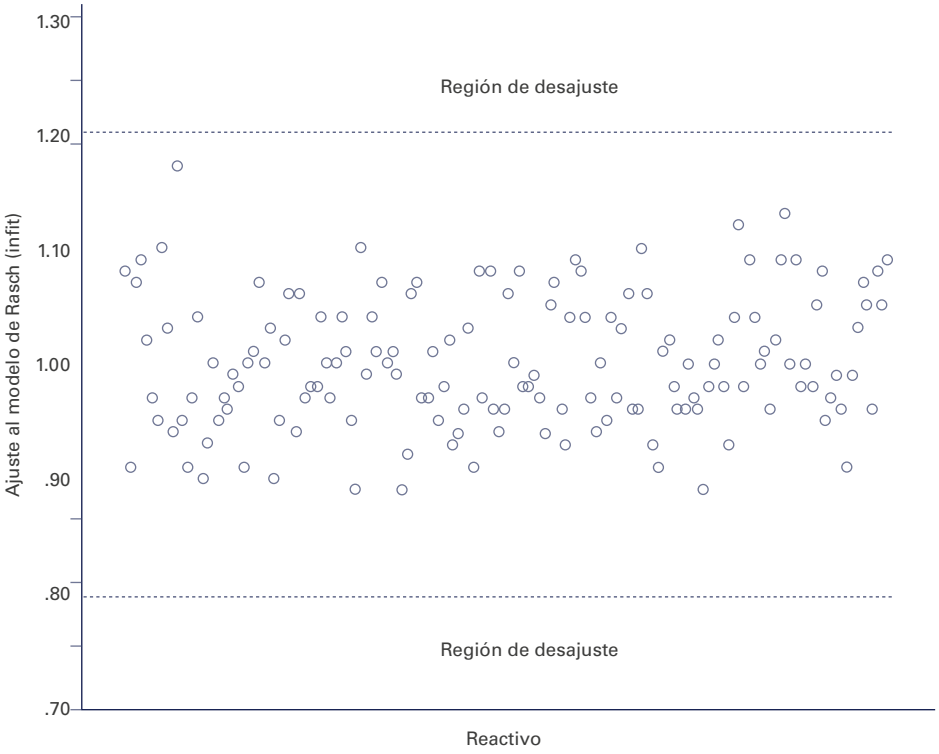
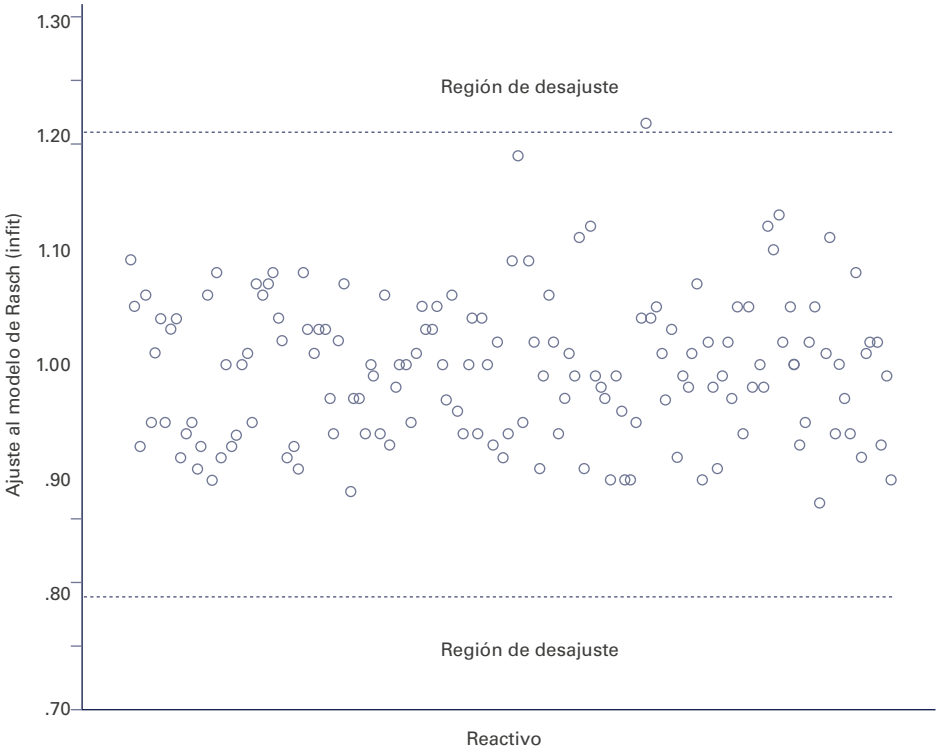


Figura 8. Distribución del ajuste al modelo de Rasch (infit) de los reactivos de Matemáticas, 3º de secundaria



Después de conjuntar los datos de la calibración y del análisis de funcionamiento diferencial, se eliminaron siete reactivos de las pruebas de Lenguaje y Comunicación de ambos niveles educativos, tres para la prueba de Matemáticas de sexto de primaria y seis para la de tercero de secundaria (INEE, 2017b).

En el anexo D se incluyen los resultados obtenidos de la calibración de los reactivos que conformaron PLANEA 2015, misma que se realizó mediante el *software* Conquest V2.

Para la calibración se utilizó el modelo condicional de respuesta al ítem en conjunción con el modelo de la población, pero sin incluir variables de condicionamiento. El método utilizado para hacer la estimación de los parámetros de los reactivos fue el de máxima verosimilitud marginal, que asume independencia condicional entre las respuestas a los reactivos y que la habilidad es un efecto aleatorio (INEE, 2017b).

Estos datos y los obtenidos en el proceso de calificación también permitieron calcular los valores de confiabilidad para los instrumentos. Tal medida resulta relevante para establecer la calidad métrica de las pruebas, en virtud de que alude al grado en que se generan puntuaciones consistentes con poco error de medición (Manzi, García y Godoy, 2017).

Se aprovechó que el modelo de la población hace posible ajustar un modelo bidimensional para las habilidades, en el que se consideran relacionadas las habilidades de un estudiante en Lenguaje y Comunicación con las de Matemáticas (INEE, 2017b).

Del ajuste del modelo bidimensional, se derivó que la correlación de los puntajes entre las dimensiones es de 0.837 para las pruebas de sexto de primaria y 0.771 para las de tercero de secundaria. Introducir este indicador en el cálculo de la confiabilidad permite obtener valores aún más altos, por lo que se tomó la decisión de utilizar un modelo bidimensional para generar los puntajes (tabla 14).

Tabla 14. Confiabilidad modelo unidimensional y bidimensional, por grado evaluado y campo disciplinar

Grado	Campo disciplinar	Confiabilidad modelo unidimensional	Confiabilidad modelo bidimensional
6º primaria	LyC	0.833	0.899
	MAT	0.858	0.911
3º secundaria	LyC	0.798	0.833
	MAT	0.790	0.825

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Modelo de puntuación de las respuestas

Para continuar con el proceso de escalamiento, es necesario suponer que existe una distribución de probabilidad asociada a la variable latente y que tiene una forma funcional conocida.

La práctica más común es asumir que las habilidades de los estudiantes provienen de una distribución normal multivariada. Entonces, el modelo poblacional del estudiante i es:

$$\theta_i = \Gamma' y_i + \in$$

Donde $\in$ se distribuye normal multivariada con media 0 y matriz de varianza y covarianza Σ entre las dimensiones. De esta forma, los parámetros a estimar están en $\alpha = (\Gamma, \Sigma)$. En este contexto las y_i son conocidas como variables de condicionamiento, que resultan indispensables para la estimación final de las habilidades. El modelo de la población se define asumiendo que existe independencia entre los sustentantes:

$$p(\theta \mid Y, \alpha, \delta) = \prod_{i=1}^{N} p(\theta_i \mid Y_i, \alpha, \delta) \quad \ldots(5)$$

En el modelo de la población se estimaron los coeficientes de regresión y la matriz de varianza y covarianza, tanto en la fase del condicionamiento, como en la estimación de parámetros.

Las variables de condicionamiento se construyeron considerando dos grupos; el primero lo constituyen las variables que están directamente en el modelo de regresión, entre ellas, las que definen subpoblaciones de interés:

- Dominios primarios de PLANEA-ELSEN. Las entidades federativas.
- Dominios secundarios de PLANEA-ELSEN. Tipos de escuelas que consideran marginación y sostenimiento.
- Sexo.
- Edad.
- Grado de multigrado (sólo en primaria).

El otro grupo de variables de condicionamiento lo constituyen las que indirectamente influyen sobre la habilidad de los estudiantes, por ejemplo:

- La forma del examen que se le asignó al sustentante.
- Estimación preliminar del promedio del rendimiento de cada estudiante en ambas asignaturas dentro de cada escuela. Es el promedio de los estimadores de máxima verosimilitud de ambas asignaturas para todos los demás estudiantes. Incluir las medias de rendimiento por escuela sirve para tomar en cuenta gran parte de la variación existente entre escuelas.

Dichas variables se pueden construir a partir de los cuestionarios de contexto de la siguiente forma:

1. Cada una de las variables de los cuestionarios de contexto A y B fue recodificada mediante el método *deviance coding*. Es decir, se construyen variables indicadoras para cada una de las categorías de respuesta de las variables de los cuestionarios de contexto; a continuación, se elige la categoría de referencia de mayor frecuencia estimada en la población. Los valores de esa categoría son recodificados a -1.
2. Después se elaboró un análisis de componentes principales para reducir la dimensionalidad. Se retuvieron las cargas factoriales que explicaran la varianza total de 95% de ambos cuestionarios de contexto. En general, la cantidad de factores que se construyen tiene que ser de al menos 200. Para sexto de primaria se construyeron 240 y para tercero de secundaria, 209.

Una vez construidos ambos grupos de variables de condicionamiento, se estimaron los coeficientes de regresión, en donde se obtuvieron dos grupos de coeficientes, uno para Lenguaje y Comunicación y otro para Matemáticas. Los coeficientes de regresión sirven para definir las medias de cada una de las subpoblaciones de interés.

Para terminar la estimación de los parámetros poblacionales, a la par de la de los coeficientes de regresión, se estimó la matriz de varianza y covarianza fijando los parámetros de dificultad de los reactivos (INEE, 2017b). Además, se realizó una corrección para las estimaciones considerando el diseño matricial de las pruebas PLANEA.

Aunque en teoría no tendría que esperarse un "efecto del cuadernillo", porque el ensamble proyectó bloques incompletos balanceados para construir cada forma, en la práctica es común encontrar algún impacto debido a que los reactivos aparecen en diferentes posiciones dentro de los cuadernillos. Así, en la estimación de los parámetros de los reactivos, los efectos de los cuadernillos fueron incluidos en el modelo de medición para prevenir que se confundieran las dificultades de los reactivos y sus efectos. El parámetro del cuadernillo se definió formalmente en el mismo sentido de los parámetros de los reactivos, reflejando la dificultad de cada forma. El modelo empleado es uno de los llamados de facetas y es el siguiente:

$$logit(p) = \theta - \delta_j - \tau_k$$

Después de la estimación de las habilidades de los estudiantes con valores plausibles, los parámetros asociados a los cuadernillos fueron sumados al valor de habilidad de los estudiantes con la finalidad de cancelar el efecto de cada uno de los cuadernillos (INEE, 2017b).

Para convertir la habilidad de los sustentantes en una puntuación fácilmente manejable y generar inferencias acerca de los resultados se tomó como base la metodología propuesta por la Organización para la Cooperación y el Desarrollo Económicos (OCDE) a fin de asignar la calificación de los participantes de la prueba PISA (Programa para la Evaluación Internacional de los Estudiantes). Los resultados de PISA se presentan por medio de escalas con una puntuación media de 500 y una desviación típica de 100. Esta prueba comparte algunas características con PLANEA, entre ellas, su diseño matricial y un objeto de medición centrado en aprendizajes clave.

Asimismo, el cálculo de las puntuaciones de PLANEA se llevó a cabo mediante las macros publicadas por la OCDE en 2009, dispuestas para las estimaciones de estadísticos univariados, frecuencias y diferencias entre subpoblaciones.

Para la estimación de estadísticos univariados y de frecuencias, la OCDE describe dos tipos de macros, uno utiliza la metodología de valores plausibles y otro es para variables no imputadas, como las respuestas a los cuestionarios de contexto. El logro de los estudiantes es una variable que no se puede observar directamente, por lo que se mide mediante las respuestas a los reactivos que cada estudiante resuelve. Mislevy (1991) advierte que los procedimientos estándar para hacer inferencias provenientes de muestreos complejos no son aplicables cuando la variable de interés no puede ser observada directamente. La anterior es otra razón por la que se adoptó la metodología de PISA para la construcción de las escalas de calificación de PLANEA, estimando las desviaciones estándar y media en escalas *logit* con cada vector de valores plausibles y con los factores de expansión de la muestra (anexo E).

Una vez cancelado el "efecto del cuadernillo", se aplicó una transformación lineal para que la media y la varianza de los datos muestrales en la escala *logit* se trasladaran a la escala de PLANEA (anexo F). Esta escala se definió con una media de 500 unidades y una desviación de 100 para cada una de las asignaturas. La transformación fue así:

$$T(\theta) = \frac{\sigma_1}{\sigma_0}\theta + \left[\mu_1 - \frac{\sigma_1}{\sigma_0}\mu_0\right]$$

En donde:

θ Valor plausible en la escala *logit*.
σ_1 Es la desviación estándar de la escala PLANEA y su valor se fija en 100.
μ_1 Es la media de la escala PLANEA y su valor se fija en 100.
σ_0 Es la desviación estándar en la escala *logit*.
μ_0 Es la media en la escala *logit*.

En la página electrónica de PLANEA se pueden consultar los resultados de convertir las puntuaciones de todos los estudiantes.[13] Las bases de datos están disponibles en formato de SPSS (Statistical Package for the Social Sciences) y se anexan las sintaxis correspondientes para etiquetar las variables y sus respectivos valores, así como para identificar los casos perdidos y dar formato a los datos (INEE, 2016b).

Además de las bases de datos, se incorporan los principales resultados de las evaluaciones con sus correspondientes errores estándar, así como la cantidad de información con la que se realizó la estimación, a saber, la Unidad Primaria y la Unidad Secundaria de Muestreo (número de escuelas y número de alumnos en la muestra, respectivamente). Las tablas de resultados resaltan en negritas las diferencias significativas por entidad federativa y para cada subpoblación. En el anexo G se profundiza en los procedimientos utilizados para la estimación de logro y se dan orientaciones para comprender las tablas correspondientes.

Para dar lectura a estos datos se debe recordar que la media de la población de estudio se fijó en 500 unidades con una desviación estándar de 100 unidades. Lo anterior permite cuantificar cualquier diferencia en términos de la desviación estándar (tamaño del efecto).

Tabla 15. Puntajes promedio por tipo de escuela, 6º de primaria, PLANEA 2015

	General pública		Indígena		Comunitaria		Privada	
	PP	ee	PP	ee	PP	ee	PP	ee
LyC	494	1.5	424	3.5	459	4.4	603	3.0
MAT	494	1.3	438	4.4	478	5.5	588	3.7

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas; PP es el puntaje promedio y ee indica el error estándar.

En la tabla 15 se incluyen las puntuaciones promedio por tipo de escuela para los dos campos disciplinares evaluados en sexto de primaria.

Éste es un ejemplo de los resultados y comparativos que pueden realizarse, tomando siempre como referencia la puntuación promedio a nivel nacional (500) y los niveles de desagregación del diseño muestral. En la tabla 16 se incluyen los resultados, ahora para tercero de secundaria, considerando las puntuaciones promedio por nivel de marginación.

[13] Ver http://www.inee.edu.mx/index.php/planea/bases-de-datos-planea

Tabla 16. Puntajes promedio por nivel de marginación, 3º de secundaria, PLANEA 2015

	Alto		Medio		Bajo	
	PP	ee	PP	ee	PP	ee
LyC	477	1.5	500	2.0	520	1.8
MAT	470	2.0	496	1.7	529	1.7

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas; PP es el puntaje promedio y ee indica el error estándar.

Es importante considerar que la muestra de estudiantes se diseñó para obtener resultados a nivel nacional, por tipo de escuela y por entidad federativa. La cantidad de alumnos seleccionados en la muestra se calculó para tener un margen de error máximo del 10% de la desviación estándar de la variable de logro de interés (medias, porcentajes, correlaciones) al nivel de entidad federativa. En los resultados con una de desagregación menor a la de entidad federativa, como los tipos de escuelas en la entidad u otras subpoblaciones, el margen de error se incrementa debido a que se dispone de menor cantidad de datos en la estimación de los parámetros poblacionales. Por lo anterior, es necesario revisar los errores estándar asociados antes de emitir juicios sobre los resultados.

Otros aspectos que deben tomarse en cuenta durante la revisión de resultados son las condiciones del contexto escolar, social y familiar que fueron recopiladas mediante los cuestionarios a alumnos, docentes y directores. El *Informe de resultados PLANEA 2015* (INEE, 2017a) se organiza de tal modo que, sólo después de haber señalado detalladamente los datos de contexto de los alumnos y las escuelas primarias y secundarias de México, da pie a la revisión de resultados generales, por tipo de escuela, sexo, edad de los alumnos, nivel de marginación, tamaño de la localidad, y entidad federativa.

Para la comprensión y la interpretación de los datos, y como parte de las estrategias de uso y difusión de los resultados de PLANEA, el Instituto y la SEP determinaron clasificar las puntuaciones en niveles de logro con una connotación cualitativa asociada a las fortalezas de los estudiantes, misma que se explica a detalladamente en la fase de *difusión*.

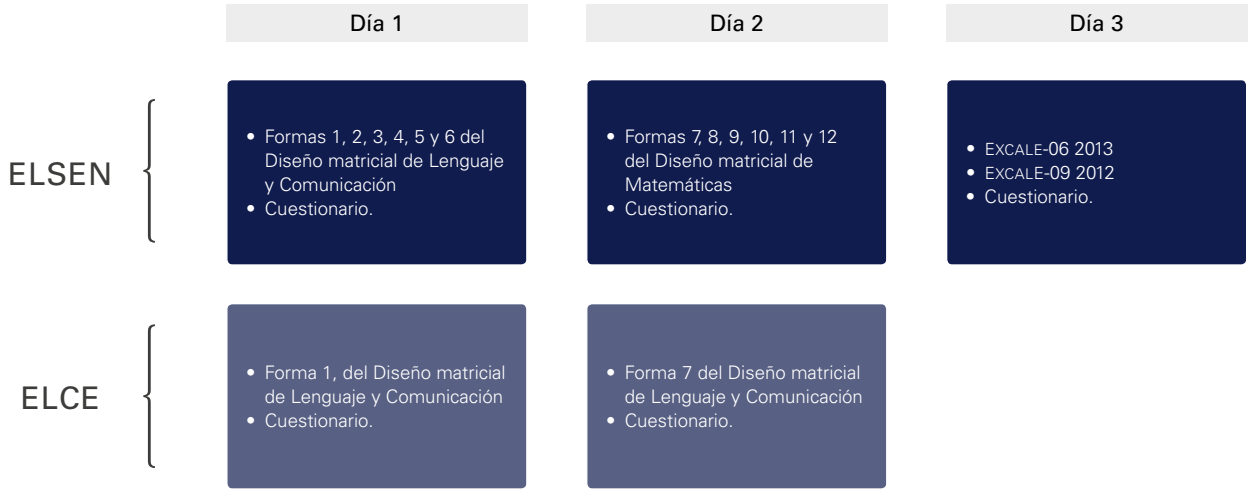
Modelos para la comparabilidad de resultados en el tiempo

Con el objetivo de definir una línea basal para las evaluaciones de PLANEA y construir la serie histórica de resultados a través del tiempo en la cual se incorporan los resultados obtenidos en los Exámenes de la Calidad y el Logro Educativos (EXCALE), en la aplicación de PLANEA de 2015, y por única ocasión, se administraron las evaluaciones EXCALE-06 2013 (sexto grado de primaria) y EXCALE-09 2012 (tercer grado de secundaria) de Español y Matemáticas a un subconjunto de alumnos de la muestra (aproximadamente 150 escuelas elegidas aleatoriamente) en un tercer día de aplicación. Esto permitió contar con una escala métrica común.

El procedimiento se conoce como *linking* y se utiliza principalmente para encontrar una regla de correspondencia aproximada entre instrumentos de evaluación que no se basan en las mismas especificaciones técnicas. Este procedimiento se utiliza en evaluaciones de logro educativo de gran renombre como la Evaluación Nacional del Progreso Educativo (NAEP, por sus siglas en inglés).

En la figura 9 se ejemplifica el esquema de aplicación para la muestra que sirve a los procedimientos de comparabilidad.

Figura 9. Diseño de aplicación para muestra de comparabilidad



Este ejercicio no implica que los resultados de las pruebas EXCALE y PLANEA sean directamente comparables, en particular porque miden constructos distintos y se expresan en diferentes escalas de logro. Únicamente permite ofrecer información cuantitativa a las autoridades educativas para hacer inferencias generales.

Difusión y uso de los resultados del instrumento de evaluación

La tarea de evaluación no consiste sólo en producir información relevante, sino también en difundirla amplia y oportunamente entre los actores interesados ofreciendo orientaciones precisas para que los resultados se comprendan, interpreten y utilicen de manera adecuada (INEE, 2016a). Estas acciones contribuyen a la promoción de una cultura de la evaluación, que la identifica como una estrategia útil para la mejora educativa.

Para que los propósitos formativos del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) se concreten es indispensable que los diferentes actores educativos y sociales conozcan y comprendan el contenido y el alcance de sus resultados. Existen dos páginas electrónicas en las que se incluyen fuentes de consulta y ligas para conocer documentos de referencia, manuales de procedimientos, fascículos informativos y los resultados de PLANEA: http://www.planea.sep.gob.mx/ y http://www.inee.edu.mx/index.php/planea

Para ampliar las interpretaciones que se hacen de los datos cuantitativos, los resultados de la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) se agrupan por niveles de logro, pues con ello se informa acerca de los conocimientos y las habilidades que poseen los estudiantes y si han alcanzado o no los aprendizajes clave del currículo. Las pruebas de ELSEN retoman las mismas escalas de puntuación y comparten los mismos contenidos que aquellas aplicadas en la modalidad de la Evaluación del Logro referida a los Centros Escolares (ELCE), por lo que sus resultados pueden compararse. Sin embargo, las pruebas de ELSEN no son comparables de manera directa con las pruebas de la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE) y de los Exámenes de la Calidad y el Logro Educativos (EXCALE), aplicadas en años anteriores.

En las siguientes páginas se explican los procedimientos que se implementaron a fin de determinar los niveles de logro y difundir los resultados para ser aprovechados por los diferentes actores educativos.

Descriptores de niveles de logro y puntos de corte

Con el propósito de contar con un sistema organizado de normas para la interpretación de los puntajes numéricos que se derivan de PLANEA, se decidió establecer cuatro niveles relativos al logro académico de los alumnos, independientemente del grado escolar o el campo disciplinar evaluado. Estos niveles se etiquetaron con los números romanos I, II, III y IV, siendo el IV el que sugiere un *logro sobresaliente de* los estudiantes. Por otra parte, el nivel III refiere de manera genérica a un *logro satisfactorio*; el nivel II, a un *logro apenas indispensable*, y el nivel I, a un *logro insuficiente* en sus aprendizajes clave.

Para establecer las fortalezas asociadas a cada nivel de logro, se llevaron a cabo reuniones con comités académicos que realizaron tareas complementarias:

1. Determinar los niveles de logro (cualitativos).
2. Establecer los puntos de corte (cuantitativos y asociados a los niveles determinados en la tarea 1).

Comité de niveles de logro y puntos de corte

Estos cuerpos colegiados estuvieron conformados por algunas de las personas que participaron en el diseño de la prueba con el propósito de aprovechar su conocimiento respecto a los análisis curriculares que se realizaron y las decisiones que se tomaron durante la selección y la operacionalización de los contenidos a evaluar. De igual manera, se contó con la participación de docentes en ejercicio provenientes de diferentes estratos educativos de diversas entidades federativas, además de especialistas en educación y personal técnico del Instituto Nacional para la Evaluación de la Educación (INEE).

En la tabla 17 se contabiliza a los participantes de estos comités por campo disciplinar y grado evaluado.

Tabla 17. Participantes en los comités de niveles de logro y puntos de corte, por grado evaluado y campo disciplinar

	6º primaria		3º secundaria	
	LyC	MAT	LyC	MAT
Total	9	14	9	18

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Los comités fueron responsables de definir teóricamente los niveles de logro por campo disciplinar. Para ello, elaboraron descriptores (definiciones operacionales) que reflejaron de forma global el tipo de aprendizaje que deben mostrar los estudiantes de cada nivel de logro.

Esta actividad requirió que los miembros de los comités realizaran un análisis cuidadoso de todos los reactivos de la prueba (y sus lecturas asociadas en el caso de Lenguaje y Comunicación), previamente ordenados por dificultad y seccionados por línea de evaluación por el personal técnico del INEE.

Para realizar este análisis se consideraron las habilidades y los conocimientos evaluados de acuerdo con las tablas de contenidos, pero también los procesos cognitivos involucrados en cada reactivo. Estos procesos pueden diferir en complejidad entre uno y otro ítem, aunque estén asociados al mismo ámbito o eje temático. Así, por ejemplo, un reactivo puede solicitar la búsqueda de un dato explícito en una lectura, o bien, de manera más compleja, puede requerir la interpretación de relaciones implícitas en diferentes partes de un texto.

A partir de un panel de discusión se acordaron:

- *Los descriptores que podrían corresponder a cada nivel, así como aquellos que pudieran considerarse limítrofes o que pertenecieran a dos niveles.*
  Para ello se solicitó a los miembros del comité que clasificaran los descriptores en cada uno de los niveles, y se identificaran los rasgos del descriptor que plantearan conflicto de clasificación al no ser claramente asimilables a una sola categoría.
- *El tipo de ejecución que implicaba cada nivel de logro.*
  Para facilitar esta tarea se tomó como referencia el análisis reticular realizado por el Comité Académico encargado del diseño de la prueba.

Para identificar los reactivos que marcan el punto de corte entre los diferentes niveles de logro para cada grado evaluado y campo disciplinar, se llevaron a cabo las siguientes tareas:

- Los participantes contestaron la totalidad de los reactivos de la prueba sin conocer los dominios o líneas de evaluación.
- Con este ejercicio se logró que los especialistas se familiarizaran con la prueba y comprobaran la calidad, la claridad, los niveles de dificultad, etc., de todos los ítems. Los reactivos se ordenaron de manera ascendente de acuerdo con los resultados de dificultad empírica de la aplicación 2015.
- Se revisaron los resultados del comportamiento estadístico de los reactivos en la aplicación nacional, así como los descriptores previamente elaborados para los niveles de logro.
- Se utilizó un formato de juicio en el que cada participante registró los reactivos marcadores (puntos de corte).

La tarea que se planteó a los participantes fue identificar los ítems que pertenecen a cada uno de los niveles de logro (comenzando por el ítem más fácil y por la categoría inferior), a partir de la pregunta: "¿un alumno del nivel X puede responder correctamente este ítem?" Este proceso requirió de dos o tres sesiones de juicio para que los especialistas compartieran argumentos acerca del grado de acuerdo/congruencia en el interior del cuerpo colegiado, y de las consecuencias que tendrían en los resultados nacionales los niveles y puntos de corte identificados.

La tabla 17 incluye los resultados nacionales de 2015 en términos de porcentaje de población, una vez que se aplicaron los puntos de corte.

Estas tablas comparativas pueden ser útiles cuando se interpretan tomando en cuenta variables de contexto como tipo de escuela, nivel de marginación, etcétera.

Tabla 18. Porcentaje de estudiantes por nivel de logro, para cada grado evaluado y por campo disciplinar, PLANEA 2015

	Primaria		Secundaria	
	LyC	MAT	LyC	MAT
Nivel I	49.5	60.5	29.4	65.4
Nivel II	33.2	18.9	46.0	24.0
Nivel III	14.6	13.8	18.4	7.5
Nivel IV	2.6	6.8	6.1	3.1

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Los reactivos marcadores acordados por los participantes sirvieron para redactar un descriptor final que conjuga los referentes empíricos con los dominios y las líneas de evaluación de PLANEA. En los cuadros 1 y 2 se puede revisar la progresión en los conocimientos y las habilidades que ilustran los descriptores conforme se avanza en el logro de aprendizaje de Lenguaje y Comunicación, y Matemáticas, respectivamente.

Los descriptores de logro están directamente relacionados con los aprendizajes clave que evalúan las pruebas y se incluyen en todos los reportes de resultados de PLANEA para indicar las fortalezas de los estudiantes que se ubican en cada nivel. La información puede utilizarse para generar mejores estrategias de enseñanza y aprendizaje, y para sugerir otras medidas de mejora si se interpretan en el marco de los datos que arrojan los cuestionarios de contexto. Algunas posibles interpretaciones, además del propio texto de los descriptores, se han compartido con el público en general por medio de los fascículos informativos, además del *Informe de resultados PLANEA 2015* (INEE, 2017a).

Cuadro 1. Descriptores de Lenguaje y Comunicación PLANEA 2015

Nivel	6° de primaria	3° de secundaria
I	Los alumnos son capaces de seleccionar información sencilla que se encuentra explícitamente en textos descriptivos. Además, comprenden textos que se apoyan en gráficos con una función evidente; distinguen los elementos básicos en la estructura de un texto descriptivo, y reconocen el uso que tienen algunas fuentes de consulta.	Los alumnos son capaces de identificar definiciones y explicaciones en artículos de divulgación científica y en anuncios publicitarios, la función y los recursos lingüísticos; de comprender el tema de un ensayo, y de identificar la rima en un diálogo teatral.
II	Los alumnos son capaces de comprender la información contenida en textos expositivos y literarios; distinguen los propósitos comunicativos de diferentes tipos de texto, y reconocen el lenguaje empleado al escribir cartas formales. Pueden elaborar inferencias simples, como el lenguaje figurado en un poema, y reconocen la estructura general de algunos textos literarios.	Los alumnos son capaces de reconocer la trama y el conflicto en un cuento e interpretar el lenguaje figurado de un poema. Organizan información pertinente y no pertinente para el objetivo de una encuesta, e identifican el propósito, el tema, la opinión y las evidencias en textos argumentativos.
III	Los alumnos son capaces de combinar y resumir información que se ubica en diferentes fragmentos de un texto, como en un mapa conceptual. Elaboran oraciones temáticas que recuperan la esencia del texto y la intención del autor. También relacionan y sintetizan información para completar un texto, pueden, por ejemplo, organizar la secuencia en un instructivo. Son capaces de realizar inferencias tales como interpretar el sentido de una metáfora en una fábula; contrastan el lenguaje de textos literarios, expositivos, periodísticos y apelativos, y pueden distinguir datos, argumentos y opiniones.	Los alumnos son capaces de interpretar hechos, identificar valores y comparar el tratamiento de un mismo tema en dos relatos míticos; reconocen las características sociolingüísticas de personajes en cuentos latinoamericanos, así como el ambiente y el contexto social en el que se desarrolla una obra teatral. Comparan géneros periodísticos y reconocen el tema en un artículo de divulgación científica. Además, pueden comprender el sentido de una oración a partir de los signos de puntuación.
IV	Los alumnos son capaces de comprender textos argumentativos, como el artículo de opinión, y pueden deducir la organización de una entrevista. Además, evalúan de manera conjunta elementos textuales y gráficos que aparecen en textos expositivos; sintetizan la información a partir de un esquema gráfico como un cuadro sinóptico, y establecen relaciones textuales que no son evidentes. Elaboran inferencias de alto nivel como evaluar el efecto poético, y analizan el contenido y la forma de textos con un tema similar. Por otra parte, discriminan el tipo de información que se solicita en un documento y reconocen las sutilezas entre el lenguaje de distintos textos.	Los alumnos son capaces de adaptar atributos biográficos a una obra de teatro y de seleccionar información relevante en un prólogo para utilizarlo en una reseña literaria. Pueden identificar secuencias argumentativas y valorar sus fundamentos en un ensayo, un artículo de opinión y un debate. Asimismo logran analizar la función de los pronombres en un texto.

Cuadro 2. Descriptores de Matemáticas PLANEA 2015

Nivel	6° de primaria	3° de secundaria
I	Los alumnos son capaces de escribir y comparar números naturales, y resolver problemas aplicando las características y propiedades básicas de triángulos, prismas y pirámides, así como aquellos que requieren leer información en gráficas de barras. Sin embargo, no son capaces de leer y realizar operaciones básicas con números naturales, representar gráficamente fracciones comunes ni identificar características como tipos de ángulos, alturas, rectas paralelas y perpendiculares en figuras y cuerpos geométricos. Tampoco pueden interpretar la descripción de una trayectoria, identificar la unidad de medida más adecuada para longitudes y áreas ni leer información explícita en gráficas de barras.	Los alumnos son capaces de resolver problemas usando estrategias de conteo básicas y comparaciones, o cálculos con números naturales. Pueden expresar en lenguaje natural el significado de fórmulas geométricas comunes y viceversa. Sin embargo, no son capaces de resolver problemas que impliquen: operaciones básicas con números decimales, fraccionarios y números con signo; el mínimo común múltiplo y el máximo común divisor, o los de valor faltante que suponen relaciones de proporcionalidad directa. Tampoco pueden calcular perímetros y áreas, o resolver ecuaciones de primer grado de la forma ax+b=c y sus expresiones equivalentes.
II	Los alumnos son capaces de leer números naturales, resolver problemas de suma con ellos, y multiplicarlos y dividirlos con decimales. Pueden representar una fracción en un modelo continuo, y reconocer la regla verbal y la pertenencia de un término a una sucesión aritmética creciente. Pueden identificar elementos geométricos como alturas, paralelas y ángulos rectos en figuras sencillas; resolver problemas utilizando las características y propiedades de cuadriláteros y pirámides; identificar unidades de medida de áreas, y resolver problemas de aplicación de perímetros. Son capaces de ubicar lugares usando sistemas de referencia convencionales en planos o mapas; resolver problemas de conversión de unidades en el Sistema Internacional de Medidas (SI), así como solucionar problemas que implican analizar o representar información en tablas o gráficas de barras, y de porcentaje y proporcionalidad del tipo "valor faltante" en diversos contextos dado el valor unitario.	Los alumnos son capaces de resolver problemas con números decimales, algoritmos elaborados como la raíz cuadrada y el máximo común divisor, y ecuaciones lineales sencillas. Pueden reconocer las relaciones de los ángulos de triángulos y los que se forman entre paralelas cortadas por una transversal, así como las secciones que se generan al cortar un cono. También son capaces de calcular el volumen de cuerpos con caras planas; reconocer y expresar, de diferentes formas, relaciones de proporcionalidad directa, y plantear relaciones sencillas de proporcionalidad inversa.
III	Los alumnos son capaces de leer y escribir números decimales, y resolver problemas aditivos con naturales o decimales y de multiplicación o división de naturales o decimales con naturales. Pueden representar una fracción en un modelo discreto, comparar fracciones y multiplicarlas por un natural. También pueden usar las fracciones para expresar una división e identificar el dividendo o divisor, así como sucesiones geométricas crecientes, a partir de la regla. Son capaces de resolver problemas utilizando las características y propiedades de ángulos, rectas, figuras y cuerpos geométricos; identificar situaciones de aplicación de perímetro; calcular la distancia real de un punto a otro en mapas, así como ubicar coordenadas y objetos en el plano cartesiano. Pueden resolver problemas directos de conversión de unidades de medida (SI e inglés) o que implican la lectura de información en portadores. Logran reconocer distintas formas de representar un porcentaje, y resolver problemas de identificación de la moda en un conjunto de datos y de proporcionalidad del tipo "valor faltante" en diversos contextos, sin dar el valor unitario.	Los alumnos son capaces de resolver problemas con números fraccionarios o con signo, o potencias de números naturales. Pueden sumar o restar expresiones algebraicas e identificar la ecuación o el sistema de ecuaciones que modelan una situación. Logran resolver problemas con el teorema de Pitágoras, la imaginación espacial (sólidos de revolución), propiedades de ángulos en círculos o triángulos, y relaciones de semejanza de triángulos. Son capaces de calcular el perímetro del círculo y de áreas de figuras compuestas; resolver problemas de cálculo de porcentajes o reparto proporcional, y modelar gráficamente un fenómeno que involucra únicamente funciones lineales.
IV	Los alumnos son capaces de comparar números decimales, y resolver problemas aditivos con números naturales, decimales y fraccionarios que implican dos o más transformaciones. Resuelven problemas que implican dividir o multiplicar números naturales por fraccionarios. Ubican una fracción en la recta numérica. Usan las fracciones para expresar el resultado de un reparto. Identifican el término siguiente en sucesiones especiales. Resuelven problemas de aplicación de áreas, así como de conversión de unidades de medida con una operación adicional. Describen rutas usando sistemas de referencia convencionales en planos o mapas. Resuelven problemas al usar información representada en tablas o gráficas de barras, de cálculo de promedio o de mediana, y de comparación de razones.	Los alumnos son capaces de calcular términos de sucesiones y multiplicar expresiones algebraicas, y resuelven problemas con números fraccionarios y decimales (combinados) usando notación científica, o una ecuación o un sistema de ecuaciones. Son capaces de solucionar problemas que suponen transformar figuras, propiedades de mediatrices, bisectrices y razones trigonométricas. Pueden calcular el área de sectores circulares y coronas, y el volumen de cuerpos redondos; resolver problemas usando estrategias de conteo; calcular la probabilidad de un evento simple, o abstraer información de tablas y gráficas. Logran modelar gráficamente un fenómeno que involucra funciones lineales y cuadráticas.

Reportes de resultados

La diseminación de los resultados, así como la promoción y la capacitación para su uso adecuado constituyen fases del proceso que resultan tan importantes como el diseño técnico de las pruebas (INEE, 2016a).

Los distintos tipos de reportes de PLANEA que se encuentran a disposición de actores educativos y sociales incluyen los resultados de las aplicaciones haciendo especial mención de los datos de contexto que deben acompañar cualquier interpretación que se derive de ellos. Enseguida se enlistan los reportes disponibles.

Reporte ELSEN

Incluye los resultados nacionales que logran los diferentes tipos de escuela del Sistema Educativo Nacional (SEN), con apoyos gráficos y textos breves. Permite apreciar las diferencias en los resultados entre los grupos sociales de contextos más y menos favorecidos con el fin de visibilizar las brechas que deberán reducirse para que todos los alumnos alcancen los aprendizajes. El INEE prepara el reporte y lo publica pocos meses después de la aplicación de la prueba.[14] Además, se presenta en el marco de la Conferencia del Sistema Nacional de Evaluación Educativa, integrada por autoridades educativas, federales y de las entidades.

Reportes ELCE

Los prepara la Secretaría de Educación Pública (SEP) con información distinta para cada tipo de destinatario:

a) Dirigidos a la comunidad escolar con los datos por escuela y de planteles similares en su entidad.
b) Dirigidos a supervisores escolares con datos de escuelas agregados y desagregados por zona escolar.
c) Dirigidos a las autoridades locales con datos de escuelas agregados y desagregados por región o municipio.

En el reporte ELCE para la comunidad escolar es posible revisar la cantidad de alumnos a los que se les aplicó la prueba, y si el porcentaje respectivo es representativo de la totalidad de estudiantes de la escuela. Posteriormente, y para cada campo disciplinar, se indica el número de alumnos que se ubicó en los diferentes niveles de logro. Las tablas con estos datos se acompañan de las descripciones de conocimientos y habilidades correspondientes a los niveles.[15]

[14] El reporte ELSEN 2015 se puede consultar en la página: http://publicaciones.inee.edu.mx/buscadorPub/P2/A/323/P2A323.pdf

[15] Los reportes ELCE se pueden consultar en: http://planea.sep.gob.mx/ba/resultados_anteriores/

Para fomentar comparativos válidos y que sean de utilidad para el centro escolar, el reporte incluye también dos cuadros en los que se pueden leer los porcentajes de alumnos en cada nivel de logro en la escuela; en otros centros con las mismas características en cuanto a tipo de escuela, entidad y grado de marginación; y en todas las escuelas de México.

Adicionalmente, se ofrece una explicación detallada de la importancia de considerar la información de contexto para comprender los resultados y reflexionar acerca de las medidas que pueden implementarse para apoyar el aprendizaje de los alumnos. En especial, se describen los indicadores de marginación y de Recursos Familiares Asociados al Bienestar (RFAB, para ubicarlos como escala en el cuestionario de contexto), y el modo en que éstos pueden impactar en el logro de aprendizaje. A partir de estas descripciones, y relacionándolas con los resultados de la escuela, la comunidad escolar puede reconocer si las tendencias en los porcentajes obtenidos para cada nivel de logro son similares a las que obtienen otros planteles en las mismas condiciones, y si es posible incidir de manera positiva en los estudiantes con los recursos que se tienen.

Además de estos reportes, la SEP difunde una serie de orientaciones que facilitan la comprensión y la interpretación de los resultados. Entre ellas, una guía para entender y analizar los reportes, y una revisión de las pruebas que incluye explicaciones acerca de las respuestas correctas y las incorrectas.

Bases de datos

Contienen el conjunto de información obtenida y procesada, de manera que sea posible realizar análisis complementarios a los que realizan la SEP y el INEE. El Instituto integra y difunde las bases de datos derivadas de ELSEN, mientras que la SEP hace lo propio con ELCE.[16]

Informe nacional

El INEE publica un informe nacional para dar cuenta del estado que guarda la educación respecto al logro de aprendizajes. Retoma los resultados más relevantes que se ofrecen en el reporte de la prueba ELSEN, y brinda mayor información sobre los elementos de contexto.[17]

Informes temáticos

La SEP, el INEE y otras instancias desarrollan informes centrados en algún tema en particular, apoyados en evidencias extraídas de los resultados de las evaluaciones.

[16] Bases de datos ELSEN 2015 en http://www.inee.edu.mx/index.php/planea/bases-de-datos-planea y ELCE 2015 en http://planea.sep.gob.mx/ba/base_de_datos_2015/

[17] Ver http://publicaciones.inee.edu.mx/buscadorPub/P1/D/246/P1D246.pdf

Importancia de los cuestionarios de contexto

Ya se ha mencionado que PLANEA considera, para la interpretación de resultados, los datos de los cuestionarios de contexto. Se parte del supuesto de que muchos factores y circunstancias influyen en el logro educativo de los estudiantes. Por una parte, se encuentra la historia personal y familiar que incluye las condiciones socioeconómicas, acceso a bienes culturales, condiciones para el estudio, situación laboral de la familia, pertenencia étnica, etcétera. Por otra parte, están las características de la escuela en términos de infraestructura, seguridad, organización, trayectorias y habilidades desarrolladas por sus profesores, así como circunstancias ambientales, sociales y políticas (INEE, 2015a).

Para contextualizar los resultados de PLANEA, en 2015 se aplicaron cuestionarios para alumnos, docentes y directores. En la tabla 19 se presentan la cantidad de reactivos y dimensiones de los cuestionarios, y algunas variables destacadas.

Su estructura completa puede consultarse en el documento rector de las pruebas.[18]

Tabla 19. Dimensiones de los cuestionarios de contexto PLANEA 2015

	Alumnos	Grupo escolar (docente)	Directores
Cantidad de reactivos	82	32	82
Dimensiones o ámbitos	Perfil del alumno Entorno familiar Entorno escolar	Expectativas académicas sobre los alumnos Atención a discapacidad	Perfil del director Entorno escolar Contexto de la convivencia
Variables destacadas	Variables sociodemográficas Antecedentes e intereses académicos Recursos familiares asociados al bienestar	Número de alumnos inscritos Prevalencia de ocho formas de discapacidad Condiciones para implementar acomodaciones	Variables sociodemográficas Perfil profesional Infraestructura y equipamiento escolar Organización escolar

Algunas de las variables que se utilizan comúnmente para analizar con mayor profundidad los resultados de logro son el tipo de escuela, el género, el tipo de sostenimiento, el nivel de marginación, la entidad federativa, el tamaño de la localidad donde está ubicada la escuela, la edad de los alumnos, entre otras.

Lo anterior permite contextualizar los datos y observar diferencias en el acceso a las oportunidades de aprendizaje para evitar hacer comparativos injustos o faltos de sustento. El *Informe de resultados PLANEA 2015* incluye diversos ejemplos para leer en un contexto adecuado todos los datos (INEE, 2017a).

[18] Ver http://planea.sep.gob.mx/content/general/docs/2015/PlaneaDocumentoRector.pdf

Los cuestionarios que se aplicaron a los alumnos incluyeron numerosos reactivos orientados a conocer sus características en los aspectos: personal, familiar, escolar y social, con la finalidad de obtener escalas que permitan analizar características no observables directamente (variables latentes). En el Anexo H se puede consultar el procedimiento de escalamiento y validación de escalas de los datos de los cuestionarios de contexto.

Por otra parte, en la página electrónica del Instituto se pueden consultar las preguntas que incluyen los cuestionarios de contexto y diversas bases de datos con los resultados.[19] Para facilitar su lectura, las bases se separan por nivel educativo y por protagonista del cuestionario, ya sean los alumnos, los directores, o los docentes titulares de los grupos escolares. Los archivos presentan, en una primera pestaña, un índice de tablas, la última fecha de actualización de los datos, e información de contacto de la Dirección de Tratamiento de Datos para que se recurra a ella en caso de requerir apoyo para la utilización de las bases. En las pestañas de resultados es posible leer las preguntas o reactivos, las categorías de respuesta y los porcentajes de alumnos ubicados en cada nivel de logro, de acuerdo con los estratos considerados en el diseño muestral.

Se reportan los datos de las poblaciones o subpoblaciones con representatividad muestral superior al 85%.

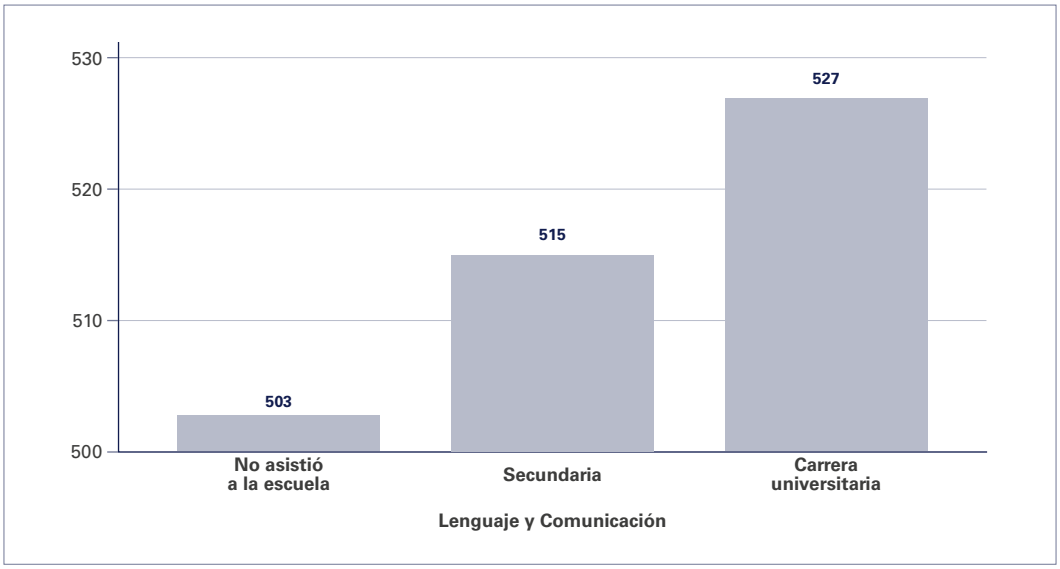
Todos los porcentajes de nivel de logro se acompañan de su error estándar, de la Unidad Primaria y la Secundaria de Muestreo (número de escuelas y número de alumnos en la muestra, respectivamente). A lo largo de las tablas, se encuentran señaladas las estimaciones con coeficientes de variación superiores a 20 y a 33%. En el primer caso sólo se hace la advertencia al lector para que reconozca que la estimación puede estar sesgada, mientras que en el segundo se omite el valor por haberse confirmado el sesgo.

Los datos de contexto tienen sentido en tanto se analicen de manera integral, es decir, no es suficiente hacer inferencias acerca de la influencia que puede tener en los resultados de logro, por ejemplo, el tamaño de la localidad, si no se relacionan los datos con los niveles de marginación o los tipos de escuela a las que asisten los alumnos. Algunas de las relaciones que pueden trazarse entre las variables que se incluyeron en los cuestionarios de contexto se ilustran en el *Informe de resultados* PLANEA *2015* (INEE, 2017a). Entre ellas, los datos de los estudiantes mexicanos confirman tendencias que se han estudiado previamente con otras pruebas a gran escala o instrumentos de investigación, como los que diseña la Organización para la Cooperación y el Desarrollo Económicos (OCDE).

Un ejemplo puede ser el impacto de la escolaridad de los padres o de las expectativas académicas de los alumnos, en sus resultados de logro. La figura 10 relaciona el puntaje promedio en Lenguaje y Comunicación, de los grupos de alumnos de tercero de secundaria, de acuerdo con la escolaridad de la madre.

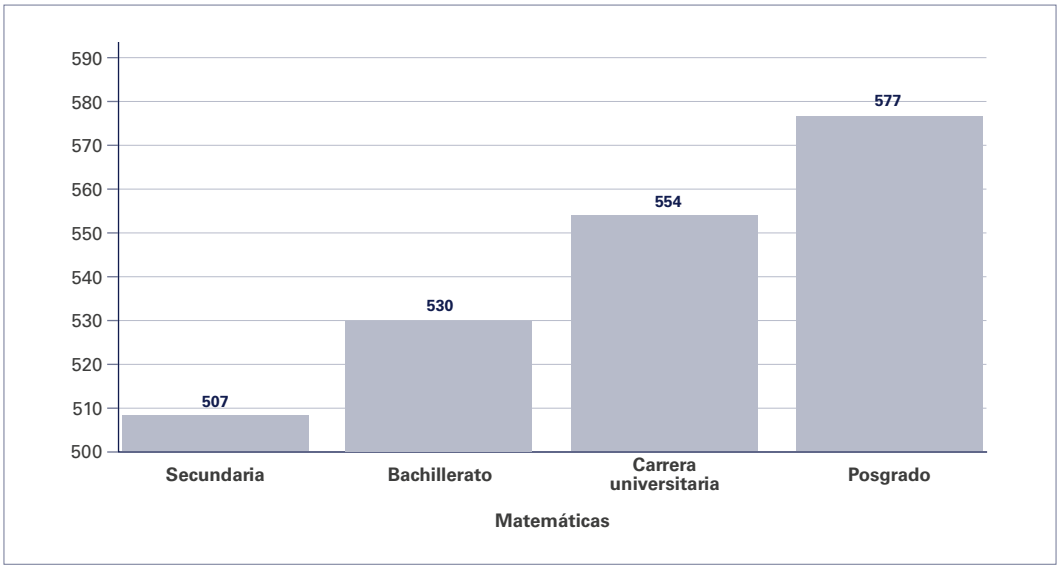
[19] Ver http://www.inee.edu.mx/index.php/planea/bases-de-datos-planea

Figura 10. Puntaje promedio de los alumnos de 3º de secundaria, Lenguaje y Comunicación. Escolaridad de la madre



Puede observarse que, a mayor escolaridad, mejores son los resultados de los alumnos. A su vez, la figura 11 permite observar que los puntajes más altos en Matemáticas de los alumnos de sexto de primaria se ubican en los alumnos con expectativas académicas mayores.

Figura 11. Puntaje promedio de los alumnos de 6º de primaria, Matemáticas. Expectativas académicas de los estudiantes



De acuerdo con la naturaleza de los datos que permitan obtener las escalas de los cuestionarios, es posible derivar recomendaciones para distintos actores educativos. En los casos que se acaban de ilustrar, la responsabilidad y posibilidad de incidir positivamente en los resultados de los alumnos reside en las autoridades educativas y políticas del país. Las oportunidades para extender la escolaridad de los padres, y entonces generar expectativas académicas más altas en sus hijos, deben potenciarse desde las instituciones públicas. De manera adicional a los beneficios económicos y sociales de impulsar la educación de los adultos, debe considerarse que entre mayores sean las oportunidades de aprendizaje y formación que tengan los padres, mejores herramientas tendrán ellos para solucionar las dudas de sus hijos o para ayudarlos con sus tareas.

Otros resultados de los cuestionarios de contexto pueden sugerir medidas remediales al interior de las escuelas, ya sea en la gestión escolar o en la práctica docente. Los datos de los cuestionarios aplicados a los tutores de grupo y a los directores que participaron en PLANEA 2015 sugieren, por ejemplo, que es necesario incrementar los materiales educativos con que cuentan los planteles para atender las necesidades especiales de los estudiantes; pueden ser aquellos con un problema de audición, visión o de aprendizaje. También sugieren la importancia de incrementar la capacitación de docentes y directores para promover y mantener una adecuada convivencia entre los alumnos, y fortalecer las redes de participación de toda la comunidad escolar.

Para complementar esta información y apoyar a los centros escolares en la implementación de mejoras, el INEE lleva a cabo la Evaluación de las Condiciones Básicas para la Enseñanza y el Aprendizaje (ECEA) con el objetivo de conocer en qué medida las escuelas de educación obligatoria del país cuentan con condiciones básicas para su operación y funcionamiento, tales como infraestructura, mobiliario, materiales de apoyo educativo, convivencia y organización escolar.[20] Las primeras tres condiciones, además de la idoneidad del personal con que cuentan las escuelas, se consideran recursos cuya distribución equitativa requiere de la intervención de las autoridades gubernamentales, en particular para las instituciones de sostenimiento público.

En complemento, lograr condiciones óptimas relacionadas con los ámbitos de gestión del aprendizaje, organización y convivencia escolar de cada plantel es responsabilidad de los directores, maestros, alumnos y padres de familia (INEE, 2016c).

En diversos fascículos informativos y en los reportes de resultados de ECEA se pueden reconocer algunas áreas de oportunidad y recomendaciones para apoyar el logro del aprendizaje de los estudiantes en la diversidad de escuelas del país.

[20] Ver http://www.inee.edu.mx/index.php/proyectos/ecea

Mantenimiento del instrumento de evaluación

Esta fase se considera indispensable cuando un instrumento de evaluación va a seguir funcionando después de una primera aplicación. Los resultados y análisis que se obtuvieron en la experiencia inicial permiten establecer un plan de trabajo para mejorar la calidad técnica de las pruebas y dar continuidad a las interpretaciones y los esfuerzos de evaluación en posteriores procesos de administración.

De acuerdo con los criterios técnicos (DOF, 2017, 28 de abril), para el mantenimiento de los instrumentos de evaluación, es necesario elaborar un plan de mejora que implica establecer un calendario con acciones encaminadas a perfeccionar el objeto de medida y robustecer los bancos de reactivos de los instrumentos.

Enseguida se describen las actividades que se han implementado al respecto tomando en consideración que, como parte de las decisiones estratégicas que determinó el Instituto Nacional para la Evaluación de la Educación (INEE) desde el inicio del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA), se consideran ajustes constantes para los instrumentos, y se plantea que en 2019 se efectúe una revisión del esquema para mejorar su diseño e implementación tras un ciclo completo de aplicación. Las modificaciones que ya se han hecho necesarias se reportan en el documento rector de la prueba, publicado en 2015 y actualizado en 2018 (INEE, 2015a; INEE, 2018).

Construcción de formas equivalentes

Una vez que las pruebas de la Evaluación del Logro referida a los Centros Escolares (ELCE) son divulgadas, las formas de examen que se seleccionaron no se vuelven a utilizar para evitar sesgos en los resultados, por lo que se implementa una estrategia de renovación y mantenimiento de las evaluaciones que permite hacer inferencias válidas y confiables a partir de los resultados de la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) y ELCE.

La estrategia implica que en las aplicaciones que se realizan cada año, a partir de 2017, se incluyan bloques de reactivos que reemplazan a los utilizados en las ELCE en el diseño matricial de las evaluaciones.

Los nuevos bloques de reactivos que se incorporen deberán calibrarse y ubicarse en la misma métrica definida en 2015. Para ello, las formas de examen del diseño matricial se aplican a una muestra de alumnos de aproximadamente 150 escuelas para efectuar el proceso de calibración. Este procedimiento se lleva a cabo en los años en donde no se aplique ELSEN. Así, en 2017 se calibraron los bloques A' y B' para sustituir los bloques A y B utilizados para la ELCE de 2015. En lo sucesivo, se tomará en cuenta el calendario de evaluaciones vigente para PLANEA.[21]

[21] Ver http://www.inee.edu.mx/index.php/planea

El esquema de renovación y mantenimiento permitirá añadir paulatinamente mejoras en los instrumentos e incorporar los cambios curriculares.

Informe técnico

El propósito de la redacción e integración de un manual técnico es contar con un documento que informe a la población en general, experta y no experta, acerca de los esfuerzos que se realizaron para atender adecuadamente los criterios técnicos que exigen los procesos de desarrollo de los instrumentos de evaluación (Downing y Haladyna, 2006).

En línea con las normas de calidad nacional e internacionales de las instancias más reconocidas en materia de evaluación, el Instituto publica el presente documento para complementar a los folletos informativos, fascículos e informes de resultados, detallando información técnica relevante como:

- Propósitos y alcances de la evaluación.
- Objeto de medición.
- Características técnicas y métricas de los instrumentos.
- Descripción de los procesos de construcción.
- Reglas de diseño.
- Algoritmo de puntuación.
- Reportes de resultados.
- Resultados de los estudios de validez.
- Materiales complementarios.

Por otra parte, de acuerdo con los criterios técnicos, se tiene planeada la actualización de este manual para reflejar las acciones de mantenimiento y los ajustes que se hagan en el marco teórico de referencia, en el objeto de medida o en el tipo de inferencias que se pueden realizar a partir de los resultados del instrumento.

Indicadores de validez

Como parte de las evidencias que se reportan en los informes técnicos para demostrar la solidez del proceso de evaluación y la validez de las inferencias que se realizan a partir de sus resultados, los criterios técnicos (DOF, 2017, 28 de abril) sugieren la necesidad de realizar estudios de diversa índole. La validez se refiere al grado en que la evidencia y la teoría apoyan las interpretaciones de los puntajes entregados por un test para un determinado propósito o uso (Manzi *et al.*, 2017).

Entre las pruebas que pueden realizarse están los análisis factoriales para demostrar que las variables latentes se agrupan en las dimensiones que se definieron desde el diseño, lo cual tiene que ver con la validez de contenido y de la estructura del instrumento (AERA, APA y NCME, 2014); o los estudios de validez convergente que pueden dar cuenta del poder predictivo del instrumento al demostrarse la relación de los resultados de la medición con otras variables teóricamente relacionadas. También pueden presentarse los resultados de un análisis de funcionamiento diferencial del instrumento o los reactivos para evidenciar la equidad de la medición (Manzi *et al.*, 2017). Esto se logra demostrando que los reactivos evalúan de manera homogénea a todos los subgrupos que componen la población objetivo.

Aquí se presentan los procedimientos y resultados de los estudios acerca de PLANEA 2015 que se han realizado en el Instituto. Además del análisis de funcionamiento diferencial de los reactivos, y un comparativo preliminar respecto a la relación entre los resultados de PLANEA y PISA (Programa para la Evaluación Internacional de los Estudiantes), se tiene previsto llevar a cabo otros estudios de validez de contenido que se reportarán de manera independiente en la página del Instituto.

Análisis de funcionamiento diferencial del instrumento

El funcionamiento diferencial de reactivos ocurre cuando diferentes grupos de estudiantes, que responden una prueba y tienen el mismo nivel de habilidad, exhiben diferente probabilidad de contestar correctamente un reactivo. En otras palabras, dicho reactivo está funcionando de manera distinta según el contexto y el grupo al que se aplique, aun cuando entre los participantes se comparta el nivel de habilidad.

Para sumar evidencias acerca de la validez del instrumento, es importante eliminar esta fuente de sesgo. El método implementado para analizar el Funcionamiento Diferencial de Reactivos (DIF) de las pruebas del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) y eliminar los reactivos con posibles defectos de construcción implicó comparar dos grupos de sustentantes (uno focal y otro de referencia), mediante la estimación de la dificultad de los reactivos para cada uno de los grupos (INEE, 2017b).

El primer análisis DIF que se realizó para PLANEA sexto de primaria implicó a los grupos de estudio definidos por la variable sexo, donde el grupo focal o minoritario fue la subpoblación de mujeres y el grupo de referencia fue la subpoblación de hombres. El segundo análisis DIF se realizó con los datos de la aplicación de tercero de secundaria para las subpoblaciones indígena (grupo focal) y no indígena (grupo de referencia). Estos análisis se llevaron a cabo para las pruebas de Lenguaje y Comunicación, y de Matemáticas. En todos los casos se determinó un mínimo de 500 participantes en cada grupo de estudio.

El modelo de Rasch para detectar DIF consistió en lo siguiente:

- Se estimaron los parámetros de dificultad de manera aislada en cada uno de los grupos de interés. La diferencia media en logro entre ambos grupos se controló fijando la dificultad promedio de los reactivos en cero:

$$\sum_{i=1}^{N} \widehat{\delta}_{i,R} = \sum_{i=1}^{N} \widehat{\delta}_{i,F} = 0$$

- Se calculó la diferencia en *logits* entre las dificultades de reactivos estimadas para ambos grupos, como una medida para determinar el DIF:

$$\widehat{\Delta}_i = \widehat{\delta}_{i,R} - \widehat{\delta}_{i,F}$$

- Se calculó el índice estandarizado que está dado por:

$$Z_i = \frac{\widehat{\Delta}_i}{\sqrt{(SE_{\hat{\delta}_{i,R}})^2 + (SE_{\hat{\delta}_{i,F}})^2}}$$

- Bajo la hipótesis de que no se presenta DIF, el valor esperado de $\widehat{\Delta}_i$ es 0, y Z_i tiene media 0 y varianza 1. Por lo cual, se reportaron solamente los reactivos que presentan una diferencia significativa, es decir, aquellos en donde: $Z_i \leq -1.96 \circ Z_i \geq 1.96$.
- Una vez identificados los reactivos con funcionamiento diferencial significativo, se procedió a clasificarlos en tres categorías, bajo el siguiente criterio (Wilson, 2005, p. 167):

Cuadro 3. Criterios DIF (aplicación operativa)

Categoría	Criterio
DIF despreciable "A"	$\left\|\widehat{\delta}_{i,R} - \widehat{\delta}_{i,F}\right\| < 0.426$
DIF intermedio "B"	$0.426 \leq \left\|\widehat{\delta}_{i,R} - \widehat{\delta}_{i,F}\right\| < 0.638$
DIF grande "C"	$\left\|\widehat{\delta}_{i,R} - \widehat{\delta}_{i,F}\right\| \geq 0.638$

En los casos en que la diferencia es significativa, es decir, se presenta DIF, el grupo focal de cada análisis se ve favorecido si $\widehat{\Delta}_i$ y Z_i son valores positivos. De esta forma fueron identificados todos los reactivos con DIF para que el equipo que elaboró las pruebas decidiera la inclusión o exclusión de los reactivos en el proceso de escalamiento.

Tomando en cuenta los resultados de estos análisis y los problemas de algunos reactivos en cuanto a discriminación, ajuste o forma de la curva característica del ítem, se determinó eliminar algunos de la calificación para asegurar que ésta cumpliera con condiciones de equidad y confiabilidad en la medición (tabla 20).

Los resultados representan una evidencia de la solidez con que fue construido el constructo y la confiabilidad de la medición, además del respeto a la diversidad que se concibió desde el diseño, el desarrollo y el ensamble de los instrumentos.

Tabla 20. Reactivos eliminados y utilizados para calificación después de calibración y análisis DIF, por grado evaluado y campo disciplinar

Grado	Campo disciplinar	Eliminados	Utilizados
6° primaria	LyC	7	143
	MAT	3	147
3° secundaria	LyC	7	143
	MAT	6	144

Nota: la abreviatura LyC hace referencia al campo disciplinar de Lenguaje y Comunicación, mientras que MAT refiere a Matemáticas.

Evidencia con base en las relaciones con otras variables

La aplicación de las pruebas PLANEA tercero de secundaria en 2015 coincidió con la de PISA que, en México, se aplicó a los estudiantes de 15 años. Esta población cursa, mayoritariamente, el tercer grado de secundaria, por lo que los resultados de ambas pruebas pueden ser comparados para buscar evidencias de validez. Esto es posible porque ambos instrumentos evalúan constructos similares, aunque es importante conocer las características de la prueba PISA para que las inferencias se hagan de manera adecuada.

Las pruebas PISA se aplican desde el año 2000 a muestras de alumnos de diferentes países que forman parte de la Organización para la Cooperación y el Desarrollo Económicos (OCDE). El objetivo es contar con datos comparables a nivel internacional para determinar qué es importante que los ciudadanos sepan y puedan hacer. Así, se lanzó un estudio trienal para evaluar hasta qué punto, los alumnos que están por concluir su educación obligatoria, cuentan con los conocimientos y habilidades fundamentales para participar en las sociedades modernas.

La prueba busca obtener información respecto a las capacidades de los estudiantes para reproducir, extrapolar y aplicar lo que han aprendido en las materias escolares básicas: Lectura, Matemáticas y Ciencias (OECD, 2016a).[22]

En este aspecto, PISA es similar a PLANEA, además de que su diseño de aplicación también es matricial e involucra a muestras representativas de estudiantes. Los resultados de PISA se expresan en seis niveles de logro y los de PLANEA en cuatro, coincidiendo en categorizar a los niveles 1 y I como los que reflejan un dominio por debajo de lo básico, o insuficiente con respecto a los aprendizajes clave. En cuanto a sus diferencias, PISA se aplica de manera electrónica y acompaña secciones de reactivos de opción múltiple con otras de respuesta construida. En ambos casos los participantes responden cuestionarios de contexto para conocer el entorno familiar, escolar y sus experiencias de aprendizaje.

Aunque en el reporte internacional y en los informes nacionales de 2015 se hace especial énfasis en los resultados de Ciencias, también es posible revisar los datos en Lectura y Matemáticas concernientes a la población mexicana. Al igual que en PLANEA, la media de desempeño a nivel internacional se fija en 500, lo que permite realizar comparaciones longitudinales entre países para derivar recomendaciones en cuanto a políticas públicas y educativas (OECD, 2016a). Ahora bien, para efectos de este apartado, el análisis de los resultados de ambas pruebas se centrará únicamente en la aplicación de 2015 y sobre todo en el aspecto cualitativo, puesto que, en el caso de PLANEA, sólo se cuenta con una sola medición y no es posible verificar que las tendencias sean las mismas que se han observado mediante las aplicaciones de PISA.

En la tabla 21 se muestra la descripción de las habilidades que poseen la mayoría de los estudiantes que participaron en PISA 2015 y PLANEA tercero de secundaria 2015 para el campo disciplinar de Lectura o Lenguaje y Comunicación (LYC).

La mayoría de los sustentantes de ambas pruebas tienen habilidades que se consideran por debajo de lo básico (nivel 1) o apenas indispensables (nivel II), de acuerdo con la definición de competencias básicas o aprendizajes clave de PISA y PLANEA, respectivamente.

Al respecto, 42% de los estudiantes mexicanos de 15 años que participaron en PISA y 46% de los participantes de PLANEA son capaces de identificar el tema principal o el propósito de un texto. También pueden reconocer información explícita y realizar asociaciones simples, similares a las que requieren establecer para comprender encuestas o documentos que utilizan en su entorno escolar o cotidiano.

[22] Ver marco de referencia en http://dx.doi.org/10.1787/9789264255425-en

Tabla 21. Habilidades de la mayoría de sustentantes PISA 2015 y PLANEA 3º de secundaria 2015, Lectura o Lenguaje y Comunicación

Prueba	Nivel de logro	Porcentaje nacional	Habilidades
PISA	Por debajo del nivel 2	42	El lector localiza uno o más piezas de información explícita, reconoce el tema principal o el propósito del autor en un texto que trata un tema familiar, o realiza asociaciones simples entre la información del texto y el conocimiento cotidiano. La información que se requiere comprender en el texto es evidente, y pocas veces exige utilizar recursos de interpretación complejos; el lector es dirigido de manera explícita para considerar los factores relevantes en la tarea y en el texto.
PLANEA	Nivel II	46	Los alumnos son capaces de reconocer la trama y el conflicto en un cuento e interpretar el lenguaje figurado de un poema. Organizan información pertinente y no pertinente para el objetivo de una encuesta, e identifican el propósito, el tema, la opinión y las evidencias en textos argumentativos.

Es interesante que, en la descripción de habilidades de PLANEA, se reconozca que este mismo porcentaje de estudiantes puede interpretar el lenguaje figurado de un poema o reconocer el conflicto en un cuento. De acuerdo con los niveles de PISA, esta interpretación tendría que darse en la lectura de poesía o de textos narrativos que no exijan recursos cognitivos complejos para su interpretación. Lo anterior está evidentemente ligado a las características de los textos y tareas que, en cada evaluación, es posible presentar a los sustentantes.

PLANEA organiza la prueba de Lenguaje y Comunicación en ámbitos (estudio, literatura, participación social), tipos de texto (como los que se trabajan en el aula) y prácticas sociales encaminadas a que los alumnos encuentren en la lectura un medio informativo y de recreación (INEE, 2015b). En el caso de PISA, y dado que se presentan reactivos de respuesta construida, es posible exponer a los alumnos a ejercicios que requieren de tareas cognitivas más elevadas, organizando la prueba de acuerdo con las situaciones o contextos en que la lectura tiene lugar, el tipo de texto (asociado a formatos continuos, discontinuos, fijos y dinámicos) y procesos cognitivos (OECD, 2016a).

Otros datos relevantes que presentan coincidencias evidentes entre PLANEA y PISA se remiten a las diferencias por género que muestran los resultados nacionales de cada prueba. En ambos casos las mujeres obtienen un puntaje promedio más alto (16 puntos de diferencia en PISA, y 28 en PLANEA), además de que mayores proporciones de hombres se ubican en el nivel más bajo de Lectura o Lenguaje y Comunicación (46% contra 37% de mujeres en PISA, 34% contra 24% en PLANEA).

Esta tendencia es opuesta a la que se observa en Matemáticas, donde tanto los hombres que participan en PISA como los que lo hacen en PLANEA obtienen mejores resultados que las mujeres: 16 puntos de diferencia en PISA y 8 en PLANEA. Lo anterior no evita que los

datos nacionales indiquen una seria deficiencia en la competencia matemática de acuerdo con ambas pruebas (tabla 22).

La mayoría de sustentantes de PISA y PLANEA presentan habilidades por debajo de lo básico e insuficientes para responder de manera adecuada a las tareas que, en el entorno cotidiano o escolar, requieren de conocimientos y destrezas matemáticas. Si bien los estudiantes pueden resolver problemas donde la información o las exigencias se presentan de manera directa, éstas deben remitirse a estrategias y rutinas básicas que, en PLANEA, se asocian únicamente con operaciones de cálculo con números naturales, o con la comprensión más rudimentaria de fórmulas geométricas. La descripción de habilidades de PLANEA va más allá y, para dar pistas a la comunidad escolar acerca de lo que se requiere reforzar en los estudiantes, también incluye lo que los alumnos no son capaces de realizar en este nivel I.

Tabla 22. Habilidades de la mayoría de sustentantes PISA 2015 y PLANEA 3º de secundaria 2015, Matemáticas

Prueba	Nivel de logro	Porcentaje nacional	Habilidades
PISA	Por debajo del nivel 2	57	Los estudiantes pueden responder problemas en contextos familiares donde la información relevante se presenta de manera directa y las preguntas están definidas claramente. Son capaces de identificar información y llevar a cabo procedimientos rutinarios en seguimiento a instrucciones directas y en situaciones explícitas. Pueden realizar acciones inmediatas cuando son obvias y siguen a determinado estímulo.
PLANEA	Nivel I	60.5	Los alumnos son capaces de resolver problemas usando estrategias de conteo básicas y comparaciones, o cálculos con números naturales. Pueden expresar en lenguaje natural el significado de fórmulas geométricas comunes y viceversa. Sin embargo, no son capaces de resolver problemas que impliquen: operaciones básicas con números decimales, fraccionarios y números con signo; el mínimo común múltiplo y el máximo común divisor, o los de valor faltante que suponen relaciones de proporcionalidad directa. Tampoco pueden calcular perímetros y áreas, o resolver ecuaciones de primer grado de la forma ax+b=c y sus expresiones equivalentes.

Como se puede ver en las tablas 21 y 22, la redacción de los niveles de logro de PLANEA está vinculada en mayor medida con las tareas específicas que se realizan en los reactivos de la prueba, mientras que la de PISA intenta referir habilidades y destrezas globales.

En cualquier caso, es posible notar una convergencia en los resultados puesto que, en ambas pruebas y para los dos campos disciplinares evaluados, se encuentra que la mayoría de la población mexicana no ha adquirido los conocimientos y habilidades que le permitirían actuar de manera exitosa a lo largo de su vida académica, personal y profesional. Las implicaciones son, entonces, similares, puesto que los datos alertan acerca de la urgencia para implementar acciones de política pública, educativa y gestión escolar que promuevan

más y mejores aprendizajes en los alumnos. En particular, los resultados de PISA y PLANEA señalan a la desigualdad de oportunidades y a la desventaja socioeconómica como factores asociados al logro en los aprendizajes (OECD, 2016a; OECD, 2016b; INEE, 2015c; INEE, 2017a). La nota de resultados por país que publicó la OCDE para PISA 2015[23] hace una descripción de los problemas que representa, en términos de oportunidades de aprendizaje, que existan problemas de ausentismo escolar, suspensión de labores, recursamiento o inmigración que están típicamente asociados al contexto socioeconómico. Por su parte, el informe de PLANEA incluye, para cada campo disciplinar y nivel educativo, un apartado referente a la equidad en el logro educativo, además de descripciones de las características contextuales y el impacto que tienen en los resultados aspectos como el trabajo infantil y la repetición escolar, y los diversos recursos familiares asociados al bienestar.

[23] Ver https://www.oecd.org/pisa/PISA-2015-Mexico-ESP.pdf

Conclusiones

Ante la necesidad de la sociedad mexicana de seguir contando con evaluaciones que sirvan de referente para diagnosticar el estado de la educación y los logros de los estudiantes, el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) surge como una herramienta que, retomando las experiencias previas en evaluación nacional e internacional, pretende ofrecer información de utilidad respecto a los aprendizajes al tiempo que implementa nuevas mecánicas en su diseño, su aplicación y su calificación. Estas innovaciones buscan asegurar la calidad técnica de los instrumentos, generar confianza en la población y acompañar a las comunidades escolares para que se supere la concepción de evaluación que sólo la valora como mecanismo de rendición de cuentas, y se aprovechen los resultados para transformar el día a día de los actos educativos en el país.

De acuerdo con Moreno (2011), esto requiere del fomento de una cultura de la evaluación que la conciba en su dimensión formativa, es decir, como un proceso de mejora de la educación. Desde su primera aplicación en 2015, PLANEA es consecuente con este propósito puesto que todas las actividades en torno a las diferentes modalidades de la prueba (Evaluación del Logro referida al Sistema Educativo Nacional —ELSEN— y Evaluación del Logro referida a los Centros Escolares —ELCE—) se organizan a fin de aportar elementos de información relevante para el monitoreo, la planificación y la operación del sistema educativo desde el ámbito gubernamental, el institucional y directamente en los centros escolares.

Desde esta perspectiva, no es importante el dato que arroja una sola prueba, o los resultados que se obtienen únicamente en términos de logro de los aprendizajes. El esquema de evaluación de PLANEA hace posible la integración periódica de información de diversas fuentes, además de su contextualización para dar sentido a los hallazgos y orientar los esfuerzos hacia una educación cada vez más equitativa y de mayor calidad.

Los resultados de PLANEA y de otros instrumentos que opera o supervisa el Instituto Nacional para la Evaluación de la Educación (INEE) buscan complementar los sistemas de evaluación internos que realizan las escuelas, las academias estatales, las direcciones regionales y las dependencias nacionales que se encargan de los diferentes niveles de la educación básica. Esto se logra con la difusión de resultados de las pruebas, pero, sobre todo, con la constante publicación de materiales informativos y la apertura para dialogar, reflexionar y analizar los datos y objetivos de las diferentes pruebas; para realimentar las prácticas pedagógicas; y para ampliar la visión que poseen las autoridades educativas, los directores y los docentes, de sus recursos y de los procesos educativos de los que son responsables (Manzi *et al.*, 2017).

Con la intención de que los logros y los resultados de las evaluaciones hagan copartícipes a todos los actores educativos, en el esquema de PLANEA se consideran responsabilidades compartidas entre el Instituto, la Secretaría de Educación Pública (SEP), las autoridades educativas y la comunidad escolar durante el desarrollo, la administración, la calificación y el control de las pruebas referidas al Sistema Educativo Nacional (SEN) y a los centros escolares.

Ésta puede ser una de las mayores virtudes de PLANEA, aunque también un área de oportunidad, ya que es necesario que año con año se perfeccionen los mecanismos de comunicación interinstitucionales y regionales, y se abran otros canales de diálogo para estar en contacto con otras instancias importantes para el éxito de las iniciativas de evaluación.

En este manual y en otros documentos técnicos se exhorta a los investigadores interesados, a los medios de comunicación y a la sociedad en general a que se comuniquen oportunamente las opiniones sobre las actividades alrededor de PLANEA y, en particular, acerca de los usos adecuados y no adecuados de sus resultados. Resulta especialmente deseable que se documenten los usos y prácticas sociales no previstas en el diseño de las pruebas para dialogar acerca de su pertinencia, de la calidad del esquema de evaluación nacional y de los elementos que pueden sumarse para contextualizar y aprovechar de mejor manera las estimaciones del logro de aprendizaje.

El propósito de PLANEA no es obtener datos precisos respecto a la habilidad individual de los estudiantes, lo que se pretende es adquirir estimaciones puntuales de grupos o poblaciones de estudiantes. La ventaja de esta aproximación es que permite monitorear el progreso de las poblaciones y evaluar un conjunto amplio de contenidos del currículo. El reto asociado es asegurar, año con año, que los procedimientos técnicos de diseño, desarrollo, aplicación, análisis y calificación de los instrumentos se llevan a cabo bajo rigurosos estándares de calidad.

Los procedimientos involucrados en las fases de aplicación y calificación adquieren una relevancia especial durante la elaboración de los informes técnicos, y de los resultados de las pruebas y los cuestionarios de contexto. La descripción y la explicación de los análisis estadísticos, los procedimientos de muestreo y de estimación, son esenciales para que los diferentes lectores de los informes hagan una interpretación adecuada de los conjuntos de datos que se derivan de PLANEA.

Con este propósito y para atender la obligación de transparencia, el INEE publica en su página electrónica todas las bases de datos, documentos técnicos y de divulgación, además de resultados y medios de contacto para consultar a su personal técnico en caso de requerir asistencia.

Por ahora, los esfuerzos de evaluación de PLANEA se han centrado en el nivel de logro en dos campos disciplinares clave: Lenguaje y Comunicación, y Matemáticas. Se tiene planeado incorporar progresivamente la evaluación en Ciencias y en Formación Cívica y Ética, además de continuar con la obtención de información relacionada con las habilidades socioemocionales de los estudiantes, y con el contexto social, escolar, familiar y personal que impacta en sus procesos de aprendizaje.

En todos los documentos asociados con las pruebas se puede notar la insistencia que hacen el Instituto y la SEP en torno a revisar los resultados de contexto para emitir inferencias acerca de los resultados de logro y las posibles acciones de apoyo a las instituciones y a los estudiantes. Aprovechar esta fortaleza de PLANEA se vuelve obligatorio en un país que se caracteriza por su diversidad en una multiplicidad de frentes. Siendo así, es importante matizar los juicios que se realizan, por ejemplo, sobre un dictamen satisfactorio o apenas

suficiente si éste se obtiene en entornos escolares con desventaja socioeconómica o con acceso limitado a otras oportunidades de aprendizaje. Lo mismo será cuando se cuente con los resultados de las aplicaciones posteriores a 2015 y se puedan hacer comparativos entre generaciones.

A partir de estos resultados y de los datos de estudios como el de Backhoff *et al.* (2015), se tendrá que considerar la adaptación de los instrumentos de tal forma que permitan evaluar de manera pertinente a la población hablante de lengua indígena o con necesidades educativas especiales.

Asimismo, será necesario actualizar el universo de medida de las pruebas en función de los nuevos planes y programas de estudio para la educación básica (SEP, 2017a).

Como parte del esquema de PLANEA, éstas y otras modificaciones se discuten de manera periódica y oportuna para mantener actualizado el Plan Nacional para la Evaluación de los Aprendizajes en la educación obligatoria.

En todo momento, se ha de tener presente que los mejores procesos evaluativos son aquellos que se inclinan por enfatizar el potencial que tiene la evaluación como factor de cambio, en tanto permite intercambiar información entre estudiantes, docentes, padres de familia, representantes de las instituciones, de los sistemas y con las autoridades educativas para que puedan planificar procesos de mejora. Lo anterior es por demás deseable puesto que hace notar el carácter formativo que debiera calificar a cualquier proceso de evaluación, ya sea que su propósito general sea evaluar poblaciones, centros escolares, programas o sistemas educativos, o si su objetivo específico es el diagnóstico, la certificación o la rendición de cuentas.

Bibliografía

Adams, R. J., Wilson, M., y Wang, W. C. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, *21*, pp. 1-23.

Adams, R. J., Wilson, M. R., y Wu, M. L. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioural Statistics*, *22*(1), pp. 46-75.

Adams, R. J., Wu, M. L., y Macaskill, G. (1997). Scaling Methodology and Procedures for the Mathematics and Science Scales. En Martin, M. y Kelly, D. (eds.). *TIMSS Technical Report Volume II: Implementation and Analysis (Primary and Middle School Years)*, pp. 111-145. Recuperado de: https://timssandpirls.bc.edu/timss1995i/TIMSSP-DF/TR2chap7.pdf

Adams, R., y Wu, M. (2002). PISA *2000 Technical Report*. Paris: OECD Publishing. AERA, APA y NCME (2014). *Standards for educational and psychological testing.* Washington, D. C.: AERA.

Andrich, D., DeJong, J., y Sheridan, B. (1997). Diagnostic opportunities with the Rasch model for ordered response categories. En: Rost, J. y Langeheine R. (Ed.), *Applications of latent trait and latent class models in the social sciences*. New York: Waxmann Verlag.

Angoff, W. (1993). Perspectives on differential item functioning Methodology. En: Holland, P. W. y Wainer, H. *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Asparouhov, T., y Muthén, B. (2009). Exploratory Structural Equation Modelling. *Structural Equation Modeling: A Multidiciplinary Journal*, *16*(3), pp. 397-438.

Backhoff, E., Solano, G., Contreras, L., Vázquez, M., y Sánchez, A. (2015) *¿Son adecuadas las traducciones para evaluar los aprendizajes de los estudiantes indígenas? Un estudio con preescolares mayas.* México: INEE.

Bentler, P., y Wu, E. (2014). *EQS 6.2 for Windows (Build 107)*. Multivariate Software, Inc.

Bock, R. D., y Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. *Psychometrika*, (46), pp. 443-459.

Bond, T. G., y Fox, C. M. (2001). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. London: Lawrence Erlbaum Associates, Inc.

Boomsma, A., van Duijn, M. A. J., y Snijders, T. A. B. (eds.) (2001). *Essays on Item Response Theory (Lecture Notes in Statistics)*. New York: Springer-Verlag.

Boone, W., Staver, J., y Yale, M. (2014). *Rasch Analysis in the Human Sciences*. Dordrecht, the Netherlands: Springer.

Bronzina, L., Chemello, G., y Agrasar, M. (2009). *Segundo Estudio Regional Comparativo y Explicativo. Aportes para la enseñanza de la Matemática.* Chile: Oficina Regional de Educación de la UNESCO para América Latina y el Caribe (OREALC/UNESCO Santiago) y Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación-LLECE.

Brown, T. (2015). *Confirmatory Factor Analysis for Applied Research* (2a ed.). New York: The Guilford Press.

Cattell, R. (1966). The Screen Test for the Number of Factors. *Multivariate Behavioral Research*, *1*(2), pp. 245-276.

Contreras R., S. y Backhoff E., E. (2014, 1 de octubre). Tendencias en el aprendizaje de la educación en México: una comparación entre ENLACE, EXCALE y PISA. *Revista Nexos*. Recuperado de: http://www.nexos.com.mx/?p=22749

Crocker, L. (2008). *Introduction to Classical and Modern Test Theory*. Ohio: Cengage Learning.

DGEP. Dirección General de Evaluación de Políticas (2015). *Lineamientos generales para la aplicación. Plan Nacional para la Evaluación de los Aprendizajes,* PLANEA *2015. Educación Básica*. México: SEP-SPEPE.

Dimitrov, D. M. (2012). *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields*. Alexandría: Wiley.

DOF. Diario Oficial de la Federación (2016, 1 de junio). Ley General de Educación. Recuperado de: https://www.gob.mx/cms/uploads/attachment/file/111212/LEY_GENERAL_DE_EDUCACI_N.pdf

DOF (2017, 28 de abril). Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. Instituto Nacional para la Evaluación de la Educación. México. Recuperado de: http://dof.gob.mx/nota_to_doc.php?codnota=5481062

Downing, S., y Haladyna, T. (eds.) (2006). *Handbook of test development.* Nueva Jersey: Lawrence Erlbaum Associates, Inc.

Dunn, G. (2004). *Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies*. Londres: Arnold.

ETS. Educational Testing Service (s. f.). *An adjustment for Sample Size in DIF Analysis*. Recuperado de: http://www.rasch.org/rmt/rmt203e.htm

Everitt, B. y Skrondal, A. (2010). *The Cambridge Dictionary of Statistics. Fourth Edition*. Cambridge: Cambridge University Press.

Guilleux, A., Blanchin, M., Hardouin, J., y Sébille, V. (2014). Power and Sample Size Determination in the Rasch Model: Evaluation of the Robustness of a Numerical Method to Non-Normality of the Latent Trait. *PLOS ONE*, 6.

Hair, J., Black, W., Babin, B. y Anderson, R. (2009). *Mutivariate Data Analysis* (séptima edición). London: Pearson.

Harrington, D. (2009). *Confirmatory Factor Analysis*. New York: Oxford, Univesity Press.

Hoyle, R. (1995). *Structural Equation Modeling. Concepts, Issues, and Applications*. Thousand Oaks, CA: SAGE Publications.

INEE. Instituto Nacional para la Evaluación de la Educación (2005). *Manual técnico para la construcción de reactivos*. México: autor. Recuperado de: http://www.inee.edu.mx/images/stories/Publicaciones/Documentos_tecnicos/De_pruebasymedicion/construccion_reactivos/Completo/mtconstrecexcalemarca.pdf

INEE (2014). *Criterios técnicos para el desarrollo y uso de instrumentos de evaluación educativa, 2014-2015*. México: autor. Recuperado de: http://www.inee.edu.mx/index.php/servicio-profesional-docente/513-reforma-educativa/1703-criterios-spd-2014

INEE (2015a). *Plan Nacional para la Evaluación de los Aprendizajes. PLANEA. Documento rector.* México: autor. Recuperado de: http://planea.sep.gob.mx/content/general/docs/2015/PLANEADocumentoRector.pdf

INEE (2015b). Folleto informativo PLANEA: Plan Nacional para la Evaluación de los Aprendizajes. México: autor. Recuperado de: http://www.inee.edu.mx/images/stories/2015/planea/PLANEA6.pdf

INEE (2015c). Plan Nacional para la Evaluación de los Aprendizajes (PLANEA). Resultados nacionales 2015. 6º de primaria y 3º de secundaria. Lenguaje y Comunicación. Matemáticas. México: autor. Recuperado de: http://publicaciones.inee.edu.mx/buscadorPub/P2/A/323/P2A323.pdf

INEE (2015d). *PLANEA 2015: Reporte del Cierre de Aplicación.* Unidad de Evaluación del Sistema Educativo Nacional, INEE, documento de trabajo.

INEE (2016a). *PLANEA. Una nueva generación de pruebas.* México: autor. Recuperado de: http://publicaciones.inee.edu.mx/buscadorPub/P2/A/321/P2A321.pdf

INEE (2016b). *Descripción del cálculo de estimaciones. PLANEA 2015*. Unidad de Evaluación del Sistema Educativo Nacional, INEE, documento de trabajo.

INEE (2016c). *Reporte general de resultados de la Evaluación de Condiciones Básicas para la Enseñanza y el Aprendizaje (ECEA) 2014/Primaria*. México: autor. Recuperado de: http://www.inee.edu.mx/images/stories/2016/ecea/resultadosECEA-2014actualizacion.pdf

INEE (2017a). *Informe de resultados PLANEA 2015. El aprendizaje de los alumnos de sexto de primaria y tercero de secundaria en México. Lenguaje y Comunicación. Matemáticas.* México: autor. Recuperado de: http://publicaciones.inee.edu.mx/buscadorPub/P1/D/246/P1D246.pdf

INEE (2017b). *Metodología de escalamiento de PLANEA 2015 para la Evaluación del Logro referida al Sistema Educativo Nacional, ELSEN*. Unidad de Evaluación del Sistema Educativo Nacional, INEE, documento de trabajo.

INEE (2018). *Plan Nacional para la Evaluación de los Aprendizajes. PLANEA. Documento rector.* México: Unidad de Evaluación del Sistema Educativo Nacional, INEE. Recuperado de: http://publicaciones.inee.edu.mx/buscadorPub/P1/E/305/P1E305.pdf

Johnson, Eugene G. y Rust, Keith F. (1992). Population Inferences and Variance Estimation for NAEP Data. *Journal of Educational Statistics*, *17*(2), pp. 175-190.

Jöreskog, K. (1993). Testing structural equation models. En: Bollen, K. A. y Long, J. S. (eds.), *Testing structural equation models*, pp. 294-316. Newbury Park, CA: SAGE Publications.

Judkins, D. (1990). Fay's Method for Variance Estimation. *Journal of Official Statistics*, *6*(3), pp. 223-239.

Kalton (1983). *Introduction to survey sampling*. Newbury Park: SAGE Publications.

Kaplan, D. (1989). Model Modification in Covariance Structure Analysis: Application of the Expected Parameter Change Statistic. *Multivariate Behavioral Research*, *24*(3), pp. 258-305.

Kaplan, D. (2000). *Structural Equation Modeling. Foundations and Extensions, Advanced Quantitative Techniques in Social Sciences*. Thousand Oaks, CA: SAGE Publications.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley y Sons.

Kline, R. (2005). *Principles and Practice of Structural Equation Modeling*. New York: The Guilford Press.

Lei, P. W. y Wu, Q. (2007). Introduction to Structural Equation Modeling: Issues and Practical Considerations. En: *Instructional Topics in Educational Measurement*, pp. 33-43.

Lepidus Carlson, B., Cox, B. G. y Bandeh, L. S. (2014). *SAS Macros Useful in Imputing Missing Survey Data.* Princeton, New Jersey: Mathematica Policy Research, Inc.

Linacre, J. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*, pp. 85-106.

Little, R., y Rubin, D. (2002). *Statistical Analysis with Missing Data, 2nd Edition.* New Jersey: John Wiley & Sons, Wiley series in probability and statistics.

Ludlow, L. H., y O'Leary, M. (1999, agosto). Scoring omitted and not-reached items: practical data analysis implications. *Educational and Psychological Measurement, 59*(4), pp. 615-629. Recuperado de: http://epm.sagepub.com/content/59/4/ 615.full.pdf+html

Manzi, J., García, M. R., y Godoy, M. I. (eds.) (2017). *Informe técnico SEPA. Sistema de Evaluación de Progreso del Aprendizaje.* Chile: Centro UC Medición, MIDE.

Martínez, R. F. (coord.) (2015). *Las pruebas ENLACE y EXCALE. Un estudio de validación.* México: INEE. Recuperado de: http://publicaciones.inee.edu.mx/buscadorPub/ P1/C/148/ P1C148.pdf

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*(2), pp. 149-174.

Millsap, R. y Olivera-Aguilar, M. (2012). Investigating Measurement Invariance Using Confirmatory Factor Analysis. En: Hoyle, R. *Handbook of Structural Equation Modeling*, pp. 380-392. New York: The Guildford Press.

Mislevy, R. (1991). Ramdomization-based inferences about latent traits from complex samples. *Psychometrika*, *56*(2), pp. 177-196.

Moreno O., T. (2011). La cultura de la evaluación y la mejora de la escuela. *Perfiles educativos*, *33*(131), pp. 116-130.

Mullis, I., Martin, M., González, E., y Chrostowski, S. (2004). *TIMSS 2003 International Mathematics Report. Findings From IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Massachusetts: International Association for the Evaluation of Educational Achievement y TIMSS & PIRLS International Study Center. Recuperado de: https://timssandpirls.bc.edu/PDF/t03_ download/T03INTLMATRPT.pdf

Murphy, M., y Schulz, W. (2006). *Sampling for National Surveys in Education*. Camberwell, Victoria: Australian Council for Educational Research.

Muthén, L. y Muthén, B. (2015). Mplus. Versión 7.31. Los Ángeles: Muthén & Muthén

OECD. Organisation for Economic Co-operation and Development (2005). *School sampling preparation manual* PISA *2006 main study*. París: OECD Publishing.

OECD (2009a). *PISA 2006 Technical Report*. París: OECD Publishing.

OECD (2009b). *PISA Data Analysis Manual: SAS*, 2a ed. París: OECD Publishing.

OECD (2014). *PISA 2012 Technical Report*. París: OECD Publishing.

OECD (2016a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. París: OECD Publishing. Recuperado de: http:// dx.doi.org/10.1787/9789264255425-en

OECD (2016b). *Nota País. Programa para la Evaluación Internacional de Alumnos (PISA). PISA 2015-Resultados México.* Recuperado de: https://www.oecd.org/pisa/PISA-2015-Mexico-ESP.pdf

Paek, I. y Wilson, M. (2011). Formulating the Rasch Differential Item Functioning Model Under the Marginal Maximum Likelihood Estimation Context and Its Comparison With Mantel–Haenszel Procedure in Short Test and Small Sample Conditions. *Educational and Psychological Measurement*, *71*(6), pp. 1023-1046.

Payton, J., Weissberg, R. P., Durlak, J. A., Dymnicki, A. B., Taylor, R. D., Schellinger, K. B., y Pachan, M. (2008). *The positive impact of social and emotional learning for kindergarten to eighth-grade students: Findings from three scientific reviews.* Chicago, IL: Collaborative for Academic, Social, and Emotional Learning.

Raykov, T., y Marcoulides, G. (2011). *Introduction to Psychometric Theory*. New York: Routledge.

Rubin, D. (1977). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Rutkowski, L., von Davier, M., y Rutkowski, D. (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. Florida: Chapman and Hall/CRC.

Santiago, P., McGregor, I., Nusche, D., Ravela, P., y Toledo, D. (2012). *OECD Reviews of Evaluation and Assessment in Education: Mexico 2012.* OECD Publishing. Recuperado de: http://dx.doi.org/10.1787/9789264172647-en

Särndal, C. E., Swensson, B., y Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Schulz, W., y Sibberns, H. (2004). *IEA Civic Education Study Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement.

SEP. Secretaría de Educación Pública (2011a). *Acuerdo número 592 por el que se establece la articulación de la Educación Básica*. México: autor.

SEP (2011b). *Plan de Estudios 2011. Educación Básica.* México: autor. Recuperado de: https://www.gob.mx/cms/uploads/attachment/file/20177/Plan_de_Estudios_2011_f.pdf

SEP (2016, 21 de julio). Modelo educativo y propuesta curricular. En: Blog Siete prioridades. Recuperado de: http://www.gob.mx/7prioridadessep/articulos/4-modelo-educativo-y-propuesta-curricular

SEP (2017a). *Aprendizajes clave para la educación integral. Planes y programas de estudio para la educación básica.* México: autor. Recuperado de: https://www.aprendizajesclave.sep.gob.mx/descargables/APRENDIZAJES_CLAVE_PARA_LA_EDUCACION_INTEGRAL.pdf

SEP (2017b). *Ruta para la implementación del modelo educativo*. México: autor. Recuperado de: https://www.gob.mx/cms/uploads/attachment/file/232636/10_Ruta_de_implementacio_n_del_modelo_educativo_DIGITAL_re_FINAL_2017.pdf

Sharon, L. (1999). *Muestreo: diseño y análisis*. México: Thomson.

StataCorp (2011). *Stata Statistical Software: Release 12. College Station*. Texas: StataCorp LP.

Streiner, D., y Norman, G. (2010). *Health Measurement Scales*, 3rd edition. Cambridge: Cambridge University Press.

Thompson, B. (2004). *Exploratory and Confirmatory Factor Analysis*. Washington, D. C.: American Psychological Association.

Valenzuela, J. R., Ramírez, M. S., y Alfaro, J. A. (2010). Cultura de evaluación en instituciones educativas. Comprensión de indicadores, competencias y valores subyacentes. *Perfiles educativos*, 33(131), pp. 42-63.

Van der Linden, W. J., y Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer Science + Business Media, LLC.

Wang, J., y Wang, X. (2012). *Structural Equation Modeling: Applications Using Mplus*. Reino Unido: Wiley.

Whittaker, T. (2012). Using the Modification Index and Standardized Expected Parameter Change for Model Modification. *The Journal of Experimental Education*, 80(1), pp. 26-44.

Wilson, M. (2005). *Constructing Measures. An Item Response Modeling Approach*. London: Lawrence Erlbaum Associates.

Wright, B. (1977). Solving measurement problems with de Rasch Model. *Journal of Educational Measurement*, *14*, pp. 97-116.

Wright, B. (1994). Rasch sensitivity and Thurstone insensitivity to graded responses. *Rasch Measurement Transactions*, *8*, pp. 382-383.

Wright, B. D., Mead, R. J., y Draba, R. (1976). *Detecting and correcting test item bias with a logistic response model (Research Memorandum No. 22)*. Chicago: MESA Psychometric Laboratory.

Wright, B. D., y Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., y Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

Wu, M., y Adams, R. (2007). *Applying the Rasch Model to Psycho-Social Measurement. A Practical Approach*. Melbourne: Educational Measurement Solutions.

Wu, M., Adams, R., y Haldane, S. (2007). *ACER Conquest® [Computer Software]*. Hawthorn, Victoria, Australia: ACER Press. Australian Council of Educational Research. University of California Berkeley.

Wu, M., Adams, R., Wilson, M., y Haldane, S. (2007). *ACER ConQuest Version 2: Generalised item response modelling software*. Camberwell: Australian Council for Educational Research.

Siglas y acrónimos

AEE	Áreas Estatales de Evaluación
AERA	American Educational Research Association
APA	American Psychological Association
CENEVAL	Centro Nacional de Evaluación para la Educación Superior, A. C.
CÍVICA	Estudio Internacional de Educación Cívica y Ciudadana
CONAFE	Consejo Nacional de Fomento Educativo
CONAPO	Consejo Nacional de Población
CPEUM	Constitución Política de los Estados Unidos Mexicanos
DGEI	Dirección General de Educación Indígena
DGEP	Dirección General de Evaluación de Políticas (de la SEP)
DIF	Funcionamiento Diferencial del Reactivo o Ítem (por sus siglas en inglés)
DOF	Diario Oficial de la Federación
ECEA	Evaluación de Condiciones Básicas para la Enseñanza y el Aprendizaje
EDC	Evaluación Diagnóstica Censal
ELCE	Evaluación del Logro referida a Centros Escolares
ELSEN	Evaluación del Logro referida al Sistema Educativo Nacional
ENLACE	Evaluación Nacional del Logro Académico en Centros Escolares
ETS	Educational Testing Service
EXCALE	Exámenes de la Calidad y el Logro Educativos
IEA	Asociación Internacional de Evaluación
INEE	Instituto Nacional para la Evaluación de la Educación
INEGI	Instituto Nacional de Estadística y Geografía
LGE	Ley General de Educación
LINEE	Ley del Instituto Nacional para la Evaluación de la Educación
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LyC	Lenguaje y Comunicación
MAT	Matemáticas
NAEP	Evaluación Nacional del Progreso Educativo (por sus siglas en inglés)
NCME	National Council on Measurement in Education
OCDE	Organización para la Cooperación y el Desarrollo Económico
OECD	Organisation for Economic Co-operation and Development
PIRLS	Estudio Internacional de Progreso en Comprensión Lectora (Progress in International Reading Literacy Study)
PISA	Programa para la Evaluación Internacional de los Estudiantes (Programme for International Student Assessment)
PLANEA	Plan Nacional para la Evaluación de los Aprendizajes
PP	Puntaje Promedio
RFAB	Recursos Familiares Asociados al Bienestar
SEN	Sistema Educativo Nacional
SEP	Secretaría de Educación Pública

SERCE	Segundo Estudio Regional Comparativo y Explicativo
SPEPE	Subsecretaría de Planeación y Evaluación de Políticas Educativas (de la SEP)
TIMSS	Tercer Estudio de las Tendencias en Matemáticas y Ciencias (Trends in International Mathematics and Science Study)
UNEG	Grupo de Evaluación de las Naciones Unidas (por sus siglas en inglés)
UPM	Unidad Primaria de Muestreo

Glosario de términos

A

Aprendizaje. Proceso mediante el cual se adquieren habilidades, destrezas y conocimientos como resultado de la experiencia, la instrucción o la observación.

C

Calidad del Sistema Educativo Nacional. Cualidad que resulta de integrar las dimensiones de pertinencia, relevancia, impacto, suficiencia, eficiencia, equidad y eficacia interna y externa.
Confiabilidad. Cualidad de las mediciones obtenidas con un instrumento, que se caracterizan por ser consistentes y estables cuando éste se aplica en distintas ocasiones.
Constructo. Elaboración teórica formulada para explicar un proceso social, psicológico o educativo cuya adecuada descripción permite que sea susceptible de ser observable o medible.

D

Directrices. Son aquellas normas o principios que proveen orientación para tomar decisiones de política educativa. Se construyen con base en evidencias que resultan de investigaciones y evaluaciones educativas. En el contexto de la nueva Ley del Instituto Nacional para la Evaluación de la Educación, hacen referencia a las recomendaciones que emite el Instituto para la autoridad educativa.

E

Equidad. Cualidad que consiste en dar a cada uno lo que se merece en función de sus méritos, necesidades o condiciones.
Escala. Puntuaciones o medidas asignadas a objetos o sucesos a partir de un modelo de medición.
Especificaciones de reactivos. Descripción detallada de las características o conductas relevantes que se esperan de los sujetos al sustentar el instrumento de evaluación, las cuales se pueden observar por medio de las tareas evaluativas o los reactivos. Tienen el papel de guiar en la elaboración y validación de los reactivos para que cuenten con los elementos necesarios para construirlos alineados al objeto de medida o constructo que se desea evaluar mediante el instrumento.
Estándar. Principio de valor o calidad en la conducción y uso de los procedimientos de evaluación, que es acordado por expertos en evaluación.
Estándar de desempeño. Es un criterio conceptual que delimita el marco interpretativo de las puntuaciones obtenidas en un instrumento de evaluación y que refiere a lo que la persona evaluada es capaz de hacer en términos de conocimientos, destrezas o habilidades en el contexto del instrumento.

J

Jueceo. Método en el cual se utiliza la opinión de expertos (denominados jueces) para determinar, entre otras cosas, la pertinencia de la validez de las tareas evaluativas o los reactivos respecto a un dominio; el establecimiento de estándares o puntos de corte; así como la calificación de reactivos de respuesta construida.

M

Matricial. Tipo de prueba diseñada para evaluar la mayor cantidad de aprendizajes del currículo, para lo cual se divide la totalidad de los reactivos en distintas versiones relacionadas entre sí. Cuando estas formas se aplican al universo de alumnos permiten conocer lo que han aprendido en conjunto acerca de los contenidos.
Muestra. Subconjunto de la población de interés que refleja las variables medidas en una distribución semejante a las de la población.

N

Necesidades educativas especiales. Características de los alumnos, físicas, cognitivas o de otro tipo, que exigen de recursos didácticos especiales, y de la elaboración y aplicación de adaptaciones curriculares.

0

Objeto de medida. Conjunto de características o atributos que se miden en el instrumento de evaluación.
Operacionalizar. Proceso de definición y disgregación de conceptos o variables, en términos claros y concretos que pueden ser verificados de manera empírica.

P

Prueba estandarizada. Instrumento de medición que se desarrolla, aplica y califica siguiendo procedimientos determinados.
Punto de corte. En instrumentos de evaluación con referencia a un estándar de desempeño, es la puntuación mínima o criterio para considerar que el nivel de desempeño de una persona cumple con lo esperado y distinguirlo de otro que no.

S

Sesgo. Distorsión que ocurre cuando, por variables no controladas ajenas a lo que se quiere medir, se obtienen resultados más altos o más bajos que los que se obtendrían con una medición correcta.

V

Validez. Juicio valorativo integrador respecto al grado en que los fundamentos teóricos y las evidencias empíricas apoyan la interpretación de las puntuaciones de los instrumentos de evaluación.
Variable latente. Término estadístico empleado para denotar una característica o cualidad que se quiere medir en un objeto de medición, pero que no es observable de manera directa.

Anexos

A. Objeto de medida de las pruebas de Lenguaje y Comunicación

La evaluación del área de Lenguaje y Comunicación del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) explora en los alumnos un conjunto de aprendizajes clave relacionados con el uso del lenguaje para organizar el pensamiento y el discurso; con la capacidad para leer, comprender, emplear, reflexionar e interesarse en diversos tipos de texto con el fin de ampliar los conocimientos, así como descubrir la importancia del lenguaje en la construcción del conocimiento y de los valores culturales.

Para evaluar este objeto de medida, en el caso de primaria se elaboraron 122 especificaciones, de las cuales 93 se consideraron "esenciales" y 29 "muy importantes" para evaluar el aprendizaje de los estudiantes. Para secundaria se hicieron 100 especificaciones, de las cuales 44 se juzgaron "esenciales", 21 "muy importantes" y 35 "importantes". En ambos niveles se elaboraron especificaciones con más de un reactivo, de manera que las pruebas incluyen un total de 150.

Las especificaciones se construyeron considerando los tres ámbitos señalados en el programa de la asignatura: estudio, literatura y participación social; y según los tipos de texto que se utilizan cotidianamente en el aula y fuera de ella para practicar los aprendizajes esperados.

Ámbito	Descripción
Estudio	Las prácticas agrupadas aquí tienen el propósito de apoyar a los alumnos en el desempeño de sus estudios, para que puedan expresarse oralmente y por escrito en un lenguaje formal y académico.
Literatura	En este ámbito las prácticas se organizan alrededor de la lectura compartida de textos literarios para transitar de una construcción personal y subjetiva del significado a una más social o intersubjetiva; ampliar los horizontes socioculturales, y aprender a valorar distintas creencias y formas de expresión.
Participación social	El propósito de estas prácticas es favorecer el desarrollo de otras formas de comprender el mundo y actuar en él. También comprende el desarrollo de una actitud crítica ante la información que se recibe de los medios de comunicación.

Tabla A1. Número de especificaciones de las pruebas de Lenguaje y Comunicación 6° de primaria y 3° de secundaria, por ámbitos, prácticas sociales del lenguaje y tipos de texto

Ámbito	Las prácticas sociales en este ámbito están encaminadas a:	Tipos de texto		Cantidad de especificaciones	
		Primaria	Secundaria	6° primaria	3° secundaria
Estudio	• Introducir a los alumnos a textos académicos tanto para apoyar su aprendizaje en diferentes disciplinas como para que aprendan a emplear los recursos de los textos expositivos con el fin de buscar y seleccionar información. • Desarrollar en los alumnos habilidades para escribir textos que les permitan recuperar información, organizar sus ideas y expresarlas clara y ordenadamente con base en la información que obtuvieron de la lectura. • Desarrollar en los alumnos habilidades de expresión oral por medio de su participación en eventos comunicativos formales, como exposiciones y debates, entre otros, en los que presentan sus conocimientos en sesiones organizadas.	Noticia con cita Artículo de divulgación Listado de preguntas para entrevista Entrevista Reportaje Reportaje con tabla Cuestionario Nota enciclopédica Texto enciclopédico Dos textos: académico e informal Índice Monografía de grupos indígenas	Artículo de divulgación científica Entrevista Monografía con gráfica Reportaje con gráfica Discusión constructiva Debate	46	30
Literatura	• Poner en contacto a los alumnos con la literatura infantil para darles un panorama más amplio de textos literarios y que logren recrearse con ellos. • Promover que compartan sus experiencias de lectura, hagan recomendaciones y tomen sugerencias de otros para elegir textos literarios. • Invitar a los alumnos a producir textos originales en los que puedan expresar su imaginación y usar los recursos lingüísticos de la literatura.	Relato histórico Fábula Poema Biografía Cuento de terror Cuento de misterio con diálogos Obra de teatro Diario personal	Cuento Mito Poema vanguardista Obra de teatro Obra de teatro del Siglo de Oro Texto para dramatizarse Prólogo	35	31
Participación social	• Que los alumnos empleen diferentes tipos textuales de la vida cotidiana para adquirir estrategias para consultar y usar periódicos, agendas, recibos, formularios, reglamentos, etcétera.	Anuncio Debate Instructivo Artículo de opinión Carta formal Cuatro textos Croquis Formulario Recado	Reglamento Formulario Diversidad lingüística Noticia y columna de un mismo tema Escaleta de radio Mensaje publicitario Artículo de opinión Encuesta Gráfica Documento administrativo	41	39
			Total de especificaciones de la prueba	122	100

En la tabla A1 se indica la cantidad de especificaciones de acuerdo con ambos criterios, y se señala en qué consiste el desarrollo de las prácticas sociales del lenguaje en cada uno de los ámbitos del programa.

Las especificaciones también se organizaron por unidades de evaluación. Estas unidades se identifican con los procesos cognitivos o habilidades de comprensión lectora y reflexión sobre la lengua que evalúan las pruebas PLANEA. Se considera que el desarrollo de estas habilidades es fundamental para la adquisición de los aprendizajes clave del currículo.

En el cuadro siguiente se incluye la definición de las unidades de evaluación que se requiere poner en práctica para exhibir una adecuada comprensión lectora.

Unidad	Definición
Extracción de información	El alumno debe obtener determinados datos de un texto; por ello busca, localiza y selecciona información relevante, o hace uso de información específica para cumplir una demanda. Los alumnos deben relacionar la información indicada en una pregunta con la que se presenta en el texto, la cual puede ser idéntica o redactada con sinónimos. Para lograr su cometido, los alumnos acceden a un espacio de información en donde se ubican los datos que necesitan; recorren ese espacio en búsqueda de la información requerida hasta encontrarla, la seleccionan y finalmente la extraen.
Desarrollo de una comprensión global	El alumno debe considerar el texto como una unidad y entender su función y propósito comunicativo, así como el tema, el contenido y la coherencia global del material leído. Debe ver el texto de manera integral, con una perspectiva que le permita captar algunas ideas generales, además de seleccionar de él lo más relevante. En relación con este proceso, el alumno requiere realizar un enlace entre un fragmento del texto y una pregunta, así como deducir el tema principal a partir de la repetición de una categoría particular de información. En este proceso de jerarquización entre ideas principales y secundarias, el alumno construye una representación del significado global del texto.
Desarrollo de una interpretación	Con base en la asociación de dos o más fragmentos del texto, el alumno debe construir una idea. La información que se debe vincular está asentada en el material de lectura, pero las relaciones entre la información pueden no ser explícitas; los alumnos demuestran que se apoyan en la cohesión y la coherencia del texto al interpretar información explícita, al reconstruir información implícita y realizar inferencias para su interpretación, o al establecer relaciones textuales y extra textuales. Algunas de las actividades que se realizan gracias al establecimiento de inferencias son: el esclarecimiento del significado de las partes del texto; la elaboración de interpretaciones para entender el mensaje y la perspectiva del autor; el desarrollo de una lectura interpretativa entre líneas advirtiendo ciertas pistas implícitas en el texto que informan al lector sobre posibles significados contextuales y sobre la mirada del autor.
Análisis del contenido y la estructura	El alumno debe saber cómo se desarrolla el texto y reflexionar sobre su contenido, organización y forma. Examinar el contenido y la estructura del texto implica evaluarlo, compararlo y contrastarlo, además de entender el efecto que tiene sobre el lector. Este proceso requiere que el alumno conecte la información encontrada en el contenido con el conocimiento externo, la cual puede provenir del propio texto o de otras ideas ofrecidas explícitamente en la pregunta. Este proceso da cuenta del impacto de ciertas características textuales y de su organización lógica.
Evaluación crítica del texto	El alumno debe alejarse del texto para evaluarlo de manera crítica, compararlo y contrastarlo contra una representación mental, además de entender el efecto que tienen la estructura, la forma y el contenido sobre la audiencia, para después hacer un juicio. Incluye la capacidad para descubrir los casos donde el texto proporciona un punto de vista parcial y una tendencia, y para reconocer el uso de técnicas de persuasión.

A su vez, también se elaboraron las siguientes definiciones para las unidades de evaluación asociadas a la habilidad de reflexión sobre la lengua.

Unidad	Definición
Reflexión semántica	El alumno debe comprender la noción de clases de palabras y reflexionar sobre su uso y el significado que éstas aportan al texto; establecer relaciones semánticas, gráficas y morfológicas entre palabras; dilucidar el significado de palabras, frases y expresiones en el contexto de un material escrito; interpretar relaciones semánticas entre oraciones o elementos oracionales unidos por enlaces o marcadores discursivos; identificar el significado que un término adquiere dentro de un texto; reconocer el artículo, el pronombre, el adjetivo o el verbo que completa un enunciado; conocer el significado y los cambios de sentido de las palabras o de su organización (antónimos, sinónimos, prefijos y homónimos), así como apreciar el cambio en el significado o el sentido de las oraciones al realizar permutaciones en el orden de las unidades.
Reflexión sintáctica y morfosintáctica	El alumno debe explorar diversos aspectos de la estructura del lenguaje escrito y reflexionar sobre su uso: las partes de la oración; los diferentes tipos de oraciones; los verbos y tiempos verbales predominantes en una redacción, y el establecimiento de concordancia de género, número, persona y tiempo en las oraciones de un texto.
Reflexión sobre la convencionalidad de la lengua	El alumno debe reflexionar sobre la puntuación, la acentuación, la ortografía y la segmentación de palabras y su importancia para la construcción del significado de un texto y su legibilidad, así como reconocer el orden alfabético como organizador de contenidos y secuencias.
Conocimiento de fuentes de información	El alumno debe ser capaz de identificar elementos editoriales de las fuentes de información como edición, editor, año de publicación, para referir una fuente o para valorar su importancia en un texto. Asimismo, debe ser capaz de seleccionar una fuente de información para hacer consultas de diversa índole: ortográficas, significados, integración o verificación de información.

En la tabla A2 se presenta la distribución de especificaciones de las pruebas de Lenguaje y Comunicación, por unidades de evaluación.

Tabla A2. Número de especificaciones de las pruebas de Lenguaje y Comunicación 6º de primaria y 3º de secundaria, por unidad de evaluación

Unidades de evaluación		Número de especificaciones	
		Primaria	Secundaria
Comprensión lectora	Extracción de información	7	2
	Desarrollo de una comprensión global	21	20
	Desarrollo de una interpretación	25	14
	Análisis de contenido y estructura	33	27
	Evaluación crítica del texto	0	20
Reflexión sobre la lengua	Reflexión semántica	11	13
	Reflexión sintáctica y morfosintáctica	7	2
	Convencionalidades lingüísticas	14	0
	Conocimiento de fuentes de información	4	2
	Total de especificaciones de la prueba	**122**	**100**

B. Objeto de medida de las pruebas de Matemáticas

La evaluación de las Matemáticas indaga en qué medida los alumnos desarrollan formas de pensar que les permiten formular conjeturas y procedimientos para la solución de problemas; generar explicaciones para lo relacionado con aspectos numéricos o geométricos, y fomentar el uso de distintas técnicas o recursos para hacer más eficientes los procedimientos de resolución.

Para evaluar este objeto de medida, en el caso de sexto de primaria se elaboraron 93 especificaciones a partir de 21 aprendizajes esperados (1 de cuarto grado, 9 de quinto y 11 de sexto). Para tercero de secundaria se elaboraron 100 especificaciones con base en 77 contenidos (conocimientos y habilidades) seleccionados en función de su importancia disciplinar y curricular.

En ambos niveles se elaboraron especificaciones con más de un reactivo, de manera que las pruebas incluyen un total de 150.

Las especificaciones se elaboraron con base en los tres ejes temáticos incluidos en los programas de la asignatura: Sentido Numérico y Pensamiento Algebraico; Forma, Espacio y Medida, y Manejo de la Información. A continuación, se incluyen breves descripciones de cada uno de los ejes.

Eje temático	Descripción
Sentido Numérico y Pensamiento Algebraico	Este eje alude al estudio de la aritmética y el álgebra. En primaria se abordan los conocimientos y habilidades relacionados con las propiedades de los números, las operaciones y su aplicación al resolver problemas en situaciones diversas. En secundaria se integran el estudio de los números con signo, y el desarrollo de habilidades para representar y efectuar cálculos con expresiones genéricas de los números (literales). Se trabajan el pensamiento algebraico, las ecuaciones y las generalizaciones; se desarrollan habilidades de representación como: saber describir relaciones matemáticas y usar un lenguaje verbal, gráfico o simbólico (despejar una ecuación y representar una expresión algebraica verbal o gráficamente).
Forma, Espacio y Medida	Este eje integra los tres aspectos esenciales del estudio de la geometría y la medición. En la primaria comprende la exploración de las características y propiedades de las figuras y los cuerpos geométricos, así como el conocimiento de los principios básicos de la ubicación espacial y el cálculo geométrico. En secundaria además se desarrollan habilidades para el trazo de elementos geométricos (altura, mediatrices, rotaciones, simetrías) y para resolver problemas con las propiedades de congruencia y semejanza de diversos polígonos. Además se aborda el cálculo de variables en las fórmulas de perímetro, área y volumen; la aplicación del teorema de Pitágoras, y las razones trigonométricas seno, coseno y tangente en la resolución de problemas.
Manejo de la Información	Este eje integra aspectos relacionados con el análisis de la información de distintas fuentes y su uso para la toma de decisiones informadas. En educación primaria se orienta hacia la búsqueda, la organización y el análisis de información para responder preguntas, y el uso eficiente de la herramienta aritmética en la interpretación y el análisis de los datos provenientes de diferentes contextos. En secundaria se incorporan las nociones de relaciones funcionales; proporcionalidad directa, inversa o múltiple, así como medidas de dispersión y probabilidad.

En la tabla B1 se indica la cantidad de especificaciones elaboradas por eje temático y tema.

Tabla B1. Número de especificaciones de las pruebas de Matemáticas de 6º de primaria y 3º de secundaria, por eje temático y temas

Eje temático	Temas	Número de especificaciones	
		6º primaria	3º secundaria
Sentido Numérico y Pensamiento Algebraico	Números y sistemas de numeración	18	5
	Problemas aditivos	11	5
	Problemas multiplicativos	15	14
	Patrones y ecuaciones	N/A	13
Total por eje		44	37
Forma, Espacio y Medida	Figuras y cuerpos	12	15
	Medida	17	16
	Ubicación espacial	5	N/A
Total por eje		34	31
Manejo de la Información	Proporcionalidad y funciones	8	22
	Análisis y representación de datos	7	4
	Nociones de probabilidad	N/A	6
Total por eje		15	32
Total de especificaciones de la prueba		93	100

Nota: N/A indica que ese tema no forma parte del currículo del nivel correspondiente.

Las especificaciones que se diseñaron para las pruebas de Matemáticas también fueron clasificadas de acuerdo con tres procesos cognitivos que refieren las operaciones mentales que los individuos ponen en práctica para establecer relaciones entre los objetos, las situaciones y los fenómenos. Se tiene como supuesto que estas operaciones mentales se ponen en juego cuando el alumno busca resolver un reactivo, y que su utilización es un indicador de la adquisición de los aprendizajes clave del currículo por parte de los estudiantes.

Enseguida se incluyen la categorización y las definiciones de los procesos cognitivos, mismas que fueron construidas tomando como referencia a los modelos de las pruebas del Segundo Estudio Regional Comparativo y Explicativo (SERCE) (Bronzina, Chemello, y Agrasar, 2009).

Proceso cognitivo	Definición
Reconocimiento de objetos y elementos matemáticos	Este proceso comprende el conocimiento de hechos, la retención memorística de objetos y propiedades matemáticas, ejecución de algoritmos, realización de cálculos.
Resolución de problemas simples	Este proceso comprende el uso de información matemática que está explicita en el enunciado, y el establecimiento de relaciones directas necesarias para llegar al resultado.
Resolución de problemas complejos	Este proceso comprende la reorganización de la información matemática presentada en el enunciado y la estructuración de una propuesta de solución, a partir de relaciones no explícitas.
	Este proceso comprende la reorganización de la información matemática presentada en el enunciado y la estructuración de una propuesta de solución, a partir de relaciones no explícitas.

En las tablas B2 y B3 se presenta el número de especificaciones de las pruebas de Matemáticas para sexto de primaria y tercero de secundaria, respectivamente, mostrando el número de especificaciones por proceso cognitivo y eje temático.

Tabla B2. Número de especificaciones de la prueba de Matemáticas de 6º de primaria, por proceso cognitivo y eje temático

Eje temático / Proceso cognitivo	Sentido Numérico y Pensamiento Algebraico	Forma, Espacio y Medida	Manejo de la Información	Totales por dominio cognitivo
Reconocimiento de objetos y elementos matemáticos	16	4	6	26
Resolución de problemas simples	12	20	6	38
Resolución de problemas complejos	16	10	3	29
Totales por eje	44	34	15	93

Tabla B3. Número de especificaciones de la prueba de Matemáticas de 3º de secundaria, por proceso cognitivo y eje temático

Eje temático / Proceso cognitivo	Sentido Numérico y Pensamiento Algebraico	Forma, Espacio y Medida	Manejo de la Información	Totales por dominio cognitivo
Reconocimiento de objetos y elementos matemáticos	10	9	5	24
Resolución de problemas simples	19	14	13	46
Resolución de problemas complejos	8	8	14	30
Totales por eje	37	31	32	100

C. PLANEA 2015. Sexto grado de primaria y tercer grado de secundaria. Diseño muestral

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

En este anexo se describen las características más importantes del diseño muestral del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) 2015 que fue elaborado considerando el logro educativo como la principal variable de diseño (o eje). El logro educativo, por su naturaleza, no es una variable que se mida directamente por lo que se le da el nombre de "latente".

A su vez, se asumió que la población que es objeto de estudio está determinada por los alumnos de sexto grado de primaria y los alumnos del tercer grado de secundaria que cursaban el ciclo escolar 2014-2015. Para estas subpoblaciones se definieron sendas muestras que son independientes entre sí.

Dominios de estudio (justificación)

Para la definición de los dominios de estudio, es decir, las subpoblaciones sobre las que se darán resultados de PLANEA, se exploraron variables con las cuales la población de alumnos del país se pudiera desagregar bajo la perspectiva de su potencial para comunicar resultados y que tuvieran las características técnicas necesarias para reconstruir consistentemente a futuro dichas subpoblaciones. Además, se consultaron los requerimientos y perspectivas de otras áreas del Instituto Nacional para la Evaluación de la Educación (INEE) con la intención de que dichos dominios se utilicen para otros proyectos y así complementar y contextualizar de mejor manera las evaluaciones que realiza el Instituto.

Las variables exploradas para la definición de los dominios de estudio fueron:

- Tipo de servicio, sea para primaria o secundaria.
- Sostenimiento de la escuela.
- Entidad federativa.
- Tamaño de la localidad donde se ubica la escuela, de acuerdo con el Instituto Nacional de Estadística y Geografía (INEGI).
- Índice de marginación del Consejo Nacional de Población (CONAPO).
- Escuela (primaria) multigrado: cualidad de la escuela de tener grupos de primaria —al menos uno— en donde se atiende más de un grado escolar en un mismo grupo (multigrado).

Las variables seleccionadas debido a sus características para hacer una partición en el marco muestral y a partir de ésta seleccionar la muestra son:

- Tipo de servicio, sea para primaria o secundaria.
- Sostenimiento de la escuela.
- Entidad federativa.
- Tamaño de la localidad (INEGI).

Es importante destacar que el diseño permitirá estimar resultados de los alumnos de escuelas multigrado y de los alumnos en escuelas en condiciones de marginalidad alta, media y baja. En términos de muestreo, a estos dominios se les conoce como dominios no planeados en el sentido de que el marco muestral no se segmenta por ellos.

Ejemplo: es posible dar resultados de logro de hombres y mujeres simplemente porque en la muestra hay suficiente cantidad de ellos como para hacer una estimación con una precisión aceptable. Para lograr que la muestra permita generar resultados de estos dominios no planeados se utilizó estratificación implícita, la cual permite obtener una muestra en la que las unidades muestrales están en las mismas proporciones en las que se encuentran en la población.

A continuación, se presentan los resultados de la exploración de las variables utilizadas en la definición de los dominios de estudio.

Tipo de servicio y sostenimiento

El tipo de servicio y el sostenimiento son las clasificaciones que la Secretaría de Educación Pública (SEP) ha usado tradicionalmente y que se encuentran disponibles en la estadística de la encuesta de centros escolares que cada año se realiza (Forma 911). Si bien hay otro tipo de clasificaciones de escuelas, como escuelas de tiempo completo, internados, etc., éstas no están sistemáticamente identificadas en la estadística de la SEP. En las siguientes tablas se muestran los totales de alumnos y de escuelas con base en esta clasificación.

Alumnos y escuelas de 6º de primaria por tipo de servicio y sostenimiento

Tipo de servicio	Sostenimiento							
	Alumnos de 6º				Escuelas con alumnos de 6º			
	Público		Privado		Público		Privado	
	Total	%	Total	%	Total	%	Total	%
Comunitarias	15 461	0.62	0	0.00	7 309	7.84	0	0.00
Generales	2 144 086	85.74	207 803	8.31	67 983	72.90	8 135	8.72
Educación indígena	133 193	5.33	25	0.001	9 820	10.53	2	0.002

Alumnos y escuelas de 3º de secundaria por tipo de servicio y sostenimiento

Tipo de servicio	Sostenimiento							
	Alumnos de 3º				Escuelas con alumnos de 3º			
	Público		Privado		Público		Privado	
	Total	%	Total	%	Total	%	Total	%
Comunitarias	10 111	0.52	0	0.00	2 433	6.59	0	0.00
Generales	831 549	42.72	152 434	7.83	7 145	19.34	4 220	11.42
Migrantes	139	0.01	0	0.00	24	0.06	0	0.00
Técnicas	526 353	27.04	11 034	0.57	4 319	11.69	272	0.74
Telesecundarias	405 173	20.82	611	0.03	18 243	49.39	16	0.04
Trabajadores	8 867	0.46	65	0.00	262	0.71	3	0.01

A partir de los datos de la tabla se puede decir que:

- Las primarias generales son las que concentran a la gran mayoría de alumnos (86.0%).
- Las primarias comunitarias tienen menos del 1.0% de alumnos del país.
- Las primarias privadas son casi en su totalidad generales.
- Las secundarias generales y técnicas concentran al 70.0% de alumnos y las telesecundarias 21.0%.
- Las secundarias comunitarias, de migrantes y de trabajadores atienden (cada una) a menos del 1.0% de alumnos del país.
- Las telesecundarias representan 50.0% de secundarias del país.
- Las secundarias privadas son principalmente generales.

La distribución de alumnos es diferente a la de las escuelas, por ejemplo: las primarias comunitarias concentran menos del 1.0% de alumnos del país, pero la cantidad de escuelas es casi del 8.0%.

Esto pone de manifiesto que para efectos del diseño de muestras se deben tener claros los objetivos de los estudios para poder hacer los ajustes necesarios. En principio, PLANEA deberá aportar información sobre el estado que guardan los aprendizajes de los estudiantes en el Sistema Educativo Nacional (SEN), es decir, se debe diseñar una muestra de alumnos que tienen una distribución muy particular. Si en futuras administraciones de PLANEA se requiere hacer inferencias sobre las características y condiciones de las escuelas del país que tienen otra distribución, se deberán realizar los ajustes necesarios.

Debido a que en el levantamiento de PLANEA 2015 no será incorporado otro estudio, el diseño de la muestra se llevó a cabo con base en la distribución de alumnos.

Entidad federativa

Se considera que, a partir de la muestra de PLANEA, se puedan dar resultados de cada una de las 32 entidades federativas con un nivel de precisión aceptable conforme a los estándares internaciones que más adelante se detallan. Al interior de ellas, y con un nivel de precisión menor, se podrán dar resultados de, a lo más, tres subpoblaciones.

Tamaño de la localidad donde se ubica la escuela

El tamaño de la localidad en donde se ubica la escuela se obtuvo a partir de la información del INEGI. En la exploración de la variable de tamaño de la localidad se encontró que los alumnos tanto de sexto de primaria como de tercero de secundaria se pueden ubicar en 4 grupos (1 a 499; 500 a 2 499; 2 500 a 99 999 y 100 000 o más habitantes). En las siguientes tablas se muestra su distribución. Una desagregación mayor, entre 2 500 y 99 999 habitantes conduciría a tener categorías vacías en la siguiente tabla. Las primarias y secundarias privadas son generales y se encuentran ubicadas principalmente en localidades de 100 000 habitantes o más.

Alumnos y escuelas de 6º de primaria por tamaño de la localidad

Tamaño de la localidad	Alumnos de 6º						Escuelas con alumnos de 6º					
	Comunitarias		Generales		E. Indígena		Comunitarias		Generales		E. Indígena	
	Total	%	Total	%	Total	%	Total	%	Total	%	Total	%
No identificable	392	0.02	8761	0.35	636	0.03	208	0.22	372	0.40	83	0.09
1 a 499	13 736	0.55	162 079	6.48	52 650	2.11	6 613	7.09	22 711	24.36	6 685	7.17
500 a 2 499	777	0.03	333 943	13.35	52 873	2.11	285	0.31	13 869	14.87	2 183	2.34
2 500 a 99 999	363	0.01	748 200	29.92	23 953	0.96	146	0.16	16 771	17.99	788	0.85
100 000 o más	193	0.01	1 098 906	43.95	3 106	0.12	57	0.06	22 395	24.02	83	0.09

Alumnos de 3º de secundaria por tamaño de la localidad

Tamaño de la localidad	Tipo de servicio											
	Comunitarias		Generales		Migrantes		Técnicas		Telesecundarias		Trabajadores	
	Total	%	Total	%	Total	%	Total	%	Total	%	Total	%
No identificable	100	0.01	5 003	0.26			2 325	0.12	1 170	0.06		
1 a 499	9 270	0.48	8 836	0.45	73	0.004	10 553	0.54	83 167	4.27		
500 a 2 499	534	0.03	41 918	2.15	12	0.001	48 912	2.51	209 380	10.76		
2 500 a 99 999	63	0.003	345 182	17.73	38	0.002	197 317	10.14	81 295	4.18	2 680	0.14
100 000 o más	144	0.01	583 044	29.96	16	0.001	278 280	14.30	30 772	1.58	6 252	0.32

Escuelas de 3º de secundaria por tamaño de la localidad

Tamaño de la localidad	Tipo de servicio											
	Comunitarias		Generales		Migrantes		Técnicas		Telesecundarias		Trabajadores	
	Total	%	Total	%	Total	%	Total	%	Total	%	Total	%
No identificable	25	0.07	85	0.23	0	0.0	32	0.09	52	0.14		
1 a 499	2 313	6.26	241	0.65	13	0.04	167	0.45	7 423	20.10		
500 a 2 499	72	0.19	939	2.54	3	0.01	908	2.46	8 782	23.78		
2 500 a 99 999	16	0.004	3 797	10.28	5	0.01	1 719	4.65	1 455	3.94	79	0.21
100 000 o más	7	0.02	6 303	17.06	3	0.01	1 765	4.78	547	1.48	186	0.50

A partir de los datos de las tablas previas se puede decir que:

- Los alumnos de sexto de primarias generales se reparten en los cuatro grupos (1 a 499; 500 a 2 499; 2 500 a 99 999 y 100 000 o más habitantes).
- Los alumnos de sexto de primarias comunitarias se concentran en localidades de 1 a 499 habitantes.
- Los alumnos de sexto de primarias de educación indígena se concentran en los grupos de 1 a 499 y 500 a 2 499 habitantes, es decir, en un ámbito rural.
- Los alumnos de tercero de secundarias generales y técnicas se concentran en los grupos de 2 500 a 99 999 y 100 000 o más habitantes.
- Los alumnos de tercero de telesecundaria se concentran en el grupo de 500 a 2 499 habitantes, seguido del de 1 a 499 y de 2 500 a 99 999 habitantes.

A continuación, se presentan dos tablas en donde se esquematiza la concentración de alumnos de sexto de primaria y de tercero de secundaria. Las celdas que se han unido (combinadas), indican que los alumnos se reparten en ellas en proporciones pequeñas y que su unión forma una subpoblación a partir de la cual se pueden dar resultados. Lo anterior no implica que en el diseño muestral se vayan a eliminar a los alumnos de las celdas vacías. Esta información sirvió para definir subpoblaciones sobre las cuales se van a dar resultados (marcadas con cruces).

Concentración de alumnos de 6º de primaria por tamaño de la localidad

	Tamaño de la localidad			
	1 a 499	500 a 2 499	2 500 a 99 999	100 000 o más
Generales	X	X	X	X
Educación Indígena	X			
Comunitarias	X			
Privadas				X

Concentración de alumnos de 3º de secundaria por tamaño de la localidad

	Tamaño de la localidad			
	1 a 499	500 a 2 499	2 500 a 99 999	100 000 o más
Generales		X	X	X
Técnicas		X	X	X
Telesecundarias	X	X	X	
Comunitarias	X			
Privadas				X

La variable tamaño de la localidad tiene la cualidad de que a partir de ella se pueden crear agrupaciones (subpoblaciones) que sean relativamente estables a través del tiempo. Es decir, las localidades se siguen clasificando de la misma forma.

Índice de marginación del Consejo Nacional de Población (CONAPO)

El índice de marginación del CONAPO es un índice con cinco categorías: muy alto, alto, medio, bajo y muy bajo. Para efectos del diseño muestral, estas categorías se colapsaron en tres, con la finalidad de tener grupos con suficiente población como para poder decir algo de ellos: alto, medio y bajo.

A continuación se muestra su distribución:

Índice de marginación CONAPO	Primaria				Secundaria			
	Alumnos de 6°		Escuelas		Alumnos de 3°		Escuelas	
	Total	%	Total	%	Total	%	Total	%
No identificable	35 799	1.43	2 248	2.41	40 358	2.07	762	2.06
Alto	981 424	39.25	53 492	57.36	691 352	35.52	20 581	55.72
Medio	757 411	30.29	20 251	21.72	558 494	28.69	7 929	21.47
Bajo	725 934	29.03	17 258	18.51	656 132	33.71	7 665	20.75

En general, el tamaño de la población está asociado a la marginación de las localidades: a menor tamaño, mayor marginación.

La desventaja de utilizar la variable del índice de marginación como variable de segmentación en el diseño muestral es que la definición del grado de marginación de las localidades se realiza con base en un análisis factorial de componentes principales de las variables que constituyen el índice. Este procedimiento puede provocar que, a futuro, la clasificación de las localidades pueda variar debido a que la desagregación se basa en las varianzas analizadas.

Escuelas multigrado

Una escuela es multigrado si al menos en un grupo se imparte más de un grado escolar. Debido a que en el marco muestral no se incluye información respecto a si en un grupo o más de las escuelas se dan clases a más de un grado, se delimitó la siguiente definición operacional con la información disponible:

- Se calcula la razón de cantidad de grados escolares que se imparten en la escuela y el total de docentes de la misma (cantidad de grados escolares que se imparten en la escuela / cantidad de total de docentes).
- Dicha razón se categoriza en valores enteros redondeados al entero superior. El 1 representa una escuela que no es multigrado, el 2 un nivel de multigrado bajo (por ejemplo, cinco docentes en la escuela para atender seis grados escolares), etcétera. El 6 representa el nivel más severo de multigrado pues un docente atiende seis grados escolares.

Definición de unidad primaria de muestreo

Las Unidades Primarias de Muestreo (UPM) están constituidas por la combinación de la clave del centro de trabajo y el turno.

Dominios de estudio (definición)

Dominios primarios

Para sexto de primaria

- **Nacional:** conformado por todos los alumnos de la población objetivo.
- **Entidades:** conformadas por la partición que generan las 32 entidades federativas.

Para tercero de secundaria

- **Nacional:** conformado por todos los alumnos de la población objetivo.
- **Entidades:** conformadas por la partición que generan las 32 entidades federativas.

Dominios secundarios

Conformados por los alumnos que se encuentran en la siguiente clasificación de UPM.

Para sexto grado de primaria:

- **Cursos Comunitarios** (CCO). UPM en primarias administradas por el Consejo Nacional de Fomento Educativo (CONAFE) y atendidas por instructores que no necesariamente tienen formación docente.

- **Generales públicas en localidades de 1 a 499 habitantes** (GRPT01). UPM en primarias generales públicas ubicadas en localidades de 1 a 499 habitantes.
- **Generales públicas en localidades de 500 a 2 499 habitantes** (GRPT02). UPM en primarias generales públicas ubicadas en localidades de 500 a 2 499 habitantes.
- **Generales públicas en localidades de 2 500 a 99 999 habitantes con alta marginación** (GRPT03M0A). UPM en primarias generales públicas ubicadas en localidades de 2 500 a 99 999 habitantes cuyo índice de marginación es alto o muy alto.
- **Generales públicas en localidades de 2 500 a 99 999 habitantes con media marginación** (GRPT03M0M). UPM en primarias generales públicas ubicadas en localidades de 2 500 a 99 999 habitantes cuyo índice de marginación es medio.
- **Generales públicas en localidades de 2 500 a 99 999 habitantes con baja marginación** (GRPT03M0B). UPM en primarias generales públicas ubicadas en localidades de 2 500 a 99 999 habitantes cuyo índice de marginación es bajo o muy bajo.
- **Generales públicas en localidades con 100 000 o más habitantes con alta marginación** (GRPT04M0A). UPM en primarias generales públicas ubicadas en localidades con 100 000 o más habitantes cuyo índice de marginación es alto o muy alto.
- **Generales públicas en localidades con 100 000 o más habitantes con media marginación** (GRPT04M0M). UPM en primarias generales públicas ubicadas en localidades con 100 000 o más habitantes cuyo índice de marginación es medio.
- **Generales públicas en localidades con 100 000 o más habitantes con baja marginació**n (GRPT04M0B). UPM en primarias generales públicas ubicadas en localidades con 100 000 o más habitantes cuyo índice de marginación es bajo o muy bajo.
- **Educación indígena no multigrado** (INDM0). UPM en primarias administradas por la Dirección General de Educación Indígena (DGEI) que no se consideran multigrado bajo la definición de la SEP.
- **Educación indígena multigrado** (INDM1). UPM en primarias administradas por la DGEI que se consideran multigrado bajo la definición de la SEP.
- **Privadas** (PRV). UPM en primarias generales públicas e indígenas de sostenimiento privado.

Para tercer grado de secundaria

- **Cursos Comunitarios** (CCO). UPM en secundarias administradas por CONAFE y atendidas por instructores que no necesariamente tienen formación docente.
- **Generales públicas en localidades de 1 a 2 499 habitantes** (GRPT12). UPM en secundarias generales públicas ubicadas en localidades de 1 a 2 499 habitantes.
- **Generales públicas en localidades de 2 500 a 99 999 habitantes** (GRPT03). UPM en secundarias generales públicas ubicadas en localidades de 2 500 a 99 999 habitantes.
- **Generales públicas en localidades con 100 000 o más habitantes con alta y media marginación** (GRPT04MAM). UPM en secundarias generales públicas ubicadas en localidades con 100 000 o más habitantes cuyo índice de marginación es medio, alto o muy alto.
- **Generales públicas en localidades con 100 000 o más habitantes con baja marginación** (GRPT04M0B). UPM en secundarias generales públicas ubicadas en localidades con 100 000 o más habitantes cuyo índice de marginación es bajo o muy bajo.
- **Técnicas en localidades de 1 a 2 499 habitantes** (TECT12). UPM en secundarias técnicas ubicadas en localidades de 1 a 2 499 habitantes.

- **Técnicas en localidades de 2 500 a 99 999 habitantes** (TECT03). UPM en secundarias técnicas ubicadas en localidades de 2 500 a 99 999 habitantes.
- **Técnicas en localidades con 100 000 o más habitantes** (TECT04). UPM en secundarias técnicas ubicadas en localidades con 100 000 o más habitantes.
- **Telesecundarias en localidades de 1 a 499 habitantes** (TELT01). UPM en telesecundarias ubicadas en localidades de 1 a 499 habitantes.
- **Telesecundarias en localidades de 500 a 2 499 habitantes** (TELT02). UPM en telesecundarias ubicadas en localidades de 500 a 2 499 habitantes.
- **Telesecundarias en localidades con 2 500 o más habitantes** (TELT34). UPM en telesecundarias ubicadas en localidades con 2500 o más habitantes.
- **Privadas** (PRV). UPM en primarias generales públicas e indígenas de sostenimiento privado.

Dominios no planeados

- **Nacional:** escuelas multigrado y no multigrado.
- **Nacional:** escuelas en localidades de alta, media y baja marginación.
- **Primarias generales públicas.** Escuelas en localidades de alta, media y baja marginación.
- **Primarias generales públicas.** Escuelas multigrado y no multigrado.
- **Primarias de educación indígena.** Escuelas en localidades de alta, media y baja marginación.
- **Secundarias generales.** Escuelas en localidades de alta, media y baja marginación.
- **Secundarias técnicas.** Escuelas en localidades de alta, media y baja marginación.
- **Telesecundarias.** Escuelas en localidades de alta, media y baja marginación.

Cobertura y exclusiones

Exclusiones de sexto de primaria:

- Se excluyeron las UPM que no contaban con algún alumno registrado en sexto grado.
- Se excluyeron las UPM que tenían como valor perdido el total de alumnos en sexto grado.
- Se excluyeron las UPM en escuelas generales públicas en las que no fue posible identificar el tamaño de la localidad.
- Se excluyeron las UPM en escuelas generales públicas en localidades de 2 500 a 99 999 habitantes en las que no fue posible identificar el grado de marginación.
- Se excluyeron las UPM en escuelas generales públicas en localidades con 100 000 o más habitantes en las que no fue posible identificar el grado de marginación.
- Se excluyeron los estratos en los que no se pudo asignar al menos una UPM en muestra.
- Se excluyen las UPM de la entidad de Oaxaca que no son escuelas comunitarias debido a que la autoridad educativa estatal manifestó que se reúnen las condiciones necesarias para llevar a cabo el levantamiento de datos.

Exclusiones de tercero de secundaria:

- Se excluyeron las UPM que no contaban con algún alumno registrado en tercer grado.
- Se excluyeron las UPM que tenían como valor perdido el total de alumnos en tercer grado.
- Se excluyeron las UPM pertenecientes a escuelas de trabajadores y de migrantes.
- Se excluyeron las UPM en escuelas generales públicas, técnicas públicas y telesecundarias públicas en las que no fue posible identificar el tamaño de la localidad.
- Se excluyeron las UPM en escuelas generales públicas en localidades con 100 000 o más habitantes en las que no fue posible identificar el grado de marginación.
- Se excluyen las UPM de la entidad de Oaxaca que no son escuelas comunitarias debido a que la autoridad educativa estatal manifestó que se reúnen las condiciones necesarias para llevar a cabo el levantamiento de datos.

Estratificación

Explícita

La estratificación explícita consiste en formar una partición de la población en la cual se seleccionan muestras independientes entre cada uno de los grupos resultantes (llamados estratos). La estratificación puede ser aplicada tanto a nivel de UPM como de alumnos. En este caso, la estratificación fue construida para las UPM.

La estratificación explícita tanto para sexto de primaria como para tercero de secundaria se define de la misma manera. Simplemente se tomó el producto cartesiano de los conjuntos que definen los dominios primarios y los secundarios con lo que en ambos casos se forman 384 estratos. Si se desea dar información de las subpoblaciones que conforman cada estrato es indispensable estimar los errores estándar asociados para valorar su utilidad.

Formación de los estratos explícitos

	Dominios secundarios
Entidades	Estratos

Implícita

La estratificación implícita se puede anidar dentro de la estratificación explícita y consiste en ordenar las unidades de la población de acuerdo con algunas variables de interés antes de seleccionar una muestra. Combinando este ordenamiento junto con un muestreo sistemático de unidades se puede obtener una muestra en la que las unidades están en las mismas proporciones en que se encuentran en la población y se pueden obtener estimaciones más confiables.

En este caso, la estratificación implícita se le aplica a las UPM dentro de los estratos explícitos, las variables de ordenamiento fueron las siguientes:

En sexto grado de primaria:

- Tamaño de localidad.
- Grado de marginación.
- Grado de multigrado.

En tercer grado de secundaria:

- Tamaño de localidad.
- Grado de marginación.

Medida de tamaño de las UPM

La medida de tamaño es la cantidad de alumnos de sexto grado que estén registrados en el turno al que corresponde la UPM.

Precisión de las muestras

La precisión que emplean estudios internacionales en el diseño de sus muestras como TIMSS (Trends in International Mathematics and Science Study), Civic Education Study de la IEA (Asociación Internacional de Evaluación) y PISA (Programme for International Student Assessment) se resume en el siguiente cuadro:

La precisión estándar requerida para dar resultados de cualquier población consta de un **tamaño de muestra efectiva** de 400 alumnos para construir intervalos de confianza del 95% al estimar medias poblacionales, porcentajes y coeficientes de correlación.

- **Medias:** $M \pm 0.1S$ (donde M es la estimación de la media y S la desviación estándar estimada).
- **Porcentajes:** $p \pm 5\%$ (donde p es el porcentaje estimado).
- **Correlaciones:** $r \pm 0.1$ (donde r es la correlación estimada).

Puede revisarse en (Mullis, Martin, González y Chrostowski, 2004, p. 114), (Schulz y Sibberns, 2004, p. 45) y (Murphy y Schulz, 2006, p. 7).

Para el diseño muestral de PLANEA, los dominios de estudio que cumplen con esta precisión son los siguientes:

- Nacional.
- Dominios primarios.
- Dominios secundarios.
- Dominios no planeados.

Cualquier estimación hecha en otras subpoblaciones diferentes a las mencionadas tendrá que valorarse con base en las estimaciones del error estándar y la desviación estándar. Debe considerarse que las subpoblaciones que se obtienen de combinar los dominios arriba citados tendrán precisiones menores respecto a los estándares previamente establecidos debido a que el tamaño de muestra disminuye, por lo que se recomienda que las estimaciones que se obtengan se utilicen con cautela.

Debe tenerse en cuenta que el tamaño de la precisión que aquí se ha presentado únicamente considera al error debido al muestreo. No se toma en cuenta el error de medida de los instrumentos o el error de cualquier otra fuente.

Método de selección de las muestras

La selección de los alumnos se llevó a cabo en dos etapas. Antes de llevar a cabo la selección de las UPM, éstas se organizaron mediante el ordenamiento serpentina —*Serpentine sort*— (2014) dentro de los estratos explícitos e implícitos. Este método ordena alternadamente de manera ascendente y descendente las UPM de tal forma que cualesquiera dos registros consecutivos en el archivo ordenado son más similares con respecto a los valores de las variables de clasificación que en la clasificación tradicional. Esta técnica reduce las estimaciones de la varianza cuando se utilizan métodos de replicación.

La primera etapa de muestreo consistió en la selección de UPM mediante un muestreo sistemático en el que las unidades tienen una probabilidad de selección proporcional a su tamaño, como se describió anteriormente.

La segunda etapa de muestreo consistió en hacer una selección aleatoria simple de alumnos tomando una cuota dependiendo del tamaño de cada UPM proveniente de la primera etapa.

Modificación del diseño muestral

Inicialmente, el diseño muestral se realizó bajo el supuesto de que en cada escuela seleccionada se censarían a los alumnos para homologar la logística de aplicación de la Evaluación del Logro referida a los Centros Escolares (ELCE) a cargo de la SEP. El objetivo de la ésta era evaluar a todos y cada uno de los alumnos e informarles sus resultados.

El INEE realizó la selección de las escuelas de la muestra e inició el proceso de validación de ésta.

Posteriormente, la SEP modificó la logística de aplicación de ELCE pasando de ser censal para los alumnos dentro de las escuelas a muestral, debido a la falta de recursos económicos. En respuesta, el INEE decidió modificar el diseño muestral pues ya no era necesario llevar a cabo un operativo, que desde el punto de vista técnico era ineficiente porque con menos alumnos se obtendrían las mismas precisiones.

Por lo anterior, los apartados *Tamaño de la muestra (alumnos) y Asignación de la muestra y tamaño de la muestra* (UPM) que se presentan a continuación corresponden a la forma en que fueron seleccionadas las escuelas bajo el supuesto de que se censarían los alumnos en cada una de las escuelas seleccionadas. En el apartado *Tamaño de las muestras al interior de las* UPM *seleccionadas*, se hace la modificación correspondiente.

Tamaño de la muestra (alumnos)

PASO 1: tamaño de muestra necesario bajo el supuesto de muestreo aleatorio simple.

Como se mencionó en la sección anterior, la precisión para estimar promedios, porcentajes y correlaciones se tomó como un valor fijo. Dicha precisión equivale a tomar el error estándar aproximadamente a 0.05 desviaciones estándar en cualquiera de los tres tipos de estimadores.

Es fácil probar que considerando muestreo aleatorio simple y eliminando la corrección por finitud se necesitan 400 alumnos para obtener dicha precisión.

En adelante denotaremos con n_d a la cantidad de alumnos necesaria para alcanzar la precisión estándar internacional considerando muestreo aleatorio simple. A dicho número se le conoce como el tamaño de muestra efectiva.

PASO 2: ajuste por el efecto de diseño ($deff_d$)

Para cada uno de los dominios primarios y secundarios se calculó el tamaño de muestra necesario (n_d) considerando el efecto de diseño ($deff_d$) y el tamaño de muestra efectiva:

$$n_d = n_0 deff_d$$

El efecto de diseño estimado de cada dominio (d) se aproximó de la siguiente forma:

$$deff_d = 1 + \left\{ \left(1 - TNR_{id}\right) \overline{n}_{id} - 1 \right\} \rho_d$$

Donde:

TNR_{id} es la tasa de no respuesta al interior de las UPM del dominio d,
$\overline{n}_{id}$ es el tamaño promedio de los conglomerados dentro en el dominio d,
ρ_d es el coeficiente de correlación intra-conglomerado del dominio d.

$$n_{final_d} = \frac{n_d}{\left(1 - TNR_{I_d}\right)\left(1 - TNR_{i_d}\right)}$$

Donde:

n_{final_d} es el tamaño de muestra de alumnos, necesario considerando el efecto de diseño y la falta de respuesta,
TNR_{I_d} es la tasa de no respuesta de las UPM en dominio d,

Asignación de la muestra y tamaño de la muestra (UPM)

Una vez definidos los dominios de estudio y fijadas las cantidades de alumnos requeridas para cada uno de ellos, se procede a calcular y alojar los tamaños definitivos de muestra de alumnos y de UPM.

Para asignar la muestra en los estratos explícitos el procedimiento fue el siguiente:

1. Se distribuye la cuota de alumnos de cada uno de los dominios primarios (las entidades) en los estratos explícitos (que corresponden a cada entidad) de forma proporcional a la cantidad de alumnos en cada estrato. Este procedimiento asegura que en casi todos los dominios secundarios se tenga el tamaño de muestra necesario excepto en CCO, GRPT03M0B, GRPT04M0A y INDMU1 para sexto de primaria y GRPT12 y TEC12 para tercero de secundaria.
2. En los dominios secundarios mencionados en el punto anterior se distribuye la cuota de alumnos de los dominios secundarios de forma proporcional a la cantidad de alumnos en cada estrato.
3. A continuación, se toma en los estratos al máximo de los tamaños de muestra que se obtienen de distribuir proporcionalmente en los dominios primarios contra el que se distribuye proporcionalmente en los secundarios.
4. La asignación de muestra final por estrato (n_h) queda determinada como el valor máximo en cada estrato resultante del proceso descrito en el punto anterior.
5. En los estratos cuya cantidad de UPM en muestra fuera menor a cinco se incrementó la cuota a cinco con el objeto de poder calcular la varianza de los estimadores a pesar de que se puedan perder UPM.

Para calcular la cantidad necesaria de UPM por estrato (n_{Ih}) se utilizó lo siguiente:

$$n_{Ih} = \frac{n_h}{\overline{n}_{ih}}$$

Donde:

$\overline{n}_{ih}$ es el tamaño promedio del conglomerado en el estrato h.

Resultando de este procedimiento la cantidad final de n_I 3 995 UPM tanto para primaria como para secundaria. Lo cual nos situó por debajo del límite presupuestal que era de 4 000 UPM para sexto de primaria y 4 000 para tercer grado de secundaria.

Tamaño de las muestras al interior de las UPM seleccionadas

Para definir el tamaño de la muestra dentro de las UPM seleccionadas en las muestras, se consideró fijar una precisión con la que fuese posible dar resultados de los promedios de cada una de las UPM en muestra. El criterio consistió en fijar en 0.1 el cociente entre el error estándar de la media y la desviación estándar de la variable eje. Con esta precisión se reducen los costos, se hace posible la viabilidad del levantamiento y se conserva la precisión fijada para los dominios de estudio.

Una vez fijada la precisión para cada UPM i en las muestras se consideró que al interior el muestreo fue aleatorio simple en una población finita. Por lo que es posible deducir el tamaño de muestra.

$$n_i = \frac{1}{\left(\frac{\sqrt{V(Y_i)}}{S_i}\right)^2 + \frac{1}{N_i}}$$

Donde:

$\sqrt{V(Y_i)}$ es el error estándar de la media de la i-esima UPM seleccionada,
S_i es la desviación estándar de la variable y de la i-esima UPM seleccionada,
n_i es el tamaño de muestra de alumnos de la i-esima UPM seleccionada,
N_i es el total de alumnos de la población objetivo en la i-esima UPM seleccionada.

A partir de esta ecuación es posible calcular las cuotas que fueron empleadas para cada UPM:

- Si el tamaño de la UPM seleccionada es menor o igual a 53, la cuota de alumnos se define como el mínimo entre 35 y el tamaño, con la finalidad de que se requiera como máximo un aplicador para muestrear toda la UPM.
- Si el tamaño de la UPM seleccionada está entre 54 y 231, la cuota de alumnos se define como el mínimo entre 70 y el tamaño, con la finalidad de que se requieran como máximo dos aplicadores para muestrear toda la UPM.
- Si el tamaño de la UPM es mayor a 231, la cuota de alumnos es de 105, con la finalidad de que requieran como máximo 3 aplicadores para muestrear toda la UPM.

De esta forma, la cantidad de aplicadores necesarios para recabar la información de ambas muestras se reduce considerablemente con respecto al diseño muestral anterior.

Ninguna de estas cuotas considera una sobre muestra para reducir la pérdida de información debida a la no respuesta al interior de las UPM seleccionadas. Esto se debe a los altos costos que representa llevar material extra.

Después de todas las modificaciones, pérdidas y exclusiones, a continuación se presentan las distribuciones de ambas muestras, tanto de alumnos como de UPM.

Tablas del tamaño y distribución de las muestras

Para sexto de primaria

- **Total de UPM: 3 734**
- **Total de alumnos: 117 060**

Para tercero de secundaria

- **Total de UPM: 3 803**
- **Total de alumnos: 167 455**

UPM en la muestra de sexto grado de primaria

Entidad	Cursos Comunitarios	Generales Públicas en localidades de 1 a 499 habitantes	Generales Públicas en localidades de 500 a 2499 habitantes	Generales Públicas en localidades de 2500 a 99999 habitantes con Alta Marginación	Generales Públicas en localidades de 2500 a 99999 habitantes con Media Marginación	Generales Públicas en localidades de 2500 a 99999 habitantes con Baja Marginación
Aguascalientes	5	13	17	5	7	5
Baja California	4	7	8	5	8	5
Baja California Sur	4	17	8	5	9	10
Campeche	6	30	16	12	5	0
Coahuila	3	23	11	5	13	9
Colima	4	19	8	10	8	0
Chiapas	62	23	12	8	5	5
Chihuahua	7	22	8	5	5	5
Distrito Federal*	0	0	0	5	5	0
Durango	6	43	14	5	5	5
Guanajuato	14	33	24	7	7	5
Guerrero	28	38	23	16	5	0
Hidalgo	13	29	23	6	7	5
Jalisco	7	32	11	5	16	10
México	9	13	16	11	6	10
Michoacán	5	47	21	15	9	5
Morelos	3	6	14	24	10	5
Nayarit	9	23	22	5	12	5
Nuevo León	2	25	5	5	7	10
Oaxaca	26	0	0	0	0	0
Puebla	15	19	20	19	5	5
Querétaro	10	23	19	8	5	5
Quintana Roo	4	10	12	5	5	5
San Luis Potosí	19	35	17	5	5	5
Sinaloa	8	34	19	5	8	5
Sonora	3	20	12	5	8	7
Tabasco	6	28	34	9	7	5
Tamaulipas	5	32	8	5	6	5
Tlaxcala	6	9	19	28	10	5
Veracruz	27	37	20	12	6	6
Yucatán	4	11	15	24	5	5
Zacatecas	0	55	26	5	11	5
NACIONAL	324	756	482	289	230	162

* A partir de 2016, cambió su denominación por Ciudad de México.

	Generales Públicas en localidades con más de 100 000 habitantes con Alta Marginación	Generales Públicas en localidades con más de 100 000 habitantes con Media Marginación	Generales Públicas en localidades con más de 100 000 habitantes con Baja Marginación	Educación Indígena No Multigrado	Educación Indígena Multigrado	Privadas	TOTAL
	5	10	17	0	0	11	**95**
	5	18	25	5	5	17	**112**
	3	5	11	0	0	12	**84**
	5	6	5	0	6	7	**98**
	5	17	28	0	0	14	**128**
	5	12	21	0	0	8	**95**
	5	5	5	23	33	5	**191**
	5	14	21	5	14	9	**120**
	5	26	31	0	0	35	**107**
	5	6	10	5	9	5	**118**
	5	6	4	0	0	9	**114**
	9	5	5	20	18	4	**171**
	5	5	5	7	15	8	**128**
	6	9	12	5	5	11	**129**
	31	12	11	5	5	14	**143**
	5	5	5	6	4	9	**136**
	5	9	7	0	4	25	**112**
	0	5	8	5	14	6	**114**
	5	13	32	0	0	12	**116**
	0	0	0	0	0	0	**26**
	5	5	5	8	14	12	**132**
	5	5	9	5	5	13	**112**
	9	14	8	5	7	13	**97**
	5	5	8	5	10	7	**126**
	5	6	11	0	4	7	**112**
	5	10	19	5	5	11	**110**
	0	5	5	5	5	5	**114**
	5	17	17	0	0	12	**112**
	0	0	0	5	0	14	**96**
	7	5	5	12	10	5	**152**
	5	6	11	5	10	11	**112**
	5	5	5	0	0	5	**122**
	175	271	366	141	101	336	**3 734**

Alumnos en la muestra de sexto grado de primaria

Entidad	Cursos Comunitarios	Generales Públicas en localidades de 1 a 499 habitantes	Generales Públicas en localidades de 500 a 2 499 habitantes	Generales Públicas en localidades de 2 500 a 99 999 habitantes con Alta Marginación	Generales Públicas en localidades de 2 500 a 99 999 habitantes con Media Marginación	Generales Públicas en localidades de 2 500 a 99 999 habitantes con Baja Marginación
Aguascalientes	20	137	546	305	403	280
Baja California	27	166	190	257	385	249
Baja California Sur	9	149	266	335	418	535
Campeche	16	293	356	463	196	0
Coahuila	12	191	303	182	548	420
Colima	6	182	197	304	400	0
Chiapas	189	297	319	401	265	226
Chihuahua	13	216	198	180	186	241
Distrito Federal*	0	0	0	315	312	0
Durango	12	297	346	197	233	205
Guanajuato	43	385	723	334	443	310
Guerrero	70	296	596	664	215	0
Hidalgo	29	368	710	212	375	311
Jalisco	17	320	286	265	729	542
México	30	304	448	643	350	618
Michoacán	8	419	502	631	389	255
Morelos	5	131	414	1 111	527	262
Nayarit	14	157	488	211	485	230
Nuevo León	4	244	173	169	353	537
Oaxaca	94	0	0	0	0	0
Puebla	46	377	623	1 122	325	315
Querétaro	25	277	553	376	272	350
Quintana Roo	15	149	284	267	305	271
San Luis Potosí	44	355	453	168	215	260
Sinaloa	23	308	413	178	365	275
Sonora	4	214	238	171	371	262
Tabasco	15	223	942	411	368	269
Tamaulipas	23	232	242	139	241	240
Tlaxcala	20	162	655	1 435	586	262
Veracruz	87	398	470	533	246	207
Yucatán	9	100	456	1 151	284	291
Zacatecas	0	491	601	157	558	280
NACIONAL	929	7 788	12 991	13 287	11 348	8 503

* A partir de 2016, cambió su denominación por Ciudad de México.

	Generales Públicas en localidades con más de 100 000 habitantes con Alta Marginación	Generales Públicas en localidades con más de 100 000 habitantes con Media Marginación	Generales Públicas en localidades con más de 100 000 habitantes con Baja Marginación	Educación Indígena No Multigrado	Educación Indígena Multigrado	Privadas	TOTAL
	297	653	963	0	0	397	**4 001**
	225	883	1 329	141	57	489	**4 398**
	126	259	454	0	0	292	**2 843**
	317	298	277	0	83	267	**2 566**
	171	886	1 541	0	0	642	**4 896**
	162	552	1 097	0	0	295	**3 195**
	259	303	265	930	392	170	**4 016**
	230	710	1 086	137	114	267	**3 578**
	304	1 513	1 820	0	0	1 480	**5 744**
	226	240	477	112	42	201	**2 588**
	315	381	245	0	0	367	**3 546**
	447	299	294	610	247	91	**3 829**
	291	295	332	231	115	194	**3 463**
	347	452	666	173	48	387	**4 232**
	1 866	686	706	201	48	536	**6 436**
	245	176	262	169	22	300	**3 378**
	199	447	374	0	19	553	**4 042**
	0	224	481	145	115	247	**2 797**
	308	657	1 702	0	0	545	**4 692**
	0	0	0	0	0	0	**94**
	307	297	315	395	166	414	**4 702**
	275	305	587	136	60	657	**3 823**
	535	794	408	98	62	483	**3 671**
	259	332	491	132	101	264	**3 074**
	237	321	635	0	49	311	**3 115**
	175	441	1 000	187	39	498	**3 600**
	0	331	307	146	66	165	**3 243**
	230	936	881	0	0	314	**3 478**
	0	0	0	230	0	383	**3 733**
	263	162	202	412	79	173	**3 232**
	260	290	541	144	76	384	**3 986**
	299	291	248	0	0	144	**3 069**
	9 175	14 414	19 986	4 729	2 000	11 910	**117 060**

UPM en muestra de tercer grado de secundaria

Entidad	Cursos Comunitarios	Generales Públicas en localidades de 1 a 2499 habitantes	Generales Públicas en localidades de 2500 a 99999 habitantes	Generales Públicas en localidades de más de 100000 habitantes con Alta y Media Marginación	Generales Públicas con más de 100000 habitantes con Baja Marginación	Técnicas en localidades de 1 a 2499 habitantes	
Aguascalientes	5	0	5	5	9	5	
Baja California	0	5	13	11	19	5	
Baja California Sur	0	5	9	3	5	5	
Campeche	16	0	5	5	5	5	
Coahuila	5	5	7	5	12	6	
Colima	4	0	5	5	10	5	
Chiapas	17	5	8	5	5	7	
Chihuahua	13	7	6	5	9	5	
Distrito Federal*	0	0	5	20	29	0	
Durango	10	5	5	5	5	5	
Guanajuato	5	5	8	5	5	5	
Guerrero	23	5	7	5	5	18	
Hidalgo	14	5	9	5	5	5	
Jalisco	12	5	16	6	8	5	
México	7	28	19	16	9	8	
Michoacán	11	5	11	4	5	6	
Morelos	6	4	12	5	5	4	
Nayarit	8	5	8	5	5	7	
Nuevo León	6	6	14	8	20	5	
Oaxaca	11	0	0	0	0	0	
Puebla	7	5	10	5	5	5	
Querétaro	9	5	5	5	5	5	
Quintana Roo	13	0	5	6	5	3	
San Luis Potosí	8	5	5	5	5	5	
Sinaloa	12	13	11	5	6	5	
Sonora	6	5	7	5	7	5	
Tabasco	9	5	10	0	5	5	
Tamaulipas	7	5	5	6	9	5	
Tlaxcala	7	4	16	0	0	5	
Veracruz	7	5	6	5	5	6	
Yucatán	13	5	15	4	6	5	
Zacatecas	5	5	4	4	4	5	
NACIONAL	276	162	271	178	237	170	

* A partir de 2016, cambió su denominación por Ciudad de México.

	Técnicas en localidades de 2 500 a 99 999 habitantes	Técnicas en localidades de con más de 100 000 habitantes	Telesecundarias en localidades de 1 a 499 habitantes	Telesecundarias en localidades de 500 a 2 499 habitantes	Telesecundarias en localidades con más de 2 500 habitantes	Privadas	TOTAL
	5	8	13	28	5	14	**102**
	5	9	7	5	5	26	**110**
	7	5	15	7	5	16	**82**
	5	5	22	24	3	12	**107**
	9	13	15	5	4	26	**112**
	7	5	12	12	3	10	**78**
	5	5	24	54	6	4	**145**
	5	11	32	9	5	13	**120**
	5	18	0	0	5	49	**131**
	5	5	53	19	5	6	**128**
	5	5	21	45	13	18	**140**
	7	5	28	35	5	5	**148**
	5	5	29	43	7	12	**144**
	10	7	21	17	5	19	**131**
	6	7	8	15	11	17	**151**
	8	5	30	35	5	13	**138**
	8	5	6	15	9	33	**112**
	6	5	29	27	5	7	**117**
	5	16	8	5	0	24	**117**
	0	0	0	0	0	0	**11**
	6	5	19	45	16	15	**143**
	5	5	15	27	7	22	**115**
	5	8	11	26	5	17	**104**
	5	5	58	31	5	8	**145**
	5	5	26	14	0	12	**114**
	7	10	25	17	5	16	**115**
	6	5	15	44	7	10	**121**
	5	10	30	13	5	20	**120**
	17	0	5	20	14	13	**101**
	5	5	31	48	13	10	**146**
	7	5	10	22	5	16	**113**
	5	5	64	31	5	5	**142**
	196	212	682	738	193	488	**3 803**

Alumnos en muestra de tercer grado de secundaria

Entidad	Cursos Comunitarios	Generales Públicas en localidades de 1 a 2 499 habitantes	Generales Públicas en localidades de 2 500 a 99 999 habitantes	Generales Públicas en localidades de más de 100 000 habitantes con Alta y Media Marginación	Generales Públicas con más de 100 000 habitantes con Baja Marginación	Técnicas en localidades de 1 a 2 499 habitantes
Aguascalientes	20	0	385	315	686	294
Baja California	0	280	945	770	1 365	279
Baja California Sur	0	198	665	210	350	249
Campeche	71	0	309	420	420	265
Coahuila	18	248	560	350	945	296
Colima	7	0	350	315	700	174
Chiapas	223	302	629	420	455	439
Chihuahua	37	355	415	420	735	246
Distrito Federal*	0	0	420	1 610	2 233	0
Durango	23	303	385	350	420	267
Guanajuato	24	306	665	455	455	350
Guerrero	147	209	525	385	455	803
Hidalgo	90	306	627	442	383	315
Jalisco	65	256	1 190	490	630	267
México	16	1 458	1 356	1 146	700	489
Michoacán	51	245	831	350	385	344
Morelos	24	245	813	420	420	280
Nayarit	29	166	509	273	308	209
Nuevo León	30	267	961	559	1 505	264
Oaxaca	52	0	0	0	0	0
Puebla	32	324	770	420	344	200
Querétaro	74	336	385	455	455	315
Quintana Roo	62	0	350	490	385	123
San Luis Potosí	36	315	383	420	385	294
Sinaloa	41	472	720	385	490	245
Sonora	16	164	522	350	595	267
Tabasco	73	297	734	0	315	315
Tamaulipas	39	221	385	490	840	272
Tlaxcala	28	229	1 120	0	0	315
Veracruz	37	305	490	385	420	372
Yucatán	63	160	1 108	385	439	203
Zacatecas	14	231	315	274	385	241
NACIONAL	1 442	8 198	19 822	13 754	18 603	8 992

* A partir de 2016, cambió su denominación por Ciudad de México.

	Técnicas en localidades de 2 500 a 99 999 habitantes	Técnicas en localidades de con más de 100 000 habitantes	Telesecundarias en localidades de 1 a 499 habitantes	Telesecundarias en localidades de 500 a 2 499 habitantes	Telesecundarias en localidades con más de 2 500 habitantes	Privadas	TOTAL
	344	595	246	905	278	651	**4 719**
	350	700	165	222	301	1 099	**6 476**
	630	420	116	199	263	457	**3 757**
	350	350	293	568	165	644	**3 855**
	735	1 050	187	86	128	1 330	**5 933**
	475	350	122	257	103	329	**3 182**
	315	420	626	1 919	404	181	**6 333**
	315	875	414	242	194	597	**4 845**
	455	1 645	0	0	288	2 658	**9 309**
	350	385	497	414	200	237	**3 831**
	420	490	488	1 636	800	870	**6 959**
	506	376	370	783	232	191	**4 892**
	420	420	461	1 420	407	439	**5 730**
	757	556	363	574	333	1 089	**6 570**
	455	595	159	460	590	938	**8 362**
	581	350	565	1 009	278	700	**5 689**
	554	420	134	499	520	1 176	**5 505**
	385	345	286	548	105	288	**3 451**
	374	1 260	85	97	0	1 370	**6 772**
	0	0	0	0	0	0	**52**
	442	420	325	1 189	968	681	**6 115**
	385	420	422	984	419	1 136	**5 786**
	385	595	131	691	305	752	**4 269**
	350	379	769	753	207	440	**4 731**
	350	420	333	295	0	582	**4 333**
	512	770	228	476	295	784	**4 979**
	525	420	375	2 012	455	493	**6 014**
	420	910	394	427	315	870	**5 583**
	1 214	0	38	542	731	463	**4 680**
	385	411	551	1 426	799	401	**5 983**
	480	350	149	551	191	705	**4 784**
	350	385	499	761	278	153	**3 886**
	14 569	17 082	9 791	21 945	10 552	22 705	**167 455**

Estimación de la precisión de las muestras

La pérdida en la cantidad de alumnos debido a la modificación del diseño muestral y las UPM no afecta considerablemente la precisión de los resultados. A continuación, se presenta una estimación gruesa que considera el efecto de diseño, las tasas de no respuesta y los tamaños de muestra.

1. Primero se calculó el tamaño de muestra efectiva.

$$n_{0d} = \frac{n_d\left(1-TNR_{Id}\right)\left(1-TNR_{id}\right)}{1+\left(\left(1-TNR_{id}\right)\overline{b}_d-1\right)\rho_d}$$

En donde:

n_{0d} es el tamaño de muestra efectiva de alumnos en el dominio d,
n_d es el tamaño de muestra obtenido para el dominio d,
TNR_{Id} es la tasa de no respuesta de las UPM en dominio d,
TNR_{Id} es la tasa de no respuesta al interior de las UPM del dominio d,
$\overline{b}_d$ es el tamaño promedio de las UPM en el dominio d,
ρ_d es el coeficiente de correlación intra-conglomerado en el dominio d.

2. El tamaño de muestra efectiva se utilizó para estimar en cada dominio el cociente entre el error estándar y la desviación estándar estimadas. Considerando que se estima un promedio, la fórmula correspondiente es la siguiente:

$$\frac{\sqrt{V(\overline{Y}_d)}}{S_d} = \sqrt{\frac{1}{n_{0d}}-\frac{1}{N_d}}$$

En donde:

$V(\overline{Y}_d)$ es la varianza asociada a la estimación de la media del dominio d,
S_d es la desviación estándar de la variable eje en el dominio d,
N_d es el tamaño de la población de alumnos del dominio d.

A continuación, se presentan las precisiones estimadas para los dominios de estudio planeados que son los primarios y los secundarios considerando los tamaños de muestra que se utilizaron primero en la planeación, después en la selección y por último en la modificación de las muestras.

Precisión esperada en dominios primarios muestra de sexto grado de primaria

Dominios Primarios Sexto grado de primaria	Planeada	Seleccionada	Modificada
	EE/DS	EE/DS	EE/DS
Aguascalientes	0.0512	0.0505	0.0508
Baja California	0.0473	0.0467	0.0473
Baja California Sur	0.0544	0.0537	0.0548
Campeche	0.0514	0.0504	0.0513
Coahuila	0.0485	0.0479	0.0489
Colima	0.0528	0.0519	0.0525
Chiapas	0.0374	0.0356	0.0373
Chihuahua	0.0474	0.0464	0.0475
Distrito Federal*	0.0480	0.0474	0.0477
Durango	0.0470	0.0453	0.0476
Guanajuato	0.0470	0.0464	0.0475
Guerrero	0.0460	0.0446	0.0457
Hidalgo	0.0457	0.0440	0.0451
Jalisco	0.0452	0.0445	0.0454
México	0.0413	0.0408	0.0418
Michoacán	0.0463	0.0446	0.0464
Morelos	0.0474	0.0467	0.0478
Nayarit	0.0486	0.0474	0.0482
Nuevo León	0.0471	0.0463	0.0477
Oaxaca	0.0430	0.0407	
Puebla	0.0451	0.0442	0.0452
Querétaro	0.0473	0.0463	0.0473
Quintana Roo	0.0503	0.0496	0.0506
San Luis Potosí	0.0453	0.0437	0.0455
Sinaloa	0.0469	0.0458	0.0483
Sonora	0.0479	0.0471	0.0480
Tabasco	0.0480	0.0469	0.0476
Tamaulipas	0.0481	0.0472	0.0478
Tlaxcala	0.0511	0.0503	0.0512
Veracruz	0.0426	0.0414	0.0431
Yucatán	0.0474	0.0467	0.0472
Zacatecas	0.0478	0.0465	0.0470

* A partir de 2016, cambió su denominación por Ciudad de México.

Precisión esperada en dominios secundarios muestra de sexto grado de primaria

Dominios Primarios Sexto grado de primaria	Planeada	Seleccionada	Modificada
	EE/DS	EE/DS	EE/DS
Cursos Comunitarios	0.0461	0.0394	0.0462
Generales Públicas en localidades de 1 a 499 habitantes	0.0213	0.0197	0.0206
Generales Públicas en localidades de 500 a 2 499 habitantes	0.0234	0.0231	0.0238
Generales Públicas en localidades de 2500 a 99999 habitantes con Alta Marginación	0.0295	0.0292	0.0300
Generales Públicas en localidades de 2500 a 99999 habitantes con Media Marginación	0.0332	0.0329	0.0335
Generales Públicas en localidades de 2500 a 99999 habitantes con Baja Marginación	0.0396	0.0389	0.0398
Generales Públicas en localidades con más de 100000 habitantes con Alta Marginación	0.0382	0.0375	0.0383
Generales Públicas en localidades con más de 100000 habitantes con Media Marginación	0.0306	0.0303	0.0308
Generales Públicas en localidades con más de 100000 habitantes con Baja Marginación	0.0263	0.0260	0.0265
Educación Indígena No Multigrado	0.0436	0.0426	0.0465
Educación Indígena Multigrado	0.0418	0.0393	0.0429
Privadas	0.0277	0.0268	0.0275

Precisión esperada en dominios primarios muestra de tercer grado de secundaria

Dominios Primarios Tercer grado de Secundaria	Planeada	Seleccionada	Modificada
	EE/DS	EE/DS	EE/DS
Aguascalientes	0.0544	0.0541	0.0553
Baja California	0.0528	0.0526	0.0531
Baja California Sur	0.0610	0.0606	0.0614
Campeche	0.0541	0.0534	0.0541
Coahuila	0.0517	0.0514	0.0527
Colima	0.0614	0.0611	0.0630
Chiapas	0.0464	0.0457	0.0466
Chihuahua	0.0506	0.0504	0.0514
Distrito Federal*	0.0483	0.0481	0.0486
Durango	0.0495	0.0484	0.0501
Guanajuato	0.0466	0.0462	0.0473
Guerrero	0.0466	0.0459	0.0465
Hidalgo	0.0468	0.0461	0.0468
Jalisco	0.0484	0.0481	0.0489
México	0.0452	0.0449	0.0455
Michoacán	0.0474	0.0468	0.0479
Morelos	0.0514	0.0511	0.0527
Nayarit	0.0522	0.0515	0.0523
Nuevo León	0.0511	0.0508	0.0516
Oaxaca	0.0468	0.0461	
Puebla	0.0470	0.0464	0.0470
Querétaro	0.0514	0.0510	0.0520
Quintana Roo	0.0540	0.0536	0.0549
San Luis Potosí	0.0467	0.0456	0.0470
Sinaloa	0.0524	0.0520	0.0528
Sonora	0.0516	0.0513	0.0523
Tabasco	0.0508	0.0503	0.0508
Tamaulipas	0.0508	0.0505	0.0511
Tlaxcala	0.0547	0.0544	0.0555
Veracruz	0.0466	0.0459	0.0466
Yucatán	0.0521	0.0516	0.0527
Zacatecas	0.0474	0.0458	0.0477

Precisión esperada en dominios secundarios muestra de tercer grado de secundaria

Dominios secundarios Tercer grado de secundaria	Planeada	Seleccionada	Modificada
	EE/DS	EE/DS	EE/DS
Cursos Comunitarios	0.0411	0.0359	0.0392
Generales Públicas en localidades de 1 a 2 499 habitantes	0.0428	0.0420	0.0438
Generales Públicas en localidades de 2 500 a 99 999 habitantes	0.0331	0.0330	0.0338
Generales Públicas en localidades de más de 100 000 habitantes con Alta y Media Marginación	0.0406	0.0403	0.0416
Generales Públicas en localidades con más de 100 000 habitantes con Baja Marginación	0.0358	0.0356	0.0361
Técnicas en localidades de 1 a 2 499 habitantes	0.0414	0.0408	0.0428
Técnicas en localidades de 2 500 a 99 999 habitantes	0.0388	0.0386	0.0397
Técnicas en localidades con más de 100 000 habitantes	0.0374	0.0373	0.0381
Telesecundarias en localidades de 1 a 499 habitantes	0.0226	0.0218	0.0225
Telesecundarias en localidades de 500 a 2 499 habitantes	0.0205	0.0203	0.0209
Telesecundarias en localidades con más de 2 500 habitantes	0.0395	0.0391	0.0402
Privadas	0.0253	0.0249	0.0253

Cualquier estimación hecha en subpoblaciones diferentes a las mencionadas tendrá que valorarse con base en el error estándar y la desviación estándar estimados. Debe considerarse que las subpoblaciones que se obtienen de combinar los dominios tendrán precisiones menores respecto a los estándares previamente establecidos debido a que el tamaño de muestra disminuye, por lo que se recomienda que las estimaciones que se obtengan se utilicen con cautela. Se reitera que debe tenerse en cuenta que el tamaño de la precisión que aquí se ha presentado únicamente considera al error debido al muestreo. **No se tomó en cuenta el error de medida que se obtendría de la utilización de valores plausibles en los instrumentos** o el error de cualquier otra fuente.

D. Características métricas de los reactivos PLANEA

Extracto del Informe de Análisis Psicométrico de Reactivos PLANEA 2015

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

Una vez que las pruebas PLANEA 2015 fueron administradas a los alumnos muestreados y sus respuestas fueron integradas a archivos electrónicos, es necesario llevar a cabo un análisis de los reactivos con la finalidad de verificar el comportamiento de cada una de las preguntas que se incluirán en los análisis definitivos. Aunque las preguntas ya fueron puestas a prueba con anterioridad en la fase de piloteo, es necesario repetir el procedimiento debido a los ajustes realizados en las propias preguntas, así como su inclusión en una nueva distribución de prueba.

En este anexo se presentan las características métricas de los reactivos de la Evaluación del Logro referida a los Centros Escolares (ELCE) 2015 de las pruebas de Lenguaje y Comunicación, y Matemáticas, para grado evaluado, así como el mapa de Wright en donde se podrá juzgar si las pruebas tienen la amplitud necesaria para la población evaluada. Los campos de las tablas son los siguientes:

Campo	Descripción
Consecutivo	Orden de los reactivos dentro de la prueba.
Ítem	Identificación de cada reactivo.
Cases ítem	Número de sustentantes que respondieron el reactivo.
Disc	Discriminación del reactivo.
Infit	Medida de ajuste.
Delta	Dificultad del reactivo en *logits*.
Score	Puntaje obtenido en la respuesta correcta del reactivo.
Count	Número de sustentantes que respondieron correctamente el reactivo.
P%	Porcentaje de sustentantes que respondieron correctamente el reactivo.
Pt Bis	Correlación Punto biserial.
t	Valor de prueba de significancia de la Pt Bis (Prueba t).
p	*p-value* de la prueba t.
PV Avg	Media de habilidad por categoría de respuesta, en *logits*, calculada con valores plausibles.
PV SD	Desviación estándar de habilidad por categoría de respuesta, en *logits*, calculada con valores plausibles.

Análisis de reactivos Lenguaje y Comunicación, 6° de primaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R01	PEA01	16 753	0.29	1.05	0.99	1	4 695	28.02	0.29	39.08	0.000	0.27	0.86
R02	PEA02	16 750	0.48	0.91	-0.58	1	10 162	60.67	0.48	71.51	0.000	0.21	0.77
R04	PEA04	16 750	0.24	1.11	0.18	1	7 423	44.32	0.24	31.90	0.000	0.12	0.81
R05	PEA05	16 749	0.30	1.05	-0.26	1	9 023	53.87	0.30	41.38	0.000	0.14	0.80
R06	PEA06	16 748	0.36	0.99	-0.99	1	11 546	68.94	0.36	49.81	0.000	0.11	0.78
R07	PEA07	16 748	0.42	0.94	-1.03	1	11 687	69.78	0.42	59.13	0.000	0.13	0.77
R08	PEA08	16 748	0.18	1.15	-0.27	1	9 041	53.98	0.18	23.14	0.000	0.05	0.79
R09	PEA09	16 748	0.31	1.05	0.65	1	5 770	34.45	0.31	42.23	0.000	0.24	0.78
R10	PEA10	16 748	0.48	0.92	0.50	1	6 289	37.55	0.48	71.01	0.000	0.39	0.79
R12	PEA12	16 747	0.47	0.93	0.16	1	7 475	44.63	0.47	68.70	0.000	0.31	0.80
R13	PEA13	16 745	0.44	0.92	-1.18	1	12 129	72.43	0.44	63.34	0.000	0.13	0.77
R14	PEA14	16 745	0.43	0.96	0.48	1	6 337	37.84	0.43	61.69	0.000	0.34	0.82
R15	PEA15	16 744	0.30	1.06	0.40	1	6 626	39.57	0.30	40.85	0.000	0.21	0.84
R16	PEA16	16 743	0.49	0.91	-0.48	1	9 794	58.50	0.49	71.94	0.000	0.23	0.77
R17	PEA17	16 742	0.45	0.94	-0.45	1	9 684	57.84	0.45	65.87	0.000	0.21	0.78
R18	PEA18	16 739	0.34	1.02	-0.33	1	9 246	55.24	0.34	46.84	0.000	0.15	0.79
R19	PEA19	16 739	0.41	0.96	-0.61	1	10 273	61.37	0.41	58.42	0.000	0.17	0.77
R20	PEA20	16 739	0.35	0.97	-1.16	1	12 073	72.12	0.35	49.07	0.000	0.09	0.77
R21	PEA21	16 737	0.40	0.97	-0.59	1	10 208	60.99	0.40	57.11	0.000	0.16	0.78
R22	PEA22	16 737	0.41	0.97	0.56	1	6 071	36.27	0.41	58.78	0.000	0.34	0.83
R23	PEA23	16 737	0.39	0.97	-0.87	1	11 163	66.70	0.39	54.49	0.000	0.13	0.78
R24	PEA24	16 736	0.49	0.90	-0.63	1	10 345	61.81	0.49	72.71	0.000	0.21	0.77
R25	PEA25	16 732	0.37	1.01	-0.35	1	9 334	55.79	0.37	50.98	0.000	0.17	0.79
R26	PEB01	16 727	0.34	1.03	0.20	1	7 340	43.88	0.34	47.16	0.000	0.21	0.82
R27	PEB02	16 717	0.30	1.06	0.19	1	7 352	43.98	0.30	41.01	0.000	0.18	0.82
R28	PEB03	16 714	0.41	0.98	0.20	1	7 337	43.90	0.41	58.48	0.000	0.27	0.81

Análisis de reactivos Lenguaje y Comunicación, 6º de primaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R29	PEB04	16 708	0.33	1.04	0.22	1	7 257	43.43	0.33	45.26	0.000	0.21	0.82
R30	PEB05	16 704	0.52	0.89	-0.29	1	9 088	54.41	0.52	78.56	0.000	0.27	0.78
R31	PEB06	16 689	0.45	0.95	0.12	1	7 583	45.44	0.45	64.81	0.000	0.29	0.81
R32	PEB07	16 670	0.30	1.06	0.21	1	7 255	43.52	0.30	41.22	0.000	0.19	0.84
R33	PEB08	16 665	0.27	1.08	0.31	1	6 904	41.43	0.27	36.35	0.000	0.17	0.82
R34	PEB09	16 661	0.46	0.94	0.09	1	7 706	46.25	0.46	67.26	0.000	0.30	0.79
R35	PEB10	16 656	0.27	1.08	0.47	1	6 343	38.08	0.27	36.61	0.000	0.18	0.81
R36	PEB11	16 646	0.40	0.98	-0.07	1	8 266	49.66	0.40	56.68	0.000	0.23	0.81
R37	PEB12	16 639	0.43	0.95	-0.67	1	10 427	62.67	0.43	60.63	0.000	0.17	0.77
R38	PEB13	16 635	0.29	1.01	1.57	1	3 144	18.90	0.29	39.11	0.000	0.39	0.91
R39	PEB14	16 625	0.31	1.05	-0.08	1	8 292	49.88	0.31	41.90	0.000	0.16	0.80
R40	PEB15	16 620	0.38	0.99	-0.28	1	9 020	54.27	0.38	53.56	0.000	0.19	0.80
R41	PEB16	16 579	0.40	0.98	0.37	1	6 664	40.20	0.40	55.77	0.000	0.29	0.83
R42	PEB17	16 568	0.33	1.03	0.70	1	5 548	33.49	0.33	44.90	0.000	0.28	0.84
R43	PEB18	16 560	0.32	1.04	0.01	1	7 928	47.87	0.32	43.80	0.000	0.18	0.81
R44	PEB19	16 548	0.38	1.00	-0.06	1	8 188	49.48	0.38	52.59	0.000	0.21	0.79
R45	PEB20	16 517	0.43	0.96	0.08	1	7 662	46.39	0.43	60.83	0.000	0.27	0.82
R46	PEB21	16 501	0.47	0.92	0.69	1	5 557	33.68	0.47	68.58	0.000	0.42	0.83
R47	PEB22	16 496	0.20	1.13	0.69	1	5 574	33.79	0.20	25.72	0.000	0.14	0.82
R48	PEB23	16 489	0.36	1.01	0.34	1	6 738	40.86	0.36	49.57	0.000	0.25	0.80
R49	PEB24	16 468	0.31	1.05	0.38	1	6 588	40.00	0.31	41.26	0.000	0.21	0.82
R50	PEB25	16 444	0.34	1.03	0.45	1	6 344	38.58	0.34	45.65	0.000	0.25	0.85

Mapa de Wright Lenguaje y Comunicación, 6º de primaria

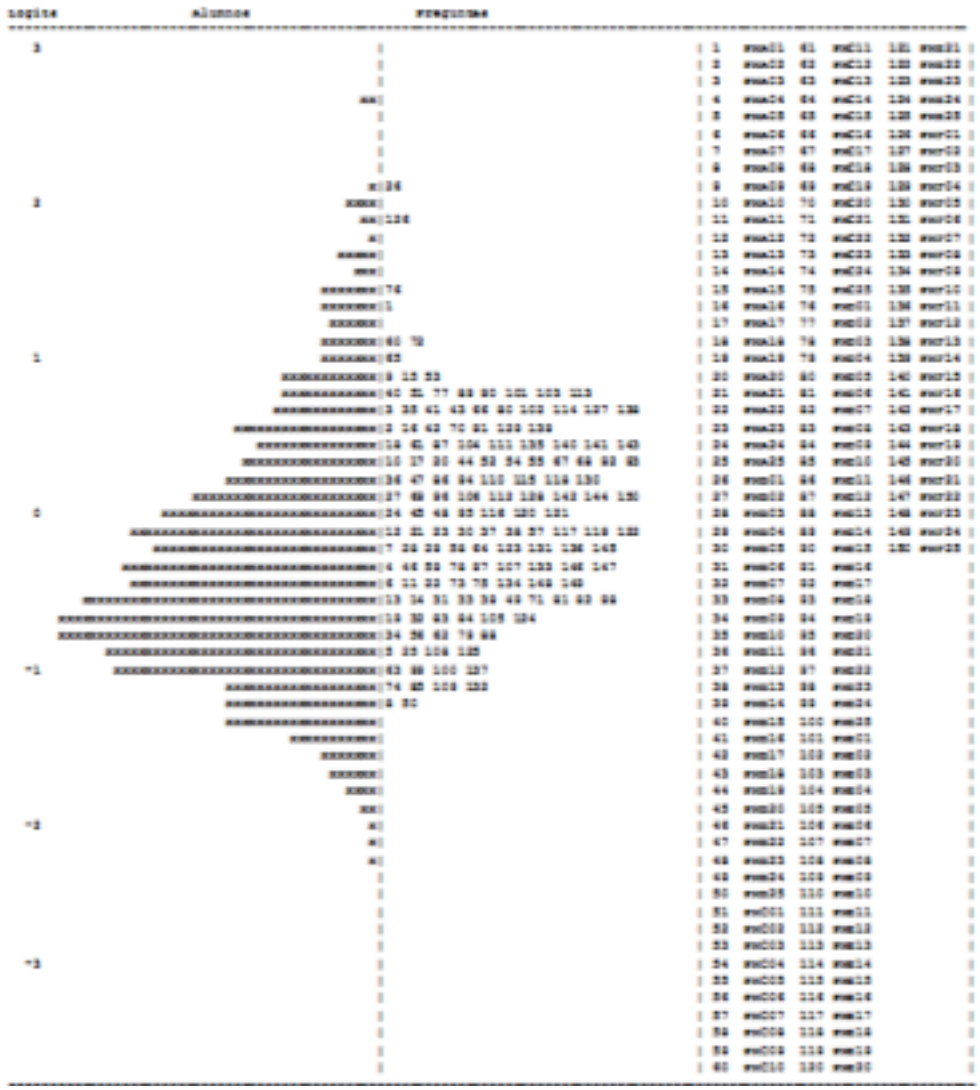
Análisis de reactivos Matemáticas, 6º de primaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R01	PMA01	16 576	0.26	1.06	1.43	1	2 987	18.02	0.26	35.35	0.000	0.16	1.07
R02	PMA02	16 576	0.35	1.04	0.46	1	5 575	33.63	0.35	47.54	0.000	0.10	0.95
R03	PMA03	16 576	0.47	0.93	0.70	1	4 841	29.20	0.47	69.38	0.000	0.30	0.93
R04	PMA04	16 576	0.31	1.06	-0.43	1	8 629	52.06	0.31	42.44	0.000	-0.06	0.91
R05	PMA05	16 576	0.41	0.96	-1.08	1	10 887	65.68	0.41	57.97	0.000	-0.06	0.88
R06	PMA06	16 576	0.35	1.03	-0.59	1	9 194	55.47	0.35	47.61	0.000	-0.05	0.91
R07	PMA07	16 576	0.36	1.04	-0.07	1	7 357	44.38	0.36	49.14	0.000	0.02	0.93
R08	PMA08	16 576	0.39	0.96	-1.43	1	11 993	72.35	0.39	53.85	0.000	-0.11	0.87
R09	PMA09	16 576	0.34	1.03	0.86	1	4 377	26.41	0.34	47.21	0.000	0.17	1.01
R10	PMA10	16 576	0.36	1.04	0.32	1	6 030	36.38	0.36	49.09	0.000	0.08	0.95
R11	PMA11	16 576	0.48	0.92	-0.52	1	8 942	53.95	0.48	70.03	0.000	0.05	0.89
R12	PMA12	16 576	0.47	0.94	-0.20	1	7 804	47.08	0.47	68.57	0.000	0.10	0.92
R13	PMA13	16 576	0.44	0.95	-0.71	1	9 604	57.94	0.44	63.67	0.000	0.00	0.88
R14	PMA14	16 576	0.49	0.91	-0.61	1	9 273	55.94	0.49	72.14	0.000	0.05	0.88
R15	PMA15	16 576	0.50	0.91	0.81	1	4 536	27.36	0.50	73.93	0.000	0.36	0.96
R16	PMA16	16 576	0.34	1.04	0.53	1	5 351	32.28	0.34	47.05	0.000	0.10	1.00
R17	PMA17	16 576	0.54	0.88	0.13	1	6 644	40.08	0.54	83.33	0.000	0.23	0.90
R18	PMA18	16 576	0.31	1.07	0.49	1	5 465	32.97	0.31	41.36	0.000	0.05	0.98
R19	PMA19	16 576	0.47	0.91	-0.97	1	10 508	63.39	0.47	68.29	0.000	-0.02	0.87
R20	PMA20	16 576	0.40	1.00	0.38	1	5 826	35.15	0.40	55.83	0.000	0.13	0.95
R21	PMA21	16 576	0.48	0.93	-0.15	1	7 644	46.11	0.48	70.32	0.000	0.12	0.90
R22	PMA22	16 576	0.48	0.92	-0.62	1	9 294	56.07	0.48	70.78	0.000	0.04	0.89
R23	PMA23	16 576	0.43	0.97	-0.30	1	8 171	49.29	0.43	61.44	0.000	0.05	0.91
R24	PMA24	16 576	0.44	0.96	-0.21	1	7 824	47.20	0.44	63.76	0.000	0.07	0.92
R25	PMA25	16 576	0.41	0.95	-1.20	1	11 271	68.00	0.41	58.68	0.000	-0.07	0.87

Análisis de reactivos Matemáticas, 6º de primaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R26	PMB01	16 575	0.17	1.08	2.04	1	1 883	11.36	0.17	22.38	0.000	0.09	1.19
R27	PMB02	16 574	0.34	1.05	0.18	1	6 488	39.15	0.34	47.14	0.000	0.04	0.93
R28	PMB03	16 574	0.31	1.08	-0.23	1	7 926	47.82	0.31	41.31	0.000	-0.04	0.91
R29	PMB04	16 573	0.30	1.08	-0.11	1	7 497	45.24	0.30	40.17	0.000	-0.03	0.93
R30	PMB05	16 573	0.35	1.05	-0.01	1	7 138	43.07	0.35	47.42	0.000	0.02	0.95
R31	PMB06	16 572	0.38	1.01	-0.40	1	8 503	51.31	0.38	52.17	0.000	-0.01	0.90
R32	PMB07	16 570	0.49	0.92	-0.33	1	8 279	49.96	0.49	72.57	0.000	0.10	0.89
R33	PMB08	16 569	0.47	0.93	-0.46	1	8 725	52.66	0.47	69.16	0.000	0.05	0.90
R34	PMB09	16 567	0.49	0.92	-0.48	1	8 789	53.05	0.49	72.52	0.000	0.07	0.90
R35	PMB10	16 566	0.27	1.10	0.70	1	4 833	29.17	0.27	36.06	0.000	0.04	0.99
R36	PMB11	16 566	0.34	1.05	0.33	1	5 989	36.15	0.34	46.63	0.000	0.07	0.96
R37	PMB12	16 565	0.36	1.03	0.00	1	7 082	42.75	0.36	50.16	0.000	0.04	0.92
R38	PMB13	16 565	0.36	1.03	-0.13	1	7 550	45.58	0.36	50.44	0.000	0.02	0.95
R39	PMB14	16 561	0.32	1.05	-0.49	1	8 825	53.29	0.32	43.99	0.000	-0.06	0.90
R40	PMB15	16 561	0.39	0.99	0.77	1	4 641	28.02	0.39	54.66	0.000	0.21	0.99
R41	PMB16	16 560	0.41	0.98	0.58	1	5 207	31.44	0.41	57.04	0.000	0.18	1.01
R42	PMB17	16 558	0.33	1.05	0.68	1	4 894	29.56	0.33	44.82	0.000	0.11	1.01
R43	PMB18	16 557	0.26	1.11	0.66	1	4 952	29.91	0.26	34.65	0.000	0.03	0.98
R44	PMB19	16 554	0.49	0.92	0.46	1	5 562	33.60	0.49	72.85	0.000	0.26	0.96
R45	PMB20	16 553	0.41	0.99	0.39	1	5 783	34.94	0.41	57.91	0.000	0.15	0.95
R46	PMB21	16 552	0.41	0.99	-0.13	1	7 546	45.59	0.41	57.35	0.000	0.05	0.93
R47	PMB22	16 549	0.41	0.99	0.38	1	5 826	35.20	0.41	57.70	0.000	0.15	0.97
R48	PMB23	16 547	0.36	1.03	0.15	1	6 561	39.65	0.36	50.04	0.000	0.06	0.94
R49	PMB24	16 542	0.37	1.01	-0.46	1	8 710	52.65	0.37	51.55	0	-0.01	0.91
R50	PMB25	16 509	0.42	0.95	-1.11	1	10 953	66.35	0.42	58.71	0.000	-0.06	0.89

Mapa de Wright Matemáticas, 6º de primaria



Análisis de reactivos Lenguaje y Comunicación, 3° de secundaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R01	SEA01	23 026	0.30	1.03	0.12	1	11 562	50.21	0.30	47.18	0.000	0.31	0.67
R02	SEA02	23 026	0.43	0.93	-0.57	1	15 046	65.34	0.43	73.06	0.000	0.32	0.63
R03	SEA03	23 025	0.38	0.94	-1.36	1	18 315	79.54	0.38	63.24	0.000	0.24	0.64
R04	SEA04	23 025	0.29	1.02	-0.49	1	14 661	63.67	0.29	46.58	0.000	0.26	0.66
R05	SEA05	23 025	0.29	1.03	-0.04	1	12 397	53.84	0.29	46.35	0.000	0.29	0.66
R06	SEA06	23 025	0.16	1.08	-0.79	1	16 067	69.78	0.16	24.82	0.000	0.19	0.66
R07	SEA07	23 025	0.34	0.96	-1.48	1	18 748	81.42	0.34	53.98	0.000	0.22	0.64
R08	SEA08	23 025	0.36	0.98	-0.45	1	14 437	62.70	0.36	57.79	0.000	0.29	0.64
R09	SEA09	23 025	0.32	0.99	-0.85	1	16 336	70.95	0.32	51.32	0.000	0.25	0.65
R10	SEA10	23 024	0.21	1.08	0.29	1	10 664	46.32	0.21	31.91	0.000	0.26	0.65
R11	SEA11	23 023	0.37	0.98	-0.45	1	14 466	62.83	0.37	59.71	0.000	0.30	0.64
R12	SEA12	23 023	0.37	0.98	-0.16	1	12 971	56.34	0.37	60.79	0.000	0.32	0.64
R13	SEA13	23 022	0.33	1.00	0.16	1	11 339	49.25	0.33	53.61	0.000	0.33	0.66
R14	SEA14	23 022	0.24	1.06	0.33	1	10 423	45.27	0.24	37.32	0.000	0.29	0.66
R15	SEA15	23 022	0.26	1.04	0.85	1	7 838	34.05	0.26	40.13	0.000	0.34	0.67
R16	SEA16	23 022	0.25	1.03	1.06	1	6 883	29.90	0.25	38.84	0.000	0.36	0.68
R17	SEA17	23 021	0.35	0.98	-0.54	1	14 883	64.65	0.35	57.16	0.000	0.28	0.64
R18	SEA18	23 020	0.40	0.95	-0.54	1	14 882	64.65	0.40	67.04	0.000	0.31	0.64
R19	SEA19	23 019	0.43	0.92	-1.19	1	17 703	76.91	0.43	71.26	0.000	0.27	0.63
R20	SEA20	23 018	0.32	1.01	-0.15	1	12 926	56.16	0.32	51.95	0.000	0.30	0.65
R21	SEA21	23 016	0.38	0.98	0.30	1	10 611	46.10	0.38	61.68	0.000	0.37	0.67
R22	SEA22	23 015	0.20	1.08	0.62	1	8 953	38.90	0.20	30.43	0.000	0.28	0.68
R23	SEA23	23 015	0.21	1.06	0.93	1	7 475	32.48	0.21	33.38	0.000	0.32	0.66
R24	SEA24	23 014	0.36	0.98	-0.39	1	14 137	61.43	0.36	58.00	0.000	0.29	0.65
R25	SEA25	23 014	0.27	1.02	1.11	1	6 682	29.03	0.27	42.05	0.000	0.39	0.70

Análisis de reactivos Lenguaje y Comunicación, 3º de secundaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R26	SEB01	23 009	0.24	1.06	-0.05	1	12 400	53.89	0.24	36.83	0.000	0.26	0.65
R27	SEB02	23 006	0.38	0.97	0.10	1	11 622	50.52	0.38	62.95	0.000	0.36	0.63
R28	SEB03	23 002	0.29	1.03	0.13	1	11 480	49.91	0.29	46.72	0.000	0.31	0.65
R29	SEB04	23 001	0.40	0.96	-0.36	1	14 017	60.94	0.40	65.69	0.000	0.32	0.64
R30	SEB05	22 994	0.33	1.01	-0.09	1	12 630	54.93	0.33	52.44	0.000	0.30	0.65
R32	SEB07	22 979	0.20	1.07	0.96	1	7 312	31.82	0.20	30.62	0.000	0.30	0.68
R33	SEB08	22 971	0.40	0.95	-0.82	1	16 148	70.30	0.40	66.48	0.000	0.28	0.64
R34	SEB09	22 959	0.53	0.88	-0.28	1	13 566	59.09	0.53	95.03	0.000	0.39	0.62
R35	SEB10	22 954	0.44	0.94	-0.03	1	12 301	53.59	0.44	74.70	0.000	0.38	0.63
R36	SEB11	22 949	0.42	0.94	-0.52	1	14 770	64.36	0.42	69.89	0.000	0.32	0.63
R37	SEB12	22 928	0.45	0.93	-0.26	1	13 444	58.64	0.45	76.92	0.000	0.35	0.64
R38	SEB13	22 917	0.32	1.01	0.64	1	8 831	38.53	0.32	50.80	0.000	0.37	0.66
R39	SEB14	22 911	0.31	1.02	0.27	1	10 733	46.85	0.31	49.06	0.000	0.33	0.66
R40	SEB15	22 906	0.15	1.08	1.21	1	6 225	27.18	0.15	22.73	0.000	0.27	0.70
R42	SEB17	22 846	0.26	1.05	-0.01	1	12 116	53.03	0.26	41.33	0.000	0.28	0.67
R43	SEB18	22 841	0.26	1.03	0.90	1	7 577	33.17	0.26	41.50	0.000	0.36	0.69
R44	SEB19	22 831	0.45	0.93	0.37	1	10 168	44.54	0.45	76.12	0.000	0.43	0.66
R45	SEB20	22 817	0.14	1.12	0.47	1	9 630	42.21	0.14	21.81	0.000	0.23	0.65
R46	SEB21	22 790	0.28	1.02	0.90	1	7 564	33.19	0.28	44.87	0.000	0.38	0.69
R47	SEB22	22 758	0.37	0.98	0.32	1	10 412	45.75	0.37	60.64	0.000	0.38	0.65
R48	SEB23	22 745	0.48	0.91	-0.09	1	12 503	54.97	0.48	82.77	0.000	0.39	0.64
R49	SEB24	22 738	0.49	0.90	-0.56	1	14 818	65.17	0.49	84.19	0.000	0.34	0.62
R50	SEB25	22 605	0.20	1.08	0.50	1	9 416	41.65	0.20	30.99	0.000	0.27	0.65

Mapa de Wright Lenguaje y Comunicación, 3º de secundaria

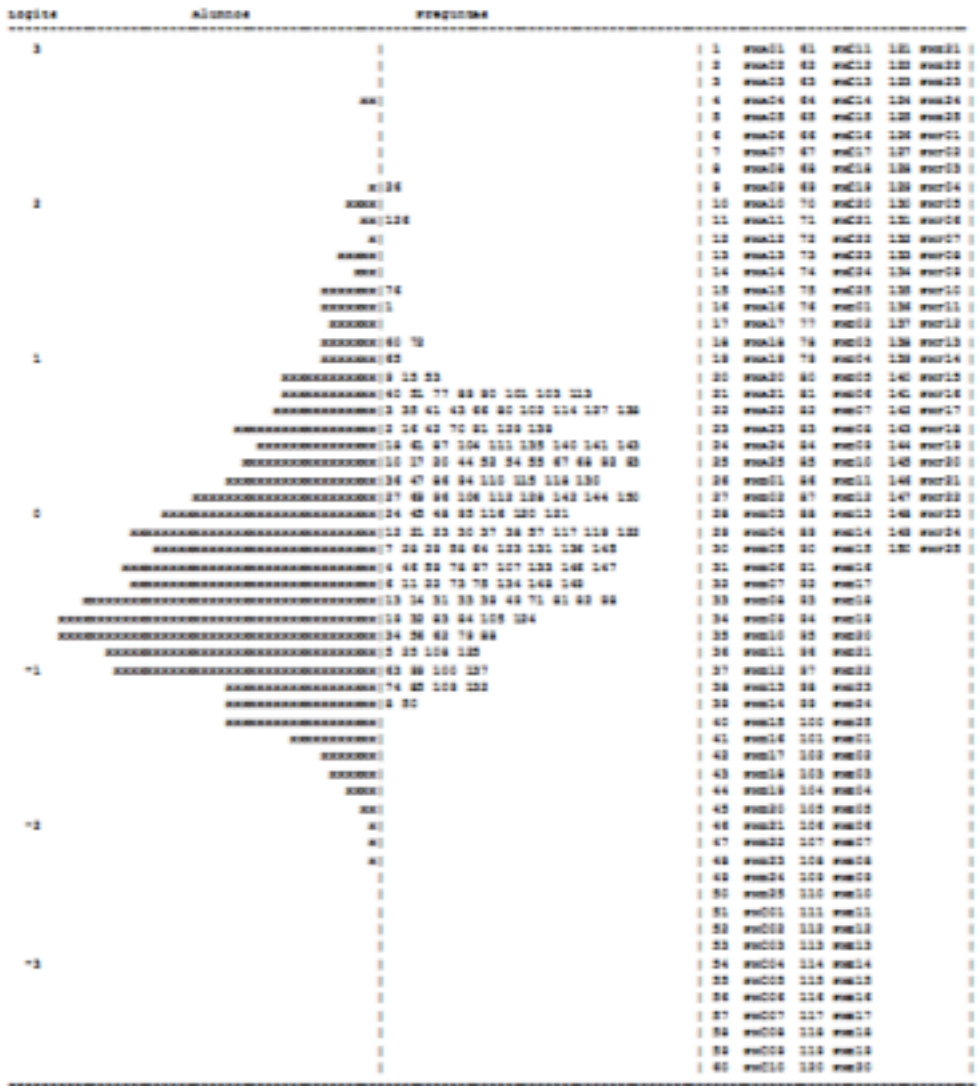
Análisis de reactivos Matemáticas, 3º de secundaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R01	SMA01	22 805	0.38	0.97	0.15	1	8 045	35.28	0.38	61.86	0.000	-0.21	0.70
R02	SMA02	22 805	0.31	1.00	-0.87	1	13 220	57.97	0.31	49.96	0.000	-0.35	0.67
R03	SMA03	22 804	0.49	0.90	0.14	1	8 054	35.32	0.49	85.50	0.000	-0.13	0.68
R04	SMA04	22 804	0.31	1.01	0.32	1	7 218	31.65	0.31	48.87	0.000	-0.25	0.72
R05	SMA05	22 804	0.27	1.03	0.14	1	8 058	35.34	0.27	42.62	0.000	-0.29	0.67
R06	SMA06	22 804	0.28	1.02	0.16	1	7 998	35.07	0.28	44.36	0.000	-0.29	0.71
R07	SMA07	22 804	0.38	0.97	0.00	1	8 754	38.39	0.38	61.75	0.000	-0.23	0.69
R08	SMA08	22 803	0.39	0.94	-1.36	1	15 619	68.50	0.39	63.94	0.000	-0.35	0.65
R09	SMA09	22 803	0.28	1.02	0.31	1	7 269	31.88	0.28	43.91	0.000	-0.27	0.70
R10	SMA10	22 803	0.29	1.02	-0.28	1	10 140	44.47	0.29	46.34	0.000	-0.32	0.68
R11	SMA11	22 800	0.18	1.08	-0.20	1	9 765	42.83	0.18	26.87	0.000	-0.39	0.67
R12	SMA12	22 800	0.39	0.96	-0.07	1	9 117	39.99	0.39	63.06	0.000	-0.24	0.69
R13	SMA13	22 800	0.36	0.97	0.89	1	4 908	21.53	0.36	58.32	0.000	-0.11	0.77
R14	SMA14	22 800	0.22	1.05	-0.57	1	11 687	51.26	0.22	34.67	0.000	-0.38	0.64
R15	SMA15	22 800	0.30	1.00	0.69	1	5 691	24.96	0.30	47.60	0.000	-0.20	0.74
R16	SMA16	22 800	0.27	1.03	-0.17	1	9 628	42.23	0.27	42.36	0.000	-0.32	0.70
R17	SMA17	22 800	0.31	1.00	-0.30	1	10 258	44.99	0.31	49.98	0.000	-0.30	0.67
R18	SMA18	22 800	0.27	1.01	1.08	1	4 286	18.80	0.27	41.65	0.000	-0.19	0.76
R19	SMA19	22 800	0.39	0.96	-0.06	1	9 060	39.74	0.39	63.10	0.000	-0.23	0.67
R20	SMA20	22 798	0.33	0.98	-1.21	1	14 901	65.36	0.33	53.12	0.000	-0.37	0.65
R21	SMA21	22 798	0.42	0.94	0.46	1	6 605	28.97	0.42	69.28	0.000	-0.13	0.70
R22	SMA22	22 797	0.46	0.92	0.08	1	8 368	36.71	0.46	77.42	0.000	-0.16	0.67
R23	SMA23	22 797	0.44	0.92	-0.95	1	13 620	59.74	0.44	74.68	0.000	-0.30	0.65
R24	SMA24	22 797	0.28	1.02	0.42	1	6 775	29.72	0.28	44.27	0.000	-0.26	0.72
R25	SMA25	22 797	0.34	0.99	-0.39	1	10 734	47.09	0.34	55.27	0.000	-0.30	0.67

Análisis de reactivos Matemáticas, 3º de secundaria

Consecutivo	Ítem	Cases ítem	Disc	Infit	Delta	Score	Count	P%	Pt Bis	t	p	PV Avg	PV SD
R26	SMB01	22 796	0.46	0.92	-0.40	1	10 766	47.23	0.46	77.60	0.000	-0.23	0.66
R27	SMB02	22 795	0.44	0.94	-0.27	1	10 121	44.40	0.44	73.14	0.000	-0.23	0.68
R28	SMB03	22 795	0.31	1.00	0.54	1	6 277	27.54	0.31	50.02	0.000	-0.22	0.72
R29	SMB04	22 793	0.30	1.01	0.02	1	8 653	37.96	0.30	47.53	0.000	-0.28	0.70
R30	SMB05	22 793	0.27	1.03	-0.13	1	9 375	41.13	0.27	41.52	0.000	-0.32	0.67
R31	SMB06	22 792	0.38	0.97	0.31	1	7 277	31.93	0.38	61.14	0.000	-0.19	0.69
R32	SMB07	22 790	0.25	1.04	-0.07	1	9 101	39.93	0.25	39.25	0.000	-0.33	0.67
R33	SMB08	22 787	0.31	1.01	0.05	1	8 505	37.32	0.31	49.46	0.000	-0.28	0.69
R34	SMB09	22 787	0.18	1.06	0.79	1	5 292	23.22	0.18	27.17	0.000	-0.32	0.74
R35	SMB10	22 787	0.33	0.98	0.89	1	4 911	21.55	0.33	52.56	0.000	-0.15	0.77
R36	SMB11	22 787	0.28	1.01	0.57	1	6 142	26.95	0.28	44.29	0.000	-0.24	0.72
R37	SMB12	22 787	0.41	0.95	-0.57	1	11 642	51.09	0.41	67.65	0.000	-0.28	0.67
R38	SMB13	22 785	0.44	0.93	0.12	1	8 172	35.87	0.44	74.81	0.000	-0.16	0.69
R39	SMB14	22 783	0.25	1.04	0.03	1	8 587	37.69	0.25	38.32	0.000	-0.32	0.69
R40	SMB15	22 779	0.18	1.08	0.02	1	8 651	37.98	0.18	28.36	0.000	-0.37	0.69
R41	SMB16	22 771	0.37	0.97	-0.72	1	12 442	54.64	0.37	60.59	0.000	-0.31	0.67
R42	SMB17	22 764	0.16	1.09	0.26	1	7 506	32.97	0.16	24.73	0.000	-0.37	0.66
R43	SMB18	22 759	0.28	1.02	-0.07	1	9 093	39.95	0.28	44.80	0.000	-0.31	0.71
R44	SMB19	22 751	0.30	1.01	0.28	1	7 394	32.50	0.30	47.64	0.000	-0.26	0.71
R45	SMB20	22 747	0.42	0.95	-0.49	1	11 225	49.35	0.42	68.80	0.000	-0.26	0.67
R46	SMB21	22 746	0.21	1.06	0.32	1	7 230	31.79	0.21	31.94	0.000	-0.33	0.69
R47	SMB22	22 739	0.14	1.08	0.86	1	5 014	22.05	0.14	21.09	0.000	-0.36	0.73
R48	SMB23	22 726	0.39	0.94	-1.20	1	14 828	65.25	0.39	64.13	0.000	-0.34	0.64
R49	SMB24	22 718	0.20	1.06	0.44	1	6 676	29.39	0.20	31.04	0.000	-0.33	0.71
R50	SMB25	22 663	0.27	1.03	0.01	1	8 630	38.08	0.27	42.26	0.000	-0.31	0.68

Mapa de Wright Matemáticas, 3º de secundaria



E. Cálculo de los factores de expansión de los alumnos para las muestras controladas por el INEE de PLANEA 2015

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

En este anexo se describe la forma en la que se definió cada componente del peso de estimación de habilidad de los estudiantes participantes en el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) 2015 y el procedimiento por el que los componentes se ensamblan en un peso final.

Propósito de la ponderación

El propósito de utilizar los factores de expansión o pesos muestrales (Kalton, 1983, p. 69; Särndal, Swensson y Wretman, 1992, p. 42; Sharon, 1999, pp. 223-227) sobre los archivos de datos para las evaluaciones a gran escala y otras encuestas, es ayudar a los usuarios de los datos a obtener estimaciones de parámetros poblacionales que:

1. No sufran sesgo resultante del uso de un diseño muestral complejo.
2. Tener un sesgo mínimo debido a la presencia de no respuesta, en la medida de lo posible.

Elementos del peso de estimación de los estudiantes (peso final)

Las Unidades Primarias de Muestreo (UPM) para PLANEA fueron constituidas por la combinación de la clave del centro de trabajo y el turno.

La extracción de los alumnos se llevó a cabo en dos etapas. En una primera etapa de muestreo las UPM fueron extraídas mediante muestreo sistemático y tuvieron una probabilidad de selección proporcional a su tamaño. En una segunda etapa se hizo una extracción aleatoria simple de alumnos en cada UPM provenientes de la primera etapa de muestreo.

Derivado de este proceso se produjeron probabilidades de selección para cada una de las UPM y otras probabilidades de selección para cada uno de los alumnos dentro de las UPM seleccionadas.

Básicamente, los pesos finales de los alumnos son el producto de los inversos de las probabilidades de selección base (o bajo el diseño muestral) mencionados arriba y uno o varios factores de ajuste.

La selección de las UPM se hizo utilizando un marco inicial construido a partir de los datos del Formato 911 (ciclo 2013-2014). La selección final de los alumnos se hizo utilizando un marco actualizado construido a partir de los datos que brindó la Secretaría de Educación Pública (SEP) al Instituto Nacional para la Evaluación de la Educación (INEE) en los que se tuvo información más reciente de las UPM y los estudiantes.

En este proceso varias UPM fueron cerradas, lo que generó pequeñas variaciones en las cantidades de alumnos. Todos estos cambios se consideran en el cálculo de los factores de expansión.

Las notaciones que se utilizarán en este documento son:

- Como es usual, las letras h, i y j se utilizan como subíndices, las letras minúsculas se refieren a la muestra mientras que las letras mayúsculas se refieren a la población.
- Existen H estratos explícitos; indexados por $h=1,\ldots,H$.
- En cada estrato explícito se extrajo una muestra S_{Ih} de n_{Ih} UPM indexadas por $i=1,\ldots,n_{Ih}$ de una población U_{Ih} de N_{Ih} que comprenden el estrato h.
- Cada UPM $i=1,\ldots,n_{Ih}$ dentro del estrato explícito h tiene una medida de tamaño denotada por N_{hi}. La medida de tamaño fue la cantidad total de alumnos por grado de cada UPM.
- En cada UPM que se considera que respondió se extrajo una muestra de m_{hi} alumnos.
- En cada UPM participante existe un total de M_{hi} grupos.

Peso base de las UPM (peso de las UPM bajo el diseño muestral)

Las UPM que contienen a los alumnos de sexto de primaria y tercero de secundaria son poblaciones con una gran variación en los valores de N_{hi}, por lo que no fue posible satisfacer estrictamente la regla de proporcionalidad entre las probabilidades de selección y los tamaños de las UPM. Para resolver este problema se definieron las probabilidades de selección de tal forma que las UPM de tamaño muy grande tuvieron probabilidad 1 de ser seleccionadas y de esta forma el resto recupera la regla de proporcionalidad (Särndal *et al.*, 1992, pp. 89-90).

Usando la notación antes descrita, para cada UPM $i=1,\ldots,n_{Ih}$ y cada estrato explícito $h=1,\ldots,H$, el peso base de la UPM está dado por:

$$WGTFAC1_{hi}=\begin{cases}1 \quad si \quad i\in A_h\\ \dfrac{\Sigma_{UIh}-A_hN_{hi}}{\left(n_{Ih}-n_{Ah}\right)N_{hi}} si\ i\in U_{Ih}-A_h\end{cases}$$

Donde A_h es el conjunto de n_{Ah} UPM de inclusión forzosa que satisfacen $n_{Ih}N_{hi}>\Sigma_{UIh}N_{hi}$

Factor de ajuste de las UPM por no respuesta

Bajo el supuesto de que la falta de respuesta de las UPM sucedió por motivos no relacionados con la variable de estudio (*Missing completely at Random*) (Little y Rubin, 2002, p. 12) un factor de ajuste es necesario dentro de cada estrato explícito (Rutkowski, von Davier y Rutkowski, 2014, p. 136).

Para cada estrato explícito $h = 1,\ldots,H$, el factor de ajuste por no respuesta está dado por:

$$WGTADJ1h = \begin{cases} \frac{n_{Ah} - d_{Ah}}{r_{Ah}} & si\ i \in A_h \\ \frac{(n_{Ih} - n_{Ah})d_n}{r_h} & si\ i \in U_{Ih} - A_h \end{cases}$$

1 *si la UPM fue cerrada*
0 *si la UPM no responde*

Donde:

r_h es la cantidad de UPM que participaron en PLANEA 2015 de n_{Ih} seleccionadas en cada estrato y que no son de inclusión forzosa.
r_{Ah} es la cantidad de UPM que participaron en PLANEA 2015 de n_{Ih} seleccionadas en cada estrato y que son de inclusión forzosa.
d_h es la cantidad de UPM que se cerraron en el marco actualizado de muestreo y que no son de inclusión forzosa.
d_{Ah} es la cantidad de UPM que se cerraron en el marco actualizado de muestreo y que son de inclusión forzosa.

Peso base del alumno (peso del alumno bajo el diseño muestral)

En cada UPM participante se extrajo una muestra aleatoria simple de estudiantes de tamaño m_{hi} .
Para cada alumno seleccionado $j = 1,\ldots,m_{hi}$ de la UPM $i = 1,\ldots,n_{Ih}$ en un estrato explícito $h = 1,\ldots,H$ el peso base está dado por:

$$WGTFAC2_{hij} = \frac{N_{\overline{hi}}}{m_{\overline{hi}}}$$

En donde $N_{\overline{hi}}$ es la cantidad total de alumnos por grado en el marco actualizado para cada UPM.

Factor de ajuste de alumnos por no respuesta

Desafortunadamente existen diversas razones por las que no todos los estudiantes seleccionados para participar en la muestra lo pueden hacer, por lo que bajo el supuesto de pérdida aleatoria se construye un nuevo factor de ajuste por no respuesta de los alumnos. Se considera que únicamente responden $m_{\underset{hi}{*}}$ de los m_{hi} extraídos (Rutkowski *et al.*, 2014, p. 138).

El factor de ajuste es el siguiente:

$$WGTADJ2_{hij} = \begin{cases} \dfrac{m_{hi}}{m_{\underset{hi}{*}}} & si\ responde \\ 0 & si\ no\ responde \end{cases}$$

Peso final del alumno

El peso final del alumno es simplemente el producto de los pesos bases junto con sus factores de ajuste por falta de respuesta y queda como sigue:

$$W\_FSTRO_{hij} = WGTFAC1_{hi} \times WGTADJ1_{h} \times WGTFAC2_{hij} \times WGTADJ2_{hij}$$

Cabe recalcar que es indispensable el uso de estos factores de expansión de los estudiantes para obtener resultados insesgados y se recomienda usarlos junto con una técnica apropiada de estimación del error de muestreo que considere métodos de replicación pues es la única vía posible para tener una buena aproximación del error estándar. Por ejemplo: en este diseño se utilizó muestreo sistemático y se sabe que para la estimación de un total es un hecho que no existe un estimador insesgado de la varianza del estimador (Särndal *et al.*, 1992, p. 83).

Peso base del grupo (peso del grupo bajo el diseño muestral)

Cuando una UPM es seleccionada para la muestra, todos los grupos presentes en ella son evaluados. Ello implica que la probabilidad de inclusión de un grupo en la muestra de grupos sea igual a la probabilidad de selección de la UPM a la que pertenece. Así, el peso base del grupo $k = 1,\ldots,M_{hi}$ de la UPM $i = 1,\ldots,n_{Ih}$ perteneciente al estrato $h = 1,\ldots,H$ es igual al peso de la UPM i del estrato h, denotado como $WGTFAC1_{hi}$ en la definición proporcionada previamente.

Factor de ajuste de grupos por no respuesta

A pesar de que todos los grupos que pertenecen a la UPM seleccionada deben ser evaluados, existen causas por las que algunos grupos no pueden participar. Suponiendo pérdida aleatoria, se construye el siguiente factor de ajuste por no respuesta para cada grupo $k = 1,\ldots, M_{*_{hi}}$:

$$WGTADJG1_{hik} = \begin{cases} \dfrac{M_{hi}}{M_{*_{hi}}} & si\ M_{*_{hi}} > M_{hi}, \\ 1 \quad si\ M_{*_{hi}} \leq M_{hi} \end{cases}$$

Donde:

$M_{*_{hi}}$ es el número de grupos que participaron en PLANEA 2015 de los M_{hi} grupos totales en la UPM i del estrato explícito h

$M_{*_{hi}}$ puede ser mayor que M_{hi} debido a la actualización del total de grupos en el marco de la SEP.

Peso final del grupo

El peso final del grupo es igual al producto del peso base del grupo por el factor de ajuste por no respuesta asociado a la UPM y estrato correspondiente:

$$GRPWGT_{hik} = WGTFAC1_{hi} \times WGTADJDG1_{hik}$$

Peso base del director (peso del director bajo el diseño muestral)

El peso base de cada director es igual al peso base de la UPM de la que proviene, ya que al seleccionar una UPM en la muestra, el director que le corresponde es seleccionado en la muestra de directores. Es decir, obtener una muestra de directores es equivalente a obtener una muestra de las UPM. Por lo tanto, en cada estrato explícito se extrae una muestra S_{Ih} de n_{Ih} directores de una población U_{Ih} de N_{Ih} directores en el estrato h.

Usando la notación antes descrita, para cada director perteneciente a la UPM $i = 1,\ldots, n_{Ih}$, y cada estrato explícito $h = 1,\ldots, H$, el peso base del director está dado por $WGTFAC1_{hi}$, tal como fue definido previamente.

Factor de ajuste de directores por no respuesta

Al igual que en el caso de las UPM y de los alumnos, existen razones por las que un director seleccionado en la muestra no puede participar en la evaluación, por lo cual es necesario aplicar un factor de ajuste. Suponiendo pérdida aleatoria, el factor de ajuste por no respuesta es:

$$WGTADJD1_h = \begin{cases} \dfrac{n_{Ah} - d_{Ah}}{r_{Ah}} & si\ i \in A_h \\ \dfrac{(n_{Ih} - n_{Ah}) - d_h}{r_h} & si\ i \in UIh - A_h, \\ \begin{matrix} 1 & si\ la\ UPM\ fue\ cerrada \\ 0 & si\ el\ director\ no\ responde \end{matrix} & \end{cases}$$

Donde:

r_h es la cantidad de directores que participaron en PLANEA 2015 de n_{Ih} seleccionados en cada estrato y que no son de inclusión forzosa.
r_{Ah} es la cantidad de directores que participaron en PLANEA 2015 de n_{Ih} seleccionados en cada estrato y que son de inclusión forzosa.
d_h es la cantidad de UPM que se cerraron en el marco actualizado de muestreo y que no son de inclusión forzosa. Notar que cuando la UPM cierra, el director de la misma pierde este título.
d_{Ah} es la cantidad de UPM que se cerraron en el marco actualizado de muestreo y que son de inclusión forzosa.

Peso final del director

El peso final del director es igual al producto del peso base del director por el factor de ajuste por no respuesta asociado al estrato correspondiente:

$$SCHWGT_{hi} = WGTFAC1_{hi} \times WGTADJD1_h .$$

F. Metodología de escalamiento de PLANEA 2015 para la Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN)

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

Introducción

En su modalidad de Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN), el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) evalúa el logro de los aprendizajes de los estudiantes, permite evaluar a un conjunto amplio de contenidos del currículo y se aplica a muestras representativas del país de todos los grados terminales de la educación obligatoria. Sus resultados sirven para monitorear al Sistema Educativo Nacional (SEN) y, en conjunto, aportan información a los tomadores de decisiones de política educativa.

El propósito de PLANEA-ELSEN no es obtener estimaciones precisas de logro (habilidad) de los estudiantes a nivel individual, sino que se pretende obtener estimaciones precisas de grupos o poblaciones de estudiantes. En este caso, se pretende monitorear el progreso de las poblaciones y con ese fin es necesario calcular varios estadísticos sobre la variable de logro, como promedios, desviaciones estándar, porcentajes, percentiles, etc. Para dicho fin, es más adecuado construir las escalas de calificación mediante la metodología de los valores plausibles (Mislevy, 1991). Esta metodología está diseñada para perseguir el propósito de la prueba y es ampliamente utilizada internacionalmente en pruebas de rendimiento como PISA (*Programme for International Student Assessment*), TIMSS (*Trends in International Mathematics and Science Study*) y PIRLS (*Progress in International Reading Literacy Study*), entre otras.

Mislevy (1991) señala que, al utilizar métodos tradicionales de escalamiento en los que primero se obtienen estimaciones individuales de logro y posteriormente, con éstas, se calculan estadísticas de grupo, se pueden generar resultados sesgados.

Además, PLANEA-ELSEN evalúa un conjunto amplio de contenidos del currículo, implementando un diseño matricial en el que cada alumno evaluado solamente contesta un subconjunto del total de reactivos que integran la evaluación.

El logro de los estudiantes es una variable que no se puede observar directamente, por lo que se mide mediante las respuestas a los reactivos que cada estudiante resuelve. Mislevy (1991) también advierte que los procedimientos estándar para hacer inferencias, provenientes de muestreos complejos, no aplican cuando la variable de interés no puede ser observada directamente. Ésta es otra razón por la que se adopta la metodología de los valores plausibles para la construcción de las escalas de calificación de PLANEA-ELSEN.

Cabe señalar que las puntuaciones derivadas de esta metodología tienen algunas propiedades paradójicas, puesto que no sirven como puntuaciones individuales y tampoco para el diagnóstico de los sujetos, sino que únicamente se pueden utilizar para la estimación precisa de parámetros de la población.

A continuación, se describe la implementación de la metodología de escalamiento con valores plausibles para PLANEA-ELSEN.

Introducción a los valores plausibles

La metodología de escalamiento con valores plausibles se combina con el escalamiento estándar con Teoría de Respuesta al Ítem, el cual supone generalmente que la habilidad es una cantidad fija. En la metodología de estimación con valores plausibles es diferente, pues se asume que la habilidad o logro de cualquier estudiante puede describirse mediante una distribución de probabilidad, de esta forma los valores plausibles se definen como posibles valores de habilidad extraídos aleatoriamente de su distribución empírica. Dicha distribución tiene información proveniente de los reactivos que respondió el estudiante y del cuestionario de contexto.

Usualmente, es suficiente hacer cinco extracciones de cada uno de los estudiantes para obtener estimaciones fiables de los parámetros de la población.

Para describir de manera más formal la metodología, es necesario introducir un poco más de notación. Supóngase que Y representa las respuestas del cuestionario de contexto de todos los estudiantes que fueron examinados en la muestra, también suponga de la misma forma que X representa las respuestas de los estudiantes a los reactivos para medir logro y finalmente suponga que θ representa los valores de la variable latente.

El principio con el que funcionan estos valores aleatorios para estimar características de la población es el siguiente:

Supóngase que los valores de θ fueran observables directamente en la muestra, entonces sería posible calcular el estadístico $Q = Q(\theta, X, Y)$ que puede ser una media, un percentil, un coeficiente de regresión o cualquier otro estadístico que depende de las respuestas de contexto o de las respuestas de los estudiantes a los reactivos; o inclusive de los valores de la variable latente. En tal caso, se pueden utilizar métodos estándares para hacer estadística.

Sin embargo, los valores de θ no pueden observarse directamente en ninguna muestra. Por lo que se propone que se asuma que los valores de θ están perdidos de manera completamente aleatoria (*Missing completely at Random*). Este supuesto hace posible la utilización de métodos de imputación múltiple para estimar Q.

Bajo esta perspectiva, Rubin (1977) afirma que el estadístico $Q(\theta,X,Y)$ puede ser aproximado por el valor esperado del estadístico condicionado a que se observan los valores de X y Y; es decir,

$$Q(\theta,X,Y) \approx E(Q|X,Y)$$

Siguiendo la definición general de valor esperado se tiene que:

$$E(Q|X,Y) = \int Q(\theta,X,Y)p(\theta|X,Y)d\theta \ldots (1)$$

En esta definición, los valores X y Y no tienen ningún problema de ser observados en cualquier muestra y se deduce que, para calcular cualquier estadístico que depende de variables latentes, se necesita extraer valores aleatorios de la distribución $p(\theta|X,Y)$, después se necesita evaluar $Q(\theta,X,Y)$ en dichos valores y finalmente se deben promediar los estadísticos evaluados en los valores aleatorios para obtener una sola estimación del estadístico de interés.

El procedimiento de estimación con valores plausibles se puede describir con cuatro etapas: el análisis preliminar del comportamiento de los reactivos que conforman la prueba, la calibración de los reactivos, el condicionamiento con la estimación de parámetros poblacionales y, finalmente, la extracción de valores plausibles. A continuación, se describirá cada una de estas etapas.

Análisis preliminar al escalamiento

Una vez que las pruebas fueron administradas a los alumnos muestreados y sus respuestas fueron integradas a archivos electrónicos, fue necesario llevar a cabo un análisis de los reactivos con la finalidad de verificar el comportamiento de cada una de las preguntas que se incluyeron en los análisis definitivos. Aunque las preguntas ya fueron puestas a prueba con anterioridad en la fase de piloteo, se repitió el procedimiento debido a los ajustes realizados en las propias preguntas.

Principalmente, el análisis se dividió en dos partes:

- Análisis con un modelo de Teoría de Respuesta al Ítem.
- Detección de funcionamiento diferencial de reactivos.

El modelo aplicado en PLANEA-ELSEN es una forma generalizada del modelo logístico simple de Rasch, este modelo se conoce como modelo *logit* multinomial con coeficientes aleatorios, descrito por Adams, Wilson y Wang (1997), e implementado mediante el *software* ConQuest 2 (Adams y Wu, 2002). En este modelo los reactivos se pueden describir por un conjunto de parámetros fijos, desconocidos y se asume que la habilidad de los sustentantes tiene una distribución de probabilidad, lo que será descrito más adelante en el modelo de la población.

El modelo logístico simple predice la probabilidad de responder correctamente un reactivo, basado en la habilidad del sustentante i y la dificultad del reactivo δ_j. Es decir,

$$p(x_{ij}\left|\theta_i,\delta_j\right.)=\frac{exp\left[x_{ij}\left(\theta_i-\delta_j\right)\right]}{1+exp\left(\theta_i-\delta_j\right)}\ldots(2)$$

Donde:

x_{ij} es la respuesta del estudiante i al reactivo j, 1 si es correcta y 0 si es incorrecta;
θ_i es el nivel de logro del estudiante i en la escala *logit*;
δ_j es el parámetro de localización del reactivo j, caracterizando la dificultad del reactivo.
La indeterminación de los parámetros del modelo se resuelve mediante la imposición de la siguiente restricción:

$$\sum_{j=1}^{n}\delta_j=0$$

En donde n es el total de reactivos de la prueba.

Análisis con el modelo logístico de Rasch

En PLANEA-ELSEN se evaluaron dos asignaturas:

- Lenguaje y Comunicación.
- Matemáticas.

En sexto de primaria y tercero de secundaria, para cada una de las asignaturas se ajustó el modelo logístico de Rasch, de manera totalmente independiente. Una vez hecho esto, el análisis consistió en revisar principalmente tres indicadores:

- Correlación punto biserial (r_{pb}).
- Ajuste al modelo de Rasch $(infit)$.
- Curva característica del reactivo.

Los reactivos fueron clasificados con base en estos indicadores de la siguiente forma:

Clasificación de los reactivos según su poder de discriminación de acuerdo con la correlación punto biserial de la respuesta correcta

Rango de discriminación	Clasificación
$r_{pb}\geq 0.15$	El reactivo discrimina de forma aceptable.
$r_{pb}<0.15$	El reactivo debe ser eliminado de la prueba.

Clasificación de los reactivos según el ajuste del modelo logístico de Rasch

Rango de discriminación	Clasificación
$0.8 \leq infit \leq 1.2$	El reactivo ajusta de forma aceptable.
$infit < .8 \circ 1.2 < infit < 2$	El reactivo desajusta moderadamente.
$infit \geq 2$	El reactivo debe ser eliminado de la prueba.

Clasificación de los reactivos en cuanto a la curva característica del reactivo

Rango de discriminación	Clasificación
La curva característica formada con los datos observados está cerca de la curva característica esperada.	El reactivo ajusta de forma aceptable.
La curva característica formada con los datos observados no se parece a la curva característica esperada.	El reactivo debe ser eliminado de la prueba.

Detección de funcionamiento diferencial de los reactivos mediante el modelo logístico de Rasch

El funcionamiento diferencial de reactivos (DIF) ocurre cuando diferentes grupos de estudiantes que responden un reactivo y tienen el mismo nivel de habilidad, no tienen la misma probabilidad de contestar correctamente, es decir, dicho reactivo está funcionando de manera distinta según el contexto y el grupo al que se aplique, después de mantener fijo el nivel de habilidad.

Es común que estos grupos sean determinados por variables como el género, características étnicas, culturales, geográficas, etc. Históricamente, se construyen dos grupos de comparación, el grupo de referencia, que consiste en los individuos para los cuales se espera que la prueba aplicada les favorezca; y el grupo focal, que consiste en los individuos para los cuales se espera que la prueba aplicada los ponga en desventaja.

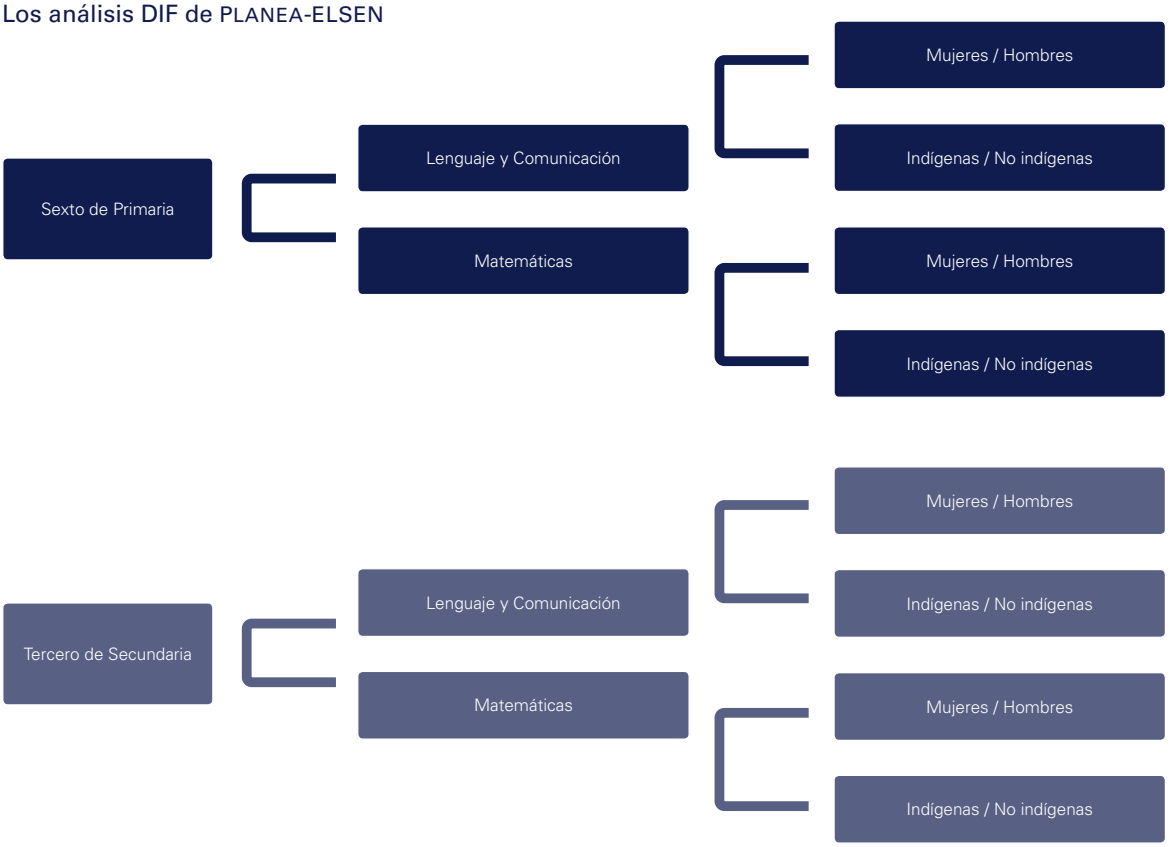
Si los datos ajustan razonablemente al modelo logístico simple de Rasch, se puede demostrar que la comparación de las probabilidades de acierto, para cada reactivo en los grupos de interés bajo la hipótesis de no DIF, es equivalente a la comparación de los parámetros de dificultad calculados independientemente en cada uno de los grupos a contrastar.

El método implementado para analizar el DIF compara dos grupos mediante la estimación de la dificultad de los reactivos para cada uno de ellos, es decir, se estiman $\widehat{\delta}_{j,F}, \widehat{\delta}_{j_R}$, y sus respectivos errores estándar, donde se considera al grupo focal (F) y al grupo de referencia (R) (Wright y Stone, 1979; Wright, Mead y Draba, 1976; Wright y Masters, 1982).

Para la prueba PLANEA-ELSEN 2015, sexto de primaria y tercero de secundaria, el primer análisis DIF que se realizó fue para los grupos de estudio definidos por la variable sexo, donde el grupo focal o minoritario fue la subpoblación de mujeres y el grupo de referencia fue la subpoblación de hombres. El segundo análisis DIF se realizó para las subpoblaciones indígena (grupo focal) y no indígena (grupo de referencia), tanto para la asignatura de Lenguaje y Comunicación como para Matemáticas.

Es importante señalar que el tamaño de las subpoblaciones se relaciona directamente con la potencia del análisis DIF de la prueba. Por lo que, previo a éste, se verificó que el tamaño de las subpoblaciones, tanto focal como de referencia, sea al menos de tamaño 500 sustentantes (Guilleux, Blanchin, Hardouin y Sébille, 2014).

Los análisis DIF de PLANEA-ELSEN



El método de Rasch para detectar DIF consistió en lo siguiente:

Se estimaron los parámetros de dificultad de manera aislada en cada uno de los grupos de interés. La diferencia media en logro entre ambos grupos se controló fijando la dificultad promedio de los reactivos en cero:

$$\sum_{j=1}^{n} \widehat{\delta}_{j,R} = \sum_{j=1}^{n} \widehat{\delta}_{j,F} = 0$$

Se calculó la diferencia en la escala *logit* entre las dificultades de reactivos estimadas para ambos grupos, como una medida para determinar el funcionamiento diferencial de reactivos:

$$\widehat{\Delta} j = \widehat{\delta}_{j,R} - \widehat{\delta}_{j,F}$$

Se calculó el índice estandarizado que está dado por:

$$Z_j = \frac{\widehat{\Delta}_j}{\sqrt{(SE\widehat{\delta}_{j,R})^2 + (SE\widehat{\delta}_{j,F})^2}}$$

Bajo la hipótesis de que no se presenta DIF, el valor esperado de $\widehat{\Delta}_j$ es 0, y Z_j tiene media 0 y varianza 1. Por lo cual, se reportan solamente los reactivos que presentan una diferencia significativa, es decir, aquellos en donde: $Z_j \leq -1.96$ o $Z_j \geq 1.96$.

Una vez identificados los reactivos con funcionamiento diferencial significativo, se clasificaron en tres categorías, bajo el siguiente criterio (Wilson, 2005, p. 167):

Categoría	Criterio
DIF despreciable "A"	$\left\|\widehat{\delta}_{j,R} - \widehat{\delta}_{j,F}\right\| < 0.426$
DIF intermedio "B"	$0.426 \leq \left\|\widehat{\delta}_{j,R} - \widehat{\delta}_{j,F}\right\| < 0.638$
DIF grande "C"	$\left\|\widehat{\delta}_{j,R} - \widehat{\delta}_{j,F}\right\| \geq 0.638$

En los casos en que la diferencia es significativa, es decir, se presenta DIF, el grupo focal de cada análisis se ve favorecido si $\widehat{\Delta}_j$ y Z_j son valores positivos.

Eliminación de reactivos

Una vez clasificados los reactivos de acuerdo con los problemas que presentaron de discriminación, ajuste, forma de la curva característica y funcionamiento diferencial, la Dirección General de Evaluación de Resultados Educativos tomó la decisión sobre qué reactivos se eliminan. Las cantidades se presentan a continuación:

Asignatura	Cantidad de reactivos eliminados	Cantidad de reactivos utilizados
Lenguaje y Comunicación sexto de primaria	7	143
Lenguaje y Comunicación tercero de secundaria	7	143
Matemáticas sexto de primaria	3	147
Matemáticas tercero de secundaria	6	144

Modelo condicional de respuesta al ítem

De los supuestos de los modelos IRT, el más importante es la independencia condicional. Bajo este supuesto, las probabilidades de los reactivos dependen solamente de θ y de su dificultad, y no son afectadas por otras variables como las características sociodemográficas. Denotando estas variables para cada uno de los estudiantes con y.

Bajo este supuesto y utilizando (2), la probabilidad conjunta de un patrón de respuesta $x' = (x_1,\ldots,x_n)$, a través de los n reactivos, está dada por:

$$p(x|\theta,\delta,y) = p(x|\theta,\delta) = \prod_{j=1}^{n} p(xj|\theta,\delta j) \quad \ldots(3)$$

Donde:

$\delta' = (\delta_1,\ldots,\delta_n)$ es el vector de parámetros de dificultad de los n reactivos.

Los parámetros de dificultad se estiman en la calibración y, posterior a ese paso, se consideran valores fijos.

La respuesta de cualquier subconjunto de reactivos induce una función de verosimilitud para θ y si, adicionalmente, suponemos que existe independencia entre los sujetos, utilizando (3), obtenemos "el modelo condicional de respuesta al ítem" o "el modelo de la variable latente".

$$p(X|\theta,\delta,Y) = p(X|\theta,\delta) = \prod_{i=1}^{N} p(x_i|\theta_i,\delta) \quad \ldots(4)$$

Donde:

N es la cantidad total de estudiantes en la muestra.

Calibración de los reactivos

Una vez depuradas las claves de respuesta, se eliminaron los reactivos con grandes deficiencias. La fase de calibración de los instrumentos consistió en la estimación de los parámetros de dificultad de los reactivos para el modelo logístico simple de Rasch.

Para la calibración, el modelo condicional de respuesta al ítem es usado en conjunción del modelo de la población, pero sin utilizar variables de condicionamiento.

El método utilizado para hacer la estimación de los parámetros de los reactivos fue el de máxima verosimilitud marginal (Bock y Aitkin, 1981), en el cual se asume independencia condicional entre las respuestas a los reactivos y que la habilidad es un efecto aleatorio. De esta forma, es posible eliminar los parámetros de las personas del proceso de estimación integrando sobre la habilidad, lo cual se reduce a maximizar el logaritmo de la verosimilitud que surge de:

$$p(X|\alpha,\delta)=\prod_{i=1}^{N}\int p(xi|\theta,\delta)p(\theta|\alpha)d\theta$$

La calibración de cada una de las escalas se realizó independientemente de los parámetros de las otras, utilizando la máxima cantidad de información disponible, es decir, con toda la muestra de los estudiantes.

Análisis multidimensional de Rasch

Después del paso de calibración, se aprovechó que el modelo de la población hace posible ajustar un modelo bidimensional para las habilidades, en el que se consideran relacionadas las habilidades de un estudiante en Lenguaje y Comunicación con Matemáticas. En la siguiente sección se explica el modelo de la población. Del ajuste del modelo bidimensional, se deriva que la correlación de los puntajes entre las dimensiones fue el siguiente:

Correlación	Matemáticas sexto de primaria	Matemáticas tercero de secundaria
Lenguaje y Comunicación sexto de primaria	0.837	
Lenguaje y Comunicación de tercero de secundaria		0.771

En ambas asignaturas la confiabilidad de las pruebas aumentó

Asignatura	Confiabilidad modelo unidimensional (EAP/PV RELIABILITY)	Confiabilidad modelo bidimensional (EAP/PV RELIABILITY)
Lenguaje y Comunicación sexto de primaria	0.833	0.899
Lenguaje y Comunicación tercero de secundaria	0.798	0.833
Matemáticas sexto de primaria	0.858	0.911
Matemáticas tercero de secundaria	0.790	0.825

En los cuadros anteriores se observó que los puntajes de las asignaturas tienen una considerable asociación lineal y que la confiabilidad de las pruebas se incrementa con respecto a los modelos unidimensionales. Por lo tanto, se tomó la decisión de utilizar un modelo bidimensional para generar los puntajes.

Modelo de la población

Para continuar con el proceso de escalamiento, es necesario suponer que existe una distribución de probabilidad asociada a la variable latente y que tiene una forma funcional conocida.

Mislevy (1991) supone que la distribución de la variable latente θ, dadas las variables sociodemográficas y, tiene una forma funcional conocida que denotaremos con $p(\theta|y,z,\alpha)$, que está caracterizada por un conjunto de parámetros desconocidos α. La práctica más común es asumir que las habilidades de los estudiantes provienen de una distribución normal multivariada.

Adams, Wilson y Wu (1997) discutieron que es muy natural reemplazar la media de la distribución normal multivariada con un modelo de regresión $\Gamma' y_i$, donde y_i es un vector de valores fijos del i-ésimo estudiante.

Entonces, el modelo poblacional del estudiante i será:

$$\theta_i = \Gamma' y_i + \in$$

Donde $\in$ se distribuye normal multivariada con media 0 y matriz de varianza y covarianza Σ entre las dimensiones. De esta forma, los parámetros a estimar están en $\alpha = (\Gamma, \Sigma)$. En este contexto las y_i son conocidas como variables de condicionamiento, que resultan indispensables para la estimación final de las habilidades.

El modelo de la población se define asumiendo que existe independencia entre los sustentantes:

$$p(\theta|Y,\alpha,\delta) = \prod_{i=1}^{N} p(\theta_i|y_i,\alpha,\delta) \quad \ldots(5)$$

En el modelo de la población se estimaron los coeficientes de regresión y la matriz de varianza y covarianza, en la fase del Condicionamiento y estimación de parámetros.

Condicionamiento y estimación de parámetros poblacionales

Las variables de condicionamiento tienen como meta maximizar la información posible para estimar las habilidades. Se construyeron considerando dos grupos, el primero, lo constituyen las variables que están directamente en el modelo de regresión:

Variables que definen subpoblaciones de interés:

- Dominios primarios de PLANEA-ELSEN. Las entidades federativas.
- Dominios secundarios de PLANEA-ELSEN. Tipos de escuelas que consideran marginación y sostenimiento.
- Sexo.
- Edad.
- Grado de multigrado (sólo en primaria).

Otras variables que influyen sobre la habilidad de los estudiantes:

- La forma del examen que se le asignó.
- Estimación preliminar del promedio del rendimiento de cada estudiante en ambas asignaturas dentro de cada escuela. Es el promedio de los estimadores de máxima verosimilitud de ambas asignaturas para todos los demás estudiantes. Incluir las medias de rendimiento por escuela sirve para tomar en cuenta gran parte de la variación existente entre escuelas.

El otro grupo de variables de condicionamiento lo constituyen las variables que indirectamente influyen sobre la habilidad de los estudiantes. Dichas variables se construyen a partir de los cuestionarios de contexto de la siguiente forma:

1. Cada una de las variables de los cuestionarios de contexto A y B fueron recodificadas mediante el método "*deviance coding*". Es decir, se construyen variables indicadoras para cada una de las categorías de respuesta de cada una de las variables de los cuestionarios de contexto, a continuación, se elige la categoría de referencia eligiendo la de mayor frecuencia estimada en la población, los valores de esa categoría son recodificados a -1.

2. Después se elaboró un análisis de componentes principales para reducir la dimensionalidad. Se retuvieron las cargas factoriales que expliquen 95% de la varianza total de ambos cuestionarios de contexto. En general, la cantidad de factores que se construyen tiene que ser de al menos 200. Para primaria se construyeron 240 y para secundaria 209.

Una vez construidos ambos grupos de variables de condicionamiento, se estimaron los coeficientes de regresión, en donde se obtuvieron dos grupos de coeficientes, uno para Lenguaje y Comunicación y otro para Matemáticas. Los coeficientes de regresión sirven para definir las medias de cada una de las subpoblaciones de interés.

Para terminar la estimación de los parámetros poblacionales, a la par de la estimación de los coeficientes de regresión, se estimó la matriz de varianza y covarianza fijando los parámetros de dificultad de los reactivos.

Extracción de valores plausibles

La distribución predictora necesaria para calcular (1), se puede aproximar numéricamente, una vez que fueron estimados los parámetros de dificultad de los reactivos, los coeficientes de regresión y la matriz de varianza y covarianza, aplicando el teorema de Bayes a θ, X, Y y utilizando (4), (5) y (6).

$$p(\theta X,Y,\alpha,\delta) = K_{\alpha\delta} p(X \mid \theta,Y,\alpha,\delta) p(\theta \mid Y,\alpha,\delta) = \prod_{i=1}^{N} K_{i\alpha\delta} p(x_i \mid \theta_i,\delta) p(\theta_i \mid y_i,x_i,\alpha)$$

Donde:

$K_{i\alpha\delta} = \dfrac{1}{p(x_i y_i \alpha,\delta)}$ es la constante de normalización que depende de α y δ pero no de θ_i.

De esta distribución se extrajeron cinco vectores de valores plausibles por asignatura.

Cálculo y cancelación del efecto del cuadernillo

El diseño matricial de las pruebas PLANEA-ELSEN 2015 fue constituido por bloques incompletos balanceados, de modo que las estimaciones de los parámetros de los reactivos y de las habilidades de los estudiantes que se obtienen en principio, no están influenciados por el efecto del cuadernillo. Sin embargo, en la práctica es común encontrar que aún hay influencia del cuadernillo debido a que los reactivos aparecen en diferentes posiciones dentro los cuadernillos.

Modelar el efecto de orden, en términos de la posición de los reactivos dentro de un cuadernillo, resultaría en un modelo muy complicado, por lo que se optó por usar el modelo implementado en los Exámenes de la Calidad y el Logro Educativos (EXCALE), el cual era muy sencillo y consistía en estimar el efecto de cada cuadernillo en la calibración, de la misma manera que si se estimara la dificultad del cuadernillo.

Es decir, en la estimación de los parámetros de los reactivos, los efectos de los cuadernillos fueron incluidos en el modelo de medición para prevenir que se confundieran las dificultades de los reactivos y los efectos de los cuadernillos. El modelo empleado es un modelo de los llamados de facetas y es el siguiente:

$$logit(p) = \theta - \delta_j - \tau_k$$

El parámetro del cuadernillo, τ_k, se definió en el mismo sentido de los parámetros de dificultad de los reactivos, reflejando la dificultad de los cuadernillos.

Posterior a la estimación de las habilidades de los estudiantes con valores plausibles, los parámetros estimados de los cuadernillos fueron sumados a las estimaciones de habilidad de los estudiantes, con la finalidad de cancelar el efecto de cada uno de los cuadernillos.

Análisis de datos con valores plausibles

Los valores plausibles pueden ser empleados para evaluar la ecuación (1), para cualquier estadístico Q, de la siguiente forma:

Usando cada uno de los vectores de valores plausibles se evalúa Q, como si los valores plausibles fueran los verdaderos valores de θ, cada una de estas evaluaciones se denotan con Q_m. Para $m = 1,\ldots,M$ con $M = 5$.

Se estima la varianza de cada una de las evaluaciones de Q_m, cada una de estas estimaciones se denotan con U_m para $m = 1,\ldots,M$.

La mejor estimación que se puede obtener de Q, a partir de los valores plausibles, es el promedio simple de los 5 valores obtenidos de los diferentes conjuntos de valores plausibles.

$$\hat{Q}_{pv} = \frac{1}{M}\sum_{m=1}^{M} Q_m$$

La estimación de la varianza total de $\hat{Q}_{pv}$ se hace sumando dos componentes.

Considere:

$$\bar{U} = \frac{1}{M}\sum_{m=1}^{M} U_m$$

$$B_M = \frac{1}{M-1}\sum_{m=1}^{M} (Q_m - \hat{Q}_{pv})^2$$

La estimación de la varianza total de $\hat{Q}_{pv}$ es:

$$\hat{V}(\hat{Q}_{pv}) = \bar{U} + (1 + M^{-1})B_M$$

El primer término contiene la varianza debida al muestreo, y el segundo la varianza debida a que las estimaciones de las habilidades no se estiman con precisión.

Un intervalo de confianza al $(1-\alpha)\%$ para $\hat{Q}_{pv}$ es $\hat{Q}_{pv} \pm t_v\left(\frac{1-\alpha}{2}\right)\sqrt{\hat{V}(\hat{Q}_{pv})}$, donde

t_v es el cuantil de una distribución t con v grados de libertad.

$$v = \left[\frac{f_M^2}{M-1} + \frac{(1-f_M)^2}{d} \right]^{-1}$$

Donde *d* son los grados de libertad para el estadístico calculado con los datos completos.

$$f_M = \frac{(1+M^{-1})B_M}{\widehat{V}(\widehat{Q}_{pv})}$$

Advertencia de posible sesgo en algunos análisis secundarios

Es común que se pretenda investigar asociaciones entre la variable latente de logro y alguna otra variable que no fue considerada en el condicionamiento. Sin embargo, esto no se debe de hacer, puesto que, si alguna variable no fue incluida como variable de condicionamiento, resulta imposible recobrar completamente el posible efecto o relación que tenga con la variable latente y, en consecuencia, las estimaciones que involucren a la variable no incluida en el condicionamiento, pueden resultar sesgadas en un grado desconocido. Este problema se encuentra ampliamente documentado en Mislevy (1991).

Definición de la escala basal

Una vez cancelado el efecto de los cuadernillos en los valores plausibles, es necesario aplicarles una transformación lineal con la finalidad de facilitar la comunicación de los resultados. Los valores plausibles se originan sobre la escala *logit*, así que la transformación lleva la media y la varianza de los valores plausibles a la nueva escala de PLANEA-ELSEN, en la que se definió que la media fuera de 500 unidades y una desviación estándar de 100, para cada una de las asignaturas.

Se utilizó la siguiente transformación:

$$T = \frac{\sigma_1}{\sigma_0}\theta + \left[\mu_1 - \frac{\sigma_1}{\sigma_0}\mu_0 \right]$$

En donde:

θ valor en la escala *logit*.
σ_1 es la desviación estándar de la escala de PLANEA-ELSEN.
μ_1 es la media de la escala de PLANEA-ELSEN.
σ_0 es la desviación estándar en la escala *logit* estimada con valores plausibles. Es decir, $\sigma_0 = \frac{1}{M}\sum \frac{M}{m} =_1 \sigma_m$ es el promedio de las desviaciones estimadas para cada uno de los vectores de valores plausibles considerando el diseño de la muestra y los factores de expansión.

μ_0 es la media en la escala logit estimada con valores plausibles. Es decir, $\mu_0 = \frac{1}{M}\sum \frac{M}{m} =_1 \mu_m$ es el promedio de las medias estimadas para cada uno de los vectores de valores plausibles considerando el diseño de la muestra y los factores de expansión.

Con la definición de la escala base termina el proceso de escalamiento de resultados.

Tratamiento de los datos faltantes en los reactivos de logro

Los reactivos que se utilizaron en PLANEA para medir el logro de los estudiantes en las asignaturas de Lenguaje y Comunicación y Matemáticas fueron de opción múltiple. Para cada una de estas opciones de respuesta se asociaron los números del 1 al 4.

En las evaluaciones de gran escala pueden ocurrir dos situaciones:

1. Los sustentantes respondieron una y sólo una de las posibles opciones de respuesta y se les asignó un puntaje tratándolos como reactivos dicotómicos, es decir, 1 si respondió correctamente y 0 cuando respondió incorrectamente.
2. Los sustentantes dejan de responder por alguna razón o responden más de una de las opciones de respuesta.

La segunda situación nunca es deseada, pues genera respuestas faltantes, sin embargo, es inevitable tener que lidiar con ella. Para PLANEA-ELSEN, estos y otros tipos de respuestas faltantes se consideraron como valores perdidos que recibieron un tratamiento distinto en los diferentes momentos del proceso de construcción de las escalas.

Como primer paso, se presentan las distintas categorías de respuestas faltantes consideradas en PLANEA-ELSEN para los reactivos correspondientes al logro:

Respuesta omitida: ocurrió cuando un examinado tuvo el tiempo y la oportunidad de responder a un reactivo, sin embargo, no lo hizo. Se le asignó el código '9'.
Respuesta múltiple: se consideró cuando un examinado respondió más de una de las opciones de respuesta. Se le asignó el código '8'.
No administrada: las pruebas de logro se aplicaron mediante un esquema de rotación de bloques de reactivos siguiendo un esquema matricial, de tal forma que a cada examinado se le asignó un subconjunto a responder de toda la prueba. A las respuestas de los reactivos que no se les asignó para responder, se les consideró de esta clase y se les asignó el código '7'.
No alcanzada: bajo las mismas condiciones de aplicación y tiempo, algunos examinados no alcanzaron a responder los reactivos del final de la sesión. A estas respuestas faltantes se les asignó el código '5'.

Una vez identificados los distintos tipos de datos faltantes, se les asignó un puntaje diferente en cada una de las etapas de la construcción de las escalas. A continuación, se describe la asignación.

Puntajes para las respuestas faltantes

Para asignar los puntajes de las respuestas faltantes se siguió la estrategia sugerida por Ludlow y O'Leary (1999), en la que el tratamiento de las respuestas faltantes se hace diferenciado en un proceso de dos etapas.

Esta estrategia es la que usan en la actualidad las pruebas TIMSS en su proceso de escalamiento (Adams, Wu y Macaskill, 1997).
En la primera etapa, en la que se hizo la estimación de los parámetros de dificultad de los reactivos bajo el modelo de Rasch (calibración), las respuestas no alcanzadas se trataron como no administradas, por lo que no se les asignó un puntaje.

La razón que describen Ludlow y O'Leary (1999) es la siguiente: si en este punto del proceso de estimación suponemos que las respuestas no alcanzadas fueran tratadas como incorrectas, ocurriría una sobreestimación de la dificultad de los reactivos, ya que los estudiantes podrían contestar a algunas de estas preguntas correctamente, teniendo el tiempo y la oportunidad de responder a ellas.

En la segunda etapa, se consideraron fijos y conocidos los parámetros de dificultad de los reactivos y se hizo la estimación de las habilidades de los examinados. En esta etapa, las respuestas no alcanzadas se consideraron como respuestas incorrectas, por lo que se les asignó un puntaje de '0'.

De acuerdo con Ludlow y O'Leary (1999), si suponemos que las respuestas no alcanzadas son tratadas como no administradas, en la fase de estimación de las habilidades, los estudiantes podrían adoptar la estrategia de ir respondiendo desde el inicio únicamente los reactivos en los que está seguro de conocer la respuesta correcta y detenerse en el momento en el que ya no se sienta seguro y dejar el resto de la prueba en blanco. Como las respuestas no alcanzadas no serían tomadas en cuenta para la estimación de la habilidad, una gran parte de ellos podría obtener muy buenos resultados en la prueba.

En ambas etapas, las respuestas omitidas y múltiples se consideraron como respuestas incorrectas, por lo que se les asignó un puntaje de '0'.

El mecanismo de administración de los reactivos, el cual se hizo mediante la rotación de los bloques bajo un diseño matricial, hace posible ignorar a las respuestas no administradas, por lo que no se les asignó un puntaje, se les consideró "*missing by design*" (Boomsma, van Duijn y Snijders, 2001).

G. Descripción del cálculo de estimaciones PLANEA 2015

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

Este anexo tiene como propósito describir los procedimientos con los que se estimaron los resultados de logro del Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) que se reportan en la página web del Instituto Nacional para la Evaluación de la Educación (INEE). Para calcular los resultados se tomó como base la metodología propuesta por la *Organisation for Economic Co-operation and Development* (OECD) aplicada a la prueba *Programme for International Student Assessment* (PISA). Describiendo así los datos, la metodología de estimación y la formación de las tablas de logro y contexto.

Datos

La Evaluación del Logro referida al Sistema Educativo Nacional (ELSEN) cuenta con las bases de datos que corresponden a las evaluaciones de los alumnos de sexto grado de primaria y tercero de secundaria aplicadas al finalizar el ciclo escolar 2014-2015 para dos asignaturas: Lenguaje y Comunicación y Matemáticas; así como a los cuestionarios de contexto que fueron aplicados a los alumnos, titulares de grupo y directores. Los datos tienen diferente información contenida en variables.

La mayor parte de las variables de alumnos contienen información referida a las respuestas dadas por los alumnos a las evaluaciones de logro y a los cuestionarios de contexto A y B de PLANEA 2015. Las variables de logro contienen valores ya calificados de las preguntas aplicadas en cada asignatura, donde uno es para la respuesta correcta y cero para indicar una respuesta incorrecta. En el caso de los datos de titulares de grupo y directores se tiene información de las respuestas al respectivo cuestionario de contexto. En general éstos contienen respuestas categóricas.

Otro tipo de información respecto a alumnos, titulares de grupo y directores es la que se presenta en la tabla G1, estas variables identifican grupos o segmentos de población dentro de los cuales se realizaron estimaciones. En la sección 4 se explicará cómo estas variables fueron utilizadas para obtener las estimaciones.

Tabla G1 Nombre de las variables que se utilizaron en el reporte de PLANEA 2015 como variables de segmentación para las estimaciones

Nombre de la variable	Variable	Categorías	
		Primaria	Secundaria
Entidad Federativa	ID_ENT	32 entidades	32 entitades
Tipo de escuela	SERV	General pública Indígena Comunitaria Privada	General pública Técnica pública Telesecundaria Comunitaria Privada
Sexo	SEXO	Mujer Hombre	Mujer Hombre
Edad normativa	EDAD_N	Edad normativa Extraedad	Edad normativa Extraedad
Edad en años cumplidos colapsada	EDAD_ACC	11 años o menos 12 años 13 años 14 años o más	14 años o menos 15 años 16 años 17 años o más
Marginación	MARGINC	Muy alta y Alta Media Baja y Muy baja	Muy alta y Alta Media Baja y Muy baja
Grupo multigrado	I_MULTIGRADO	No Sí	No Sí
Rural-Urbano	RURALIDAD	Rural Urbano	Rural Urbano
Tamaño de localidad	TAM_LOC_PRIM TAM_LOC_SEC	1 a 499 hab. 500 a 2 499 hab. 2 500 a 99 999 hab. 100 000 o más hab.	1 a 2 499 hab. 2 500 a 99 999 hab. 100 000 o más hab.

Respecto a los datos provenientes del diseño muestral, los pesos del muestreo están construidos para cada población: W_FSTUWT para alumnos, W_FGRPWT para titulares de grupo y W_FSCHWT para directores. En el caso de los pesos de replicación, se construyeron 100 réplicas por cada variable de peso muestral mediante el método de remuestreo BRR (*Balanced Repeated Replication*). En los datos se identificaron con la variable W_FSTR para alumnos, W_FGRR para titulares de grupo y W_FSCR para directores seguida de los dígitos que numeran cada réplica.

Los valores plausibles son cinco variables por cada asignatura provenientes del escalamiento de las respuestas de logro. Esto es, se cuenta con cinco diferentes valores plausibles en la muestra para realizar la estimación de los estadísticos. La metodología de estimación se presenta en la sección 2.2.

Otras variables son las de niveles de logro provenientes de las clasificaciones hechas a los valores plausibles (PV) en cuatro niveles especificados por la Dirección General de Evaluación de Resultados Educativos. Para ello, hay una variable de nivel de logro asociada a cada variable de valores plausibles que comprende la clasificación de éstos según el nivel del logro que corresponda al puntaje: 1: nivel I, 2: nivel II, 3: nivel III y 4: nivel IV.

Otras dos clasificaciones asociadas a logro son: los alumnos que alcanzan al menos el nivel II de logro y los alumnos que alcanzan al menos el nivel III de logro. Las cuales son variables dicotómicas donde 0: no alcanza el nivel y 1: alcanza el nivel. Mismas que se pueden definir como la proporción de alumnos que alcanzan al menos el nivel II o el nivel III; o bien, como el porcentaje de alumnos que alcanza al menos el nivel II o nivel III (considerando sólo el valor 1 para su cálculo). Las variables involucradas en logro se muestran en la tabla G2.

Tabla G2. Variables asociadas al logro

Variable	Raíz de la variable		Categorías
	Lenguaje y Comunicación	Matemáticas	
Logro	LYCPV	MATPV	
Niveles de logro	LYCNVL	MATNVL	Nivel I Nivel II Nivel III Nivel IV
Alumnos que alcanzan al menos el nivel II	LYCNVLII	MATNVLII	Nivel I Al menos el nivel II
Alumnos que alcanzan al menos el nivel III	LYCNVLIII	MATNVLIII	Niveles I y II Al menos el nivel III

Estimación de descriptivos y su error estándar

Debido a la obtención de la población objetivo por muestreo, es necesario considerar la estimación del error de muestreo, lo cual se lleva a cabo mediante remuestreo. Por otro lado, la construcción de la variable de estudio consiste de la variable latente de logro obtenida mediante el escalamiento de los valores plausibles; es necesario describir la metodología que se requiere para llevar a cabo estimaciones con valores plausibles. Ambas metodologías se explican a continuación, además de la combinación de éstas.

Método BRR-FAY para estimar el error debido al muestreo

En un diseño muestral existe un gran número de posibles muestras. Si se calculara un estimador de algún estadístico en alguna de estas muestras, se esperaría obtener el mismo valor de estimación dado un intervalo de confianza. Cada estimación del estadístico tiene asociado un riesgo de error o incertidumbre muestral, la varianza muestral corresponde a la medida de dicha incertidumbre. Un método para calcular la varianza debida al muestreo es mediante el uso de la replicación repetida balanceada (BRR, *Balanced Repeated Replication*), la cual es una alternativa de estimación cuando el estimador no es una función lineal o, debido al diseño muestral, no existe o no se tiene un estimador insesgado que defina la varianza muestral.

En PLANEA 2015, el diseño proporcional al tamaño (PPT) se utilizó para elegir la primera etapa de muestreo, en este diseño muestral no existe un estimador para la varianza, por lo cual, fue necesario utilizar otras alternativas. En este caso, es la metodología de remuestreo BRR con factor de corrección Fay (Judkins, 1990), con constante $k = 0.5$ y con 100 pesos de replicación.

De modo que usando los BRR la estimación : sea $\widehat{Q}$ el estadístico que se desea conocer: (media, moda, varianza), π^i (porcentaje de la categoría $_i$), diferencia de la estimación entre subpoblaciones, etc.

Por cada peso de replicación se calcula un estimador, sean ($(\widehat{Q}_1, \widehat{Q}_2, \ldots, \widehat{Q}_{99}, \widehat{Q}_{100})$) los 100 estimadores. El estimador final es $\hat{Q} = \frac{1}{100}\sum_{r=1}^{100}\widehat{Q}_r$, donde $\widehat{Q}$ es el estimador deseado y $\widehat{Q}_r$ es la estimación proveniente del $r-ésimo$ peso de replicación. La varianza es $\hat{\sigma}^2{}_{(muestreo)} = \frac{1}{100(1-k)^2}\sum_{r=1}^{100}(\widehat{Q} - \widehat{Q}_r)^2$.

Metodología para estimar con valores plausibles

Los valores plausibles se pueden describir como los posibles valores que puede tomar una variable latente extraída bajo un modelo de probabilidad. En el caso de PLANEA 2015, se calcularon cinco valores plausibles para la variable de logro en ambas asignaturas, es decir, por cada alumno en la muestra se calcularon cinco posibles valores del puntaje de logro en cada asignatura a partir de la distribución de logro a la que pertenece el alumno. De este modo, los valores plausibles permiten obtener estimaciones eficientes de las estadísticas poblacionales.

Para estimar con valores plausibles, se calculó la media de los cinco estadísticos derivados de cada valor plausible. Es decir, sea Q el estadístico que se desea conocer, entonces el estadístico estimado es $\hat{Q} = \frac{1}{M}\sum_{m=1}^{M}\widehat{Q}^{(m)}$, donde $\widehat{Q}^{(m)}$ denota la estimación del estadístico calculado con el valor plausible $m = 1,2,\ldots,M$, donde M es el número de valores plausibles, en este caso $M = 5$.

La varianza debida a la imputación se calculó como $\hat{\sigma}^2(imp) = \frac{1}{M-1}\sum_{m=1}^{M}(\widehat{Q}^{(m)} - Q)^2$, donde $\widehat{Q}^{(m)}$ es el estadístico del valor plausible $_m$ y $\widehat{Q}$ es el estimador final. Por definición, el error de imputación debe ser lo más pequeño posible y se calcula como la raíz cuadrada de la varianza de imputación.

Metodología de estadísticos con valores plausibles y remuestreo

De acuerdo con el PISA *Data Analysis Manual* (OECD, 2009b, pp. 120-121) el cálculo de un estadístico con cinco valores plausibles $(M = 5)$ se lleva a cabo en seis pasos:

1. El estadístico y su error estándar se calculan para cada valor plausible con el peso muestral y cada uno de los pesos de replicación, es decir, se obtienen 101 estimadores para cada valor plausible. Así, se obtienen los estadísticos $\widehat{Q}^{(1)}, \widehat{Q}^{(2)}, \widehat{Q}^{(3)}, \widehat{Q}^{(4)}, \widehat{Q}^{(5)}$

que son estimados con el peso muestral. De los 100 pesos de replicación con los que se estima cada valor plausible, se estiman cinco varianzas muestrales denotadas por $\widehat{\sigma^2_{(Q^{(1)})}}, \widehat{\sigma^2_{(Q^{(2)})}}, \widehat{\sigma^2_{(Q^{(3)})}}, \widehat{\sigma^2_{(Q^{(4)})}}, \widehat{\sigma^2_{(Q^{(5)})}}$. Cada una de estas varianzas se calculan como se indica en la sección 2.1.

2. El estadístico estimado finalmente es igual al promedio de los cinco estimadores, es decir $\hat{Q} = \frac{1}{5}(\widehat{Q^{(1)}} + \widehat{Q^{(2)}} + \widehat{Q^{(3)}} + \widehat{Q^{(4)}} + \widehat{Q^{(5)}})$.

3. La varianza muestral final es igual al promedio de las cinco varianzas muestrales, como sigue $\hat{\sigma^2_{(Q)}} = \frac{1}{5}(\widehat{\sigma^2_{(Q^{(1)})}} + \widehat{\sigma^2_{(Q^{(2)})}} + \widehat{\sigma^2_{(Q^{(3)})}} + \widehat{\sigma^2_{(Q^{(4)})}} + \widehat{\sigma^2_{(Q^{(5)})}})$.

4. La varianza de imputación, también llamada varianza de error de medida, se calcula como $\widehat{\sigma^2_{(imp)}} = \frac{1}{4}\sum_{m=1}^{5}(\widehat{Q^{(m)}} - Q)^2$.

5. La varianza final es la combinación de la varianza debida a la imputación y la varianza muestral de la siguiente manera: $\widehat{\sigma^2} = \widehat{\sigma^2_{(\hat{Q})}} + (1 + \frac{1}{5})\widehat{\sigma^2_{(imp)}}$.

6. El error estándar es igual a la raíz cuadrada del error de varianza, es decir, $\hat{\sigma} = \sqrt{\widehat{\sigma^2_{(\hat{Q})}} + (1 + \frac{1}{5})\widehat{\sigma 2_{(imp)}}}$.

Cálculo del error estándar

En esta sección se describen, en forma general, los códigos fuente hechos en el programa SAS® que se utilizaron en la estimación del error estándar para cada uno de los estadísticos calculados asociados al logro en las dos asignaturas: Lenguaje y Comunicación y Matemáticas, de los alumnos de sexto de primaria y tercero de secundaria, así como los utilizados en la estimación de los porcentajes reportados de los cuestionarios de contexto.

El cálculo de las estimaciones reportadas se llevó a cabo mediante las macros publicadas por la OCDE (2009). De las cuales se describen las que se emplearon para calcular las estimaciones de estadísticos univariados, frecuencias y diferencias entre subpoblaciones.

Para la estimación de estadísticos univariados y de frecuencias se describen dos tipos de macros, una utilizando la metodología de valores plausibles y otra para variables no imputadas. En este último caso, en donde no se utilizan valores plausibles como lo es el análisis de las respuestas a los cuestionarios de contexto. La forma de estimar el error estándar es como se describe en la sección 2.1 y para estimar usando los valores plausibles se siguen los seis pasos descritos en la sección 2.3.

Macro-función en SAS

Una macro-función en SAS definida por el usuario es una función o procedimiento donde se obtiene un resultado. Esta función puede recibir parámetros definidos en el encabezado y devolver valores o archivos con los resultados obtenidos. De este modo, para hacer

uso de las macro-funciones de PISA sólo es necesario definir los parámetros solicitados para que se realice la estimación requerida.

Debido a que se utilizaron 100 réplicas bajo el método de remuestreo BRR-FAY para calcular el error estándar de muestreo, fue necesario modificar cada macro de PISA considerando el nuevo número de réplicas ya que este cambio repercute en el promedio de las 100 réplicas. Esto se muestra en la figura G1.

Figura G1. Modificaciones en la macro-función de SAS según el número de réplicas

En la estimación de error de muestreo por BRR-FAY, el término $\frac{1}{G(1-K)^2}$ es el que se modifica	$\widehat{\sigma^2}_{(muestreo)} = \frac{1}{G(1-k)^2}\sum_{r=1}^{100}(\widehat{Q}-\widehat{Q}_r)^2$
Sintaxis en la macro-función en SAS que estima con 80 réplicas, el término $\frac{1}{90(1-k)^2}=\frac{1}{20}$ es el que se modifica	DATA BRR_TEMP4; MERGE BRR_TEMP2 BRR_TEMP3; BY &BYVAR; VARI=((PV-STAT)**2)*(1/20); RUN;
Cambio en la sintaxis de la macro-función en SAS que estima con 100 réplicas, el término $\frac{1}{100(1-k)^2}=\frac{1}{25}$	DATA BRR_TEMP4; MERGE BRR_TEMP2 BRR_TEMP3; BY &BYVAR; VARI=((PV-STAT)**2)*(1/25); RUN;

Para usar las macro-funciones de PISA en SAS, se importó el archivo de datos según el procedimiento mostrado en la figura G2.

En las siguientes líneas, en la sentencia DATA se hicieron asignaciones a nuevas variables y se mantuvieron aquellas que se utilizaron. A la variable NACIONAL se le asignó el valor de cero, pues las macro-funciones necesitan una variable que distinga a la muestra para estimar toda la población.

Figura G2. Lectura de datos en SAS, para la información de alumnos de sexto de primaria y la asignatura de Lenguaje y Comunicación

```
PROC IMPORT OUT= WORK.LyC_06
            DATAFILE= "D:\PLANEA 2015"
            DBMS=SAV REPLACE;


DATA BASE_LyC_06;
SET LyC_06;
W_FSTUWT=W_FSTR0;
NACIONAL=0;
KEEP NACIONAL SERVICIO SEXO EDAD EDAD_N EDAD_ACC
ID_ENT AB020 LYCpv1-LYCpv5 LYCNVL1-LYCNVL5 LYCNVLII1-
LYCNVLII5 LYCNVLIII1-LYCNVLIII5 w_fstr0-w_fstr100;
RUN;
```

En las secciones siguientes se describe cada una de las cinco macro-funciones de SAS® empleadas para obtener la estimación del estadístico descriptivo y su respectivo error estándar de muestreo. Éstos fueron: medias y frecuencias con y sin valores plausibles, y la estimación de diferencias de un mismo estadístico entre dos subpoblaciones.

PROC_MEANS_PV

Esta macro se utilizó para calcular estadísticos de una variable definida mediante valores plausibles con su error dado por el diseño muestral mediante el remuestreo BRR con factor Fay de 0.5 y por el error de imputación. Dentro de la macro-función se indican los estadísticos a estimar.

Los argumentos que utiliza la macro-función son:

- INFILE: es el nombre del archivo de entrada con los datos a analizar.
- REPLI_ROOT: corresponde a la raíz del nombre de los pesos (muestral y de replicación). El peso muestral se renombra como w_fstr0. En el caso de PLANEA 2015 la raíz es w_fstr.
- BYVAR: es (son) la(s) variable(s) de segmentación que indica la población de la cual se va a estimar.
- PV_ROOT: es la raíz del nombre de las variables de los valores plausibles.
- STAT: es el estadístico a estimar y puede ser alguno de los siguientes.
    - SUMWGT: para la suma de los pesos.
    - MEAN: para la media.
    - VAR: para la varianza.
    - STD: para la desviación estándar.
    - CV: para el coeficiente de variación.
    - MEDIAN: para la mediana.
    - Q1: para el primer cuartil.

    - Q3: para el tercer cuartil.
    - QRANGE: para el rango intercuartil.
- LIMIT= no: si "LIMIT: =no" entonces los siguientes dos argumentos no son necesarios, por lo que no es necesario declararlos en la entrada de la macro-función.
- LIMIT_CRITERIA: argumento en blanco.
- ID_SCHOOL: argumento en blanco.
- OUTFILE: es el nombre del archivo de salida, contiene el resultado de las estimaciones y su error estándar correspondiente.

En el ejemplo de la figura G3 se presenta la estimación de la media del logro de los estudiantes de sexto de primaria para Lenguaje y Comunicación, por tipo de escuela y sexo del alumno con su respectivo error estándar. El resultado se guarda en un archivo llamado "Media_LyC".

Figura G3. Código en SAS para calcular las estimaciones de la media de logro en Lenguaje y Comunicación por tipo de servicio y sexo

```
%include
"D:\MacrosPISA\proc_means_pv_100BRR.sas";

%BRR_PROCMEAN_PV(INFILE = LYC_06,
                 REPLI_ROOT = W_fstr,
                 BYVAR = SERVICIO SEXO,
                 PV_ROOT = LYCPV,
                 STAT = MEAN,
                 LIMIT= NO,
                 OUTFILE = Media_LyC);
```

PROC_FREC_PV

Las estimaciones de porcentaje de niveles de logro o porcentaje de alumnos que alcanzan al menos el nivel II o nivel III se calcularon utilizando la macro PROC_FREC_PV.

Los argumentos de la macro son:

- INFILE
- REPLI_ROOT
- BYVAR
- PV_ROOT
- LIMIT = no
- OUTFILE

El uso de la macro-función PROC_FREC_PV se ejemplifica en la figura G4, contiene el código en SAS que llama a la macro-función que estima el porcentaje de alumnos de sexto grado de primaria en cada uno de los niveles de logro en la asignatura Lenguaje y Comunicación por entidad federativa y sexo del alumno. Los resultados se guardan en un archivo llamado "NVL_LyC".

Figura G4. Código en SAS para estimar el porcentaje de alumnos en cada nivel de logro de Lenguaje y Comunicación por entidad federativa y sexo

```
%include "D:\MacrosPISA\proc_freq_pv_100BRR.sas";

%BRR_FREQ_PV(INFILE=LYC_06,
             REPLI_ROOT=W_FSTR,
             BYVAR= ID_ENT SEXO,
             PV_ROOT=LYCNVL,
             LIMIT=NO,
             OUTFILE= NVL_LyC);
```

PROC_MEANS_NO_PV

Esta macro-función es análoga a la PROC_MEANS_PV, pero hace estimaciones de estadística univariada sin hacer uso de la metodología de valores plausibles. Es decir, el cálculo se basa en la descripción de la sección 2.1.

Los argumentos que utiliza son:

- INFILE
- REPLI_ROOT
- VAR
- STAT
- BYVAR
- LIMIT = no
- OUTFILE

Por ejemplo, si se quiere calcular la desviación estándar de la edad de los alumnos, se puede usar el código de la figura G5. La variable EDAD, contiene la edad del alumno al momento de la aplicación (junio de 2015), considerando el mes de nacimiento.

Figura G5. Código en SAS para calcular la desviación estándar en la edad de los alumnos de sexto de primaria, por sexo

```
%include "D:\MacrosPISA\proc_means_no_pv100BRR.sas";

%BRRPROCMEAN(INFILE = LyC_06,
             REPLI_ROOT = w_fstr,
             BYVAR = SEXO,
             VAR = EDAD,
             STAT = STD,
             LIMIT= NO,
             OUTFILE = Edad_DE);
```

PROC_FREC_NO_PV

Cuando el objetivo fue estimar porcentajes por categoría, por ejemplo, porcentaje de hombres y mujeres con cierta característica, se usó la macro-función PROC_FREC_NO_PV.

Los argumentos de esta macro-función son:

- INFILE
- REPLI_ROOT
- BYVAR
- VAR
- LIMIT = no
- OUTFILE

Por ejemplo, la pregunta AB020= *¿Cuántos focos hay en tu casa (incluyendo los de las lámparas)?* cuenta con cinco categorías, "Ninguno", "De 1 a 5", "De 6 a 10", "De 11 a 15" y "Más de 15". En la figura G6 se muestra el código fuente utilizado para calcular el porcentaje estimado de alumnos que contestaron cada una de las categorías de la pregunta AB020, así como el error estándar de cada estimación de porcentaje por servicio para alumnos de sexto de primaria en la asignatura de Lenguaje y Comunicación. El resultado se guarda en un archivo llamado "Prc_AB020".

Figura G6. Código en SAS para calcular el porcentaje de alumnos en cada categoría de la pregunta AB020 del cuestionario de contexto, sin usar valores plausibles

```
%include "D:\MacrosPISA\proc_freq_no_pv_100BRR.sas";

%BRR_FREQ(INFILE=LyC_06,
          REPLI_ROOT=w_fstr,
          BYVAR=SERVICIO,
          VAR=AB020,
          LIMIT=NO,
          OUTFILE=Prc_AB020);
```

PROC_DIF_PV

En ocasiones fue necesario valorar la diferencia de la media y su error estándar entre las estimaciones de dos subpoblaciones. Por ejemplo, si se quiere conocer si existe diferencia significativa en la media de logro entre los alumnos de edad normativa y los de extraedad. En este caso, la macro-función de PISA PROC_DIF_PV, estima la diferencia usando los valores plausibles entre dos subpoblaciones. Es importante señalar que la macro-función también permite estimar diferencias de la mediana, desviación estándar, entre otros.

Los argumentos que utiliza la macro-función son:

- INFILE
- REPLI_ROOT
- BYVAR
- PV_ROOT
- COMPARE: es la variable categórica sobre la cual se hará la diferencia.
- CATEGORY: son las categorías de la variable COMPARE de las que se quiere contrastar. Se deben de escribir separadas por un espacio.
- STAT puede ser: SUMWGT, MEAN, VAR, STD, CV, MEDIAN, Q1, Q3 ó QRANGE.
- LIMIT= no.
- OUTFILE

En la figura G7 se muestra un ejemplo en donde se calculan las diferencias de las desviaciones estándar del logro en Matemáticas entre hombres y mujeres, para los alumnos de sexto grado de primaria por tipo de servicio escolar.

Figura G7. Código en SAS para estimar diferencias entre las desviaciones estándar entre hombres y mujeres dentro de los diferentes tipos de servicio escolar

```
%include "D:\MacrosPISA\proc_dif_pv_100BRR.sas";

%BRR_PROCMEAN_DIF_PV(INFILE = LyC_06,
                     REPLI_ROOT = w_fstr,
                     BYVAR = SERVICIO,
                     PV_ROOT = LYCPV,
                     COMPARE = SEXO,
                     CATEGORY = 1 2,
                     STAT = STD,
                     OUTFILE = Dif_DE_Sexo);
```


Cabe señalar que, cuando se desea calcular diferencia entre más de dos poblaciones se considera el contraste tomando las categorías por pares. Por ejemplo, si se quiere la diferencia de la media de logro entre las cinco categorías de la variable AB020 en el argumento CATEGORY de la llamada a la macro se tendrá: CATEGORY = 1 2 3 4 5, (las categorías se separan por un espacio). En este caso las diferencias se estiman por pares para todos los pares posibles.

Estimaciones para logro

Una vez que en el capítulo anterior se han descrito las macros propuestas por PISA, se explicará su aplicación en el cálculo de los valores reportados en los resultados de logro.

Los resultados de logro se reportan en cuatro archivos de Excel:

I) Lenguaje y Comunicación de alumnos de sexto grado de primaria.
II) Lenguaje y Comunicación de alumnos de tercer grado de secundaria.
III) Matemáticas de alumnos de sexto grado de primaria.
IV) Matemáticas de alumnos de tercer grado de secundaria.

A su vez, en estos archivos hay nueve tablas con estructura similar entre archivos, así, bastará describir la construcción de las nueve tablas que están contenidas en un archivo para dar un bosquejo general de los cuatro archivos de Excel de resultados en las estimaciones de logro, las cuales están publicadas en el portal web del INEE.[1]

Tablas 1, 2 y 6

En las tablas 1, 2 y 6 se reporta el puntaje promedio y la desviación estándar del logro, para lo cual se utilizó la macro PROC_MEANS_PV (figura G3). Los argumentos se definieron con base en la declaración de la figura G2.

En el caso de las tablas 1 y 2 también se reportan las diferencias entre los puntajes de logro de hombres y mujeres y entre los alumnos con edad normativa y extraedad. Para obtener estas estimaciones se utilizó la macro PROC_DIF_PV (figura 7).

Tablas 3, 4, 5, 7 y 8

Las tablas que reportan los porcentajes de nivel de logro en Lenguaje y Comunicación y Matemáticas de los alumnos de sexto grado de primaria y tercer grado de secundaria, son las 3, 4, 5, 7 y 8 (tabla 2). Para las estimaciones de porcentajes de estas tablas, se utilizó la macro PROC_FREC_PV descrita en la sección 3.3. De los argumentos para identificar con base en la declaración en la macro de SAS (figura G2), se consideraron las siguientes definiciones en la macro.

Para las estimaciones por nivel de logro PV_ROOT=LYCNVL, debido a que con esta raíz se identifican las variables de los niveles de logro que obtuvieron los estudiantes. Para el porcentaje de alumnos que alcanzan al menos el nivel II, PV_ROOT=LYCNVLII. Y para el porcentaje de los que alcanzan al menos el nivel III, PV_ROOT=LYCNVLIII. En la figura G2 el ejemplo que se muestra corresponde a la estimación por niveles de logro.

En las tablas 4 y 5 se reporta la diferencia entre hombres y mujeres y entre alumnos en edad normativa y extraedad, respectivamente. Para ello, se usó la macro PROC_DIF_PV considerando PV_ROOT= LYCNVLII y PV_ROOT= LYCNVLIII como variables de proporción, así se buscó la diferencia de la proporción (o diferencia de la media) entre los estudiantes que alcanzaron al menos el nivel II o nivel III, según el caso. Posteriormente, se multiplicó por cien la diferencia y su respectivo error estándar para presentar la diferencia porcentual en cada caso.

[1] Ver http://www.inee.edu.mx/index.php/planea/bases-de-datos-planea

En la figura G8 se ejemplifica la llamada a la macro para obtener una de las diferencias de la tabla 5: porcentaje de estudiantes que alcanzan al menos el nivel II de logro en Lenguaje y Comunicación de sexto grado de primaria, por tipo de escuela, entre edad normativa y extraedad.

Figura G8. Código en SAS para calcular diferencias en estimaciones

```
%MBRR_PROCMEAN_DIF_PV(INFILE = LyC_06,
                      REPLI_ROOT = w_fstr,
                          BYVAR = SERVICIO,
                        PV_ROOT = LYCNVLII,
                        COMPARE = EDAD_N,
                       CATEGORY = 1 2,
                           STAT = MEAN,
                        OUTFILE = Dif_Media_EN);
```

Tabla 9

En la tabla 9 se reportan los porcentajes de acierto a nivel nacional y por tipo de escuela: general pública, indígena, comunitaria y privada, para cada uno de los reactivos de la prueba según la asignatura; así como la dificultad de cada uno. Los porcentajes (p+) fueron calculados como el número de alumnos que contestaron correctamente un reactivo respecto al total de alumnos en la muestra por cada grupo de segmentación, como se muestra en la ecuación (1).

$$p^{+} = 100\frac{C}{C+I+\mathrm{O}} \tag{1}$$

Donde:

C es el número de alumnos que contestaron correctamente.

I es el número de alumnos que contestaron incorrectamente.

O es el número de alumnos que omitieron su respuesta.

Estimaciones de contexto

La información de los cuestionarios de contexto A y B de alumnos, titulares de grupos y directores, ya descrita en la sección 1, se compone de variables con respuestas categóricas. Además, se han incluido en los resultados del cuestionario B de alumnos las variables: EDAD_ACC, EDAD_ANT, EDAD_EES, EDAD_N y SEXO. Para la estimación de esta información se utilizó la macro descrita en la sección 3.5.

Los resultados de contexto se reportan en seis archivos de Excel:

I) Cuestionarios A y B de alumnos de sexto grado de primaria.
II) Cuestionario de titulares de grupo de sexto grado de primaria.
III) Cuestionario de directores de sexto grado de primaria.
IV) Cuestionarios A y B de alumnos de tercer grado de secundaria.
V) Cuestionario de titulares de grupo de tercer grado de secundaria.
VI) Cuestionario de directores de tercer grado de secundaria.

El reporte de resultados de los cuestionarios de contexto de PLANEA 2015 se conforma del porcentaje de respuesta de cada una de las opciones de respuesta por pregunta y su respectivo error estándar debido al muestreo. Calculando las estimaciones a nivel nacional y segmentando por tipo de escuela, tamaño de población y nivel de marginación.

H. Procedimiento de escalamiento y validación de escalas de los datos de los cuestionarios de contexto. PLANEA 2015

Elaborado por: Unidad de Evaluación del Sistema Educativo Nacional, Dirección General de Medición y Tratamiento de Datos, Dirección de Tratamiento de Datos

Las Evaluaciones de Logro referidas al Sistema Educativo Nacional (ELSEN) consideran en su evaluación los cuestionarios A y B de contexto de alumnos, los cuales se componen de numerosos reactivos orientados a conocer las características de los alumnos (en los aspectos personal, familiar, escolar y social) con la finalidad de obtener escalas que permitan analizar características no observables directamente.

Este anexo está dividido en secciones que describen cómo fueron construidos los índices simples y de escala que son posibles derivar con los datos de dichos cuestionarios. Además, se describe la metodología de escalamiento y los análisis que se realizaron para la validación de los constructos. Asimismo, se describen recomendaciones para los reactivos, desde el Modelo de Crédito Parcial (MCP), para garantizar su adecuado funcionamiento.

Índices simples

Son los construidos sólo con transformaciones aritméticas o recodificaciones de uno o más reactivos. En este caso $Edad\ del\ alumno$, $Edad\ Normativa$, $Edad\ Anticipada$, $Extraedad\ Severa$, $Edad\ en\ años\ cumplidos\ por\ categoría$.

Edad del alumno

El índice $Edad$ se construye con las variables EDAD_AC $(Edad_{AC})$ y MES_NAC (Mes_{Nac}) de los datos. Este índice es la edad del alumno en años al momento de la aplicación de la evaluación. El cálculo depende de si el alumno nació en un mes previo o posterior al mes de aplicación. Si el alumno nació en un mes previo al mes de aplicación (ya fue su cumpleaños), a su edad en años cumplidos se le suma la proporción del año que hay de diferencia entre su mes de nacimiento y el mes de aplicación. Si el alumno nació en un mes posterior al mes de aplicación (no ha sido su cumpleaños), a su edad en años cumplidos se le suma uno y se le resta la proporción del año de la diferencia entre su mes de nacimiento y el mes de aplicación.

La ecuación para calcular $Edad$ es la siguiente:

$$Edad = \begin{cases} \dfrac{12 * Edad_{AC} + |Mes_{Nac} - Mes_{aplicacion}}{12}, si\ Mes_{Nas} \leq Mes_{aplicacion} \\ \dfrac{12 * Edad_{AC} + 12 - |Mes_{Nac} - Mes_{aplicacion}}{12}, si\ Mes_{aplicacion} < Mes_{Nas} \leq 12 \\ Edad_{AC}, \quad si\ Mes_{Nac}\ está\ perdido \\ Perdido, \quad si\ Edad_{Ac}\ está\ perdid \end{cases}$$

Donde $Edad_{Ac}$ es la edad del alumno en años cumplidos al momento de la aplicación de la evaluación. Mes_{Nac} es el mes de nacimiento del alumno en formato numérico (del 1 al 12). $Mes_{aplicacion}$ es el número correspondiente al mes que se aplicó la evaluación que en este caso fue el número 6 (junio).

A partir de esta información $Edad$ fue clasificada como $Edad\ Normativa$ (EDAD_N), $Edad\ Anticipada$ (EDAD_ANT), $Extraedad\ Severa$ (EDAD_EES) y $Edad\ en\ años\ cumplidos\ por\ categoría$ (EDAD_ACC), con base en el artículo 65, fracción I de la Ley General de Educación, que establece: "*La edad mínima para ingresar a la educación básica en el nivel preescolar es de 3 años, y para nivel primaria 6 años, cumplidos al 31 de diciembre del año de inicio del ciclo escolar*".

A continuación se presenta la descripción de las clasificaciones de $Edad$.

Edad normativa

El índice EDAD_N o $Edad\ Normativa$ clasifica a los alumnos que ingresaron al primer grado de primaria hasta con un año más de los 6 años cumplidos al 31 de diciembre que especifica la ley. Este índice para alumnos con $Edad$ a lo más de 12.5 años (12 años 6 meses) para sexto grado de primaria y $Edad$ a lo más de 15.5 años (15 años 6 meses) para tercero de secundaria al momento de la aplicación toma el valor 1 y se clasifica como $Edad\ Normativa$. Los alumnos que no tienen $Edad\ Normativa$ se clasifican como $Extra-edad$ y el índice toma el valor 2. Si la respuesta en EDAD_AC fue omitida, también se omite el valor.

Edad anticipada

El índice EDAD_ANT, $Edad\ Anticipada$, toma el valor 1 y se clasifica como tal si el alumno ingresó a primero de primaria con menos de la edad legal mínima de 6 años cumplidos al 31 de diciembre del ciclo escolar. Esto equivale a $Edad$ menor a 11.5 años (11 años, 6 meses) en sexto grado de primaria y $Edad$ menor a 14.5 años (14 años 6 meses) para tercero de secundaria en la fecha de aplicación. El índice toma el valor 0 si el alumno no tiene $Edad\ Anticipada$. Se omite el valor cuando EDAD_AC fue omitido.

Extraedad severa

El índice EDAD_EES o *Extraedad Severa* es una variable indicadora para los alumnos con dos años o más de rezago en el sistema escolar. Esto es, 8 años o más al 1 de septiembre cuando cursa primero de primaria por lo que el corte es en *Edad* de 14.75 años (14 años 9 meses) para sexto grado en primaria y en *Edad* de 17.75 (17 años 9 meses) para tercero de secundaria al momento de la aplicación. El índice toma el valor 1 y se clasifica como *Extraedad Severa* en valores de *Edad* mayores al punto de corte por grado escolar y toma el valor 0 en valores menores al punto de corte por grado escolar. Se omitió su valor cuando EDAD_AC fue omitido.

Para calcular los índices anteriores es necesario contar los meses transcurridos desde el 31 de diciembre al mes de aplicación para *Edad Normativa* y *Edad Anticipada*, así como desde el mes de septiembre al mes de aplicación para *Extraedad Severa*.

Edad en años cumplidos por categoría

El índice de *Edad en años cumplidos por categoría* se denota por EDAD_ACC en la base de datos pública y es una recategorización de EDAD_AC que corresponde a las siguientes categorías:

Categoría EDAD_ACC	EDAD_AC Sexto de primaria	EDAD_AC Tercero de secundaria
1	11 o menos	14 o menos
2	12	15
3	13	16
4	14 o más	17 o más

En caso de que el valor EDAD_AC fuera omitido, también se descartó el valor EDAD_ACC.

Índices de escalas

Los índices de escala se construyen a partir de un proceso de medición para variables latentes. La medición se puede pensar como un proceso que involucra tres elementos: un objeto de medición, un conjunto de números y un sistema de reglas. Las reglas sirven para asignar números o magnitudes al objeto de medición. El objeto de medición puede ser una variable observable (como edad o sexo) o una variable latente, por ejemplo, habilidad para algo o actitud hacia algo. Cualquier variable latente se puede ver como un continuo de magnitudes que se incrementan en una dirección dada, digamos de izquierda a derecha si el continuo se representara con una línea recta. Una variable latente usualmente se define mediante indicadores observables como son las respuestas a los reactivos en un instrumento (Dimitrov, 2012).

Los cuestionarios de contexto en PLANEA están conformados por reactivos con categorías de respuesta ordinal. Éstas representan actitud, satisfacción o percepción hacia algo (por ejemplo qué tan seguido ocurre algo). Para cada reactivo, los alumnos deben responder seleccionando una categoría que indique el nivel de su actitud, satisfacción o percepción en relación con la afirmación. La categoría seleccionada por el estudiante se asocia con un valor numérico que representa el puntaje del alumno para ese reactivo.

Las reglas utilizadas para asignar magnitudes al objeto de medición conforman un proceso iterativo de ajuste del MCP (Masters, 1982) y validación de constructo. Esta última incluye: análisis de la dimensionalidad de las escalas con análisis factorial exploratorio y confirmatorio, análisis de la consistencia interna de las escalas (Crocker, 2008) y equidad de escalas mediante el análisis del comportamiento diferencial de reactivos (Wu *y et al.*, 2007).

A continuación, se describen las escalas construidas para PLANEA 2015, las cuales son los objetos de medición que, en este caso, son rasgos o constructos latentes en los alumnos. Cada escala lleva el nombre del rasgo latente que se pretende medir. Se describen los reactivos que sirvieron como indicadores observables en cada caso y los puntajes que se asignaron a las categorías de cada reactivo.

Escala "habilidades para la convivencia escolar (HCE)"

Esta escala se compone de 30 reactivos del cuestionario A del alumno. Los puntajes para cada categoría de los reactivos se asignaron de la siguiente forma respetando el principio de orden ascendente del valor de las categorías de modo que vayan en dirección a la escala.

Escala "recursos familiares asociados al bienestar (RFAB)"

Esta escala se construye con los reactivos del cuestionario B de alumno: AB015, AB020, AB022, AB023, TV y AB026-AB033. La variable TV es la unión de las variables AA024 y AA025 codificadas como: "no tener televisión" (0), "tener televisión" (1) y "tener televisión y servicio de paga"(2).

Todos los reactivos se codificaron de modo que el puntaje más bajo es cero (categoría base) que indica ausencia del rasgo latente o la menor cantidad posible del rasgo que se va a medir en el alumno y el valor aumenta en uno conforme la categoría de respuesta denota mayor cantidad del rasgo latente en el alumno. Se dice que se asignan puntajes que aumentan en el sentido de la escala.

Metodología de escalamiento y validación de constructo

El escalamiento está basado en los procedimientos del Programme for International Student Assessment (PISA) (OECD, 2014; OECD, 2009a).

Procedimiento de escalamiento

El escalamiento o construcción de escalas de medición del constructo se realizó bajo el MCP de Masters (1982) perteneciente a la familia de modelos Rasch utilizados en la teoría de respuesta al ítem. El MCP se utiliza para modelar probabilidades entre categorías adyacentes de respuestas al reactivo. Tradicionalmente, la escala de respuestas al reactivo es de cuatro categorías, a las cuales se asignan puntuaciones 0, 1, 2 y 3 donde valores mayores indican mayor cantidad del constructo, atributo o rasgo latente que se va a medir (Wilson, 2005).

El Modelo de Crédito Parcial

El Modelo de Crédito Parcial (MCP) basado en Van der Linden y Hambleton (1997) permite analizar respuestas que se registran en dos o más categorías ordenadas. Este modelo tiene parámetros separados de personas y reactivos que permiten hacer comparaciones. Todos los parámetros del modelo son *localizaciones* sobre una variable o rasgo latente.

Los parámetros del reactivo i $\delta_{i1},\delta_{i2},\ldots,\delta_{im_i-1}$ para cada categoría representan la posición sobre la escala del rasgo latente medido en la que un alumno tiene la misma probabilidad de responder en la categoría x y en la categoría $x-1$ ($ieP(x_{ni}=x|\theta_n,\delta_{ix})=P(x_{ni}=x-1|\theta_n,\delta_{ix})$). Los valores de los parámetros de reactivo no necesariamente se encuentran ordenados con respecto al puntaje. En evaluaciones de logro δ_{ik} se denomina de *dificultad* del reactivo.

El parámetro para alumnos (o cantidad de rasgo sobre la escala *logit*) θ_n es la localización modelada del alumno n sobre la escala del rasgo latente medido. En evaluaciones de logro θ_n se denomina *habilidad* del alumno.

Todos los parámetros son ubicaciones sobre la escala del rasgo latente medido y pueden utilizarse para mapear el significado cualitativo de la variable y para interpretar los parámetros de los alumnos.

Otra manera de definir los parámetros de los reactivos del modelo es $\delta_{ik}=\delta_i,+\tau_{ik}$, donde δ_i es el valor esperado de las localizaciones $\delta_{i1},\delta_{i2},\ldots,\delta_{im_i-1}$ para el reactivo i y $\delta_i,+\tau_{ik}$ es el umbral de localización o umbral de Rasch-Andrich, donde τ_{ik} es un umbral que compensa la ubicación de las categorías respecto a δ_i. Cuando el modelo restringe a las localizaciones como combinación linealmente independiente, esto es, $\Sigma\delta_{ik}=0$, entonces, los umbrales de localización también son linealmente independientes y $\sum\tau_{ik}=0$.

Sea el $i-ésimo$ reactivo con puntajes de categoría $0,1,\ldots,m_i$, la probabilidad de que el $n-ésimo$ alumno obtenga el puntaje x, con localización θ_n, está dado por:

$$P(x_{ni}=x|\theta_n,\delta_{ik})=\frac{exp\sum_{k=0}^{x}(\theta_n-\delta_{ik})}{\sum_{h=0}^{m_i}exp\sum_{k=0}^{h}(\theta_n-\delta_{ik})}$$

Este modelo considera que:

- Las categorías de cada reactivo están ordenadas de menor a mayor puntaje y se descomponen en pares de categorías adyacentes. Se aplica un modelo dicotómico para cada par, de manera que el alumno tiene mayor cantidad del rasgo latente cuando obtiene mayor puntaje.
- La probabilidad asignada a cada categoría es positiva, esto es, la probabilidad de que un alumno alcance la puntuación correspondiente a esa categoría es mayor a cero y la probabilidad total de cada reactivo es la suma de todas las probabilidades de categoría. Esto indica que para que un alumno alcance la probabilidad de obtener k puntos más que $k-1$ puntos es porque ha superado la etapa de obtener $k-1$ puntos. Lo que se dice es que este alumno "ha pasado a través de" la etapa de $k-1$ puntos. Pero además ha pasado a través de la etapa de los puntos $k-2, k-3, \ldots, 0$.

Cuando se estiman los parámetros para los reactivos y se cumplen con los criterios de validación que se describen en la sección 4, se utiliza estimación de máxima verosimilitud ponderada para obtener puntajes individuales para los alumnos. Estos puntajes se denotan wle (por sus siglas en inglés) y se obtienen mediante el software ACER Conquest®.

En la base de datos final, los parámetros estimados wle y $\widehat{\delta}_{ik}$ se presentan trasladados. Este traslado se realiza con el procedimiento que se describe en la sección 3.1.3.

Estadísticas de ajuste al Modelo de Crédito Parcial

Correlación punto biserial

La correlación punto biserial para la categoría k en el reactivo i se calcula con una variable indicadora y_{ikn} que vale 1 para los alumnos que anotan el puntaje de la categoría k en el reactivo i. La correlación punto biserial se calcula entre el conjunto de valores y_{ikn} y el correspondiente S_n. Donde S_n se define en ConQuest® como el puntaje crudo para un alumno, el cual consiste en la suma del puntaje alcanzado por éste dividido por el puntaje máximo posible que el alumno podría haber alcanzado. Sólo aquellos casos que respondieron al reactivo son incluidos en el cálculo (Wu *et al.* 2007, p. 149).

Estadísticos de ajuste basados en residuales

Los estadísticos de ajuste basados en residuales son una forma de índices de ajuste para reactivos y para alumnos, derivados de las diferencias entre el puntaje observado y el puntaje esperado del reactivo (residuales) por cada reactivo y por cada alumno. Wright (1977) propuso varios de estos estadísticos basados en los residuales estandarizados.
El residual estandarizado es la diferencia del puntaje obtenido por el alumno a un determinado reactivo (puntaje observado) menos el puntaje esperado dividido por la varianza del puntaje (Wu y Adams, 2007).

Outfit (cuadrado medio del error). Es la suma de los cuadrados de los residuales estandarizados de todos los alumnos dividida por el número total de alumnos, se puede demostrar que el valor esperado de este estadístico es uno.

Infit (cuadrado medio del error ponderado). Es la suma del cuadrado del residual ponderado por la varianza de la respuesta al reactivo, se puede demostrar que el valor esperado de este estadístico es uno.

De acuerdo con Boone, Staver y Yale (2014) cuando no hay valores atípicos es indistinto valorar el *outfit* y el *infit*. En nuestro caso se valoró el *infit*.

Rango de valores aceptables para el *infit*

Existen diversos autores que han establecido rangos de aceptabilidad para el ajuste basado en los residuales estandarizados, para fines de interpretación en este documento el tipo de ajuste y rango a considerar es el reportado por Wilson (2005), quien establece que si bien no existe un límite absoluto de lo que es un buen valor cuadrado medio ponderado (*infit*) indica que 0.75 es un valor inferior razonable, así como 1.33 es un valor superior.

Es importante notar que estos estadísticos de ajuste no indican qué tan cercana está la curva observada de la esperada, sino más bien que la pendiente de la curva observada es la misma que la de la esperada o teórica, es decir, el cuadrado medio de ajuste no mostrará la falta de ajuste si la curva característica observada del reactivo está cerca o lejos de la curva característica esperada del reactivo (Wu y Adams, 2007).

Traslado de la escala

Las mediciones de las escalas se reportan con media 50 y desviación estándar 10 para evitar el uso de números negativos y disminuir la cantidad de decimales. Esta transformación de la escala es posible ya que cualquier transformación lineal preserva la interpretación de la probabilidad de los valores *logits* (Wilson, 2005). A continuación, se describe el procedimiento de traslado de escala basado en la descripción de Wright y Stone (1979).

Los valores originales de la escala se trasladan mediante una trasformación lineal de tipo:

$$y = \alpha + \gamma x$$

En la cual x está en la escala *logit* original mientras que y está en la escala que se denomina trasladada donde γ representa el origen de la escala trasladada que es el valor de y donde la escala *logit* original x vale 0, γ denota una constante de espaciamiento que determina las unidades de la escala trasladada.

Si denotamos a y como B para los valores de los alumnos en la escala trasladada y como D para los valores de los reactivos en la escala trasladada, θ y δ son los valores originales para alumnos y reactivos respectivamente. Por lo que la ecuación de la recta para cada valor en la escala trasladada sería:

$$B = \alpha + \gamma\theta \qquad \text{para alumnos}$$

$$D = \alpha + \gamma\delta \qquad \text{para reactivos}$$

Las desviaciones estándar (De) de las mediciones para los alumnos y los reactivos cumplen que:

$$De(B) = \gamma De(\theta)$$

Se realiza una transformación preliminar utilizando los valores de la media m y desviación estándar s de la escala *logit* original para obtener valores θ' y δ' con media 0 y desviación estándar 1 de la forma:

$$\theta' = (\theta - m)/s$$

$$\delta' = (\delta - m)/s$$

Se toma el valor de origen $\alpha = 50$ y de espaciamiento $\gamma = 10$ para, finalmente, obtener los valores correspondientes a la escala trasladada con las siguientes ecuaciones:

$$B = \alpha + \gamma\theta' = \alpha + \gamma(\theta - m)/s$$

$$D = \alpha + \gamma\delta' = \alpha + \gamma(\delta - m)/s$$

Con esto se tiene que la escala trasladada tiene media 50 y desviación estándar 10.

Análisis de la dimensionalidad de las escalas con análisis factorial exploratorio y confirmatorio

El análisis factorial es una técnica estadística que determina la estructura subyacente entre las variables en el análisis. Es decir, es factible identificar grupos de variables dependiendo de sus correlaciones, a cada conjunto de variables identificadas se le denomina factor.

Un factor es una combinación lineal de las variables observadas. Los factores representan las dimensiones (constructos) subyacentes que resumen o representan a las variables observadas (Hair, Black, Babin y Anderson, 2009).

El análisis factorial se puede llevar a cabo desde dos perspectivas, análisis factorial exploratorio (AFE) y análisis factorial confirmatorio (AFC). El AFE permite identificar la existencia de factores en el conjunto total de reactivos de interés, suele ser utilizado cuando los

constructos y los reactivos asociados con ellos no cuentan con un vasto historial de investigación; y el AFC se usa cuando se han identificado factores que se pueden asociar a un constructo de interés y se desea probar la unidimensionalidad de éstos.

Las escalas de la prueba PLANEA 2015 (HCE y RFAB) fueron sometidas a estos análisis con el fin de corroborar que los reactivos candidatos a formar las escalas fueran unidimensionales.

Análisis factorial exploratorio

El AFE es útil para encontrar una forma de concentrar la información obtenida de las respuestas de los reactivos (variables observadas) en un número pequeño de factores (variables latentes). En este proceso se suele perder información (variabilidad), es por ello que también se busca que haya una pérdida mínima de información, es decir, que los factores logren recuperar la mayor variabilidad de los reactivos.

El modelo empleado para medir observaciones en relación a k factores, es el Modelo Lineal de Análisis Factorial (MLAF) que es:

$$x = \Lambda_x \xi + \in$$

Donde x es el vector de variables observadas (reactivos) cuya dimensión es $q \times 1, q$ es el número de reactivos/preguntas en el análisis para N estudiantes. La matriz Λ_x llamada matriz de cargas factoriales, cuya dimensión es $q \times k$, contiene las cargas/coeficientes de regresión para cada factor, las entradas de las matriz son λ_{ij} que representan la carga al factor i del reactivo j. El vector ξ de dimensión $k \times 1$ es el vector de factores. Finalmente $\in$ es un vector $q \times 1$ de variables únicas que contienen tanto el error de medición como el error específico.

Los supuestos a cumplir en torno a este modelo son:

$$E(\zeta) = 0 \qquad E(\in) = 0 \qquad cov(\zeta, \in) = 0$$

Donde cov (.,.) es el operador covarianza. Bajo estos supuestos, la matriz de covarianzas de los datos observados puede ser escrita en la forma de la ecuación fundamental de análisis factorial, de modo que:

$$\left.\right] = cov(xx')$$

$$= \Lambda_x E(\xi\xi')\Lambda_x + E(\in\in')$$

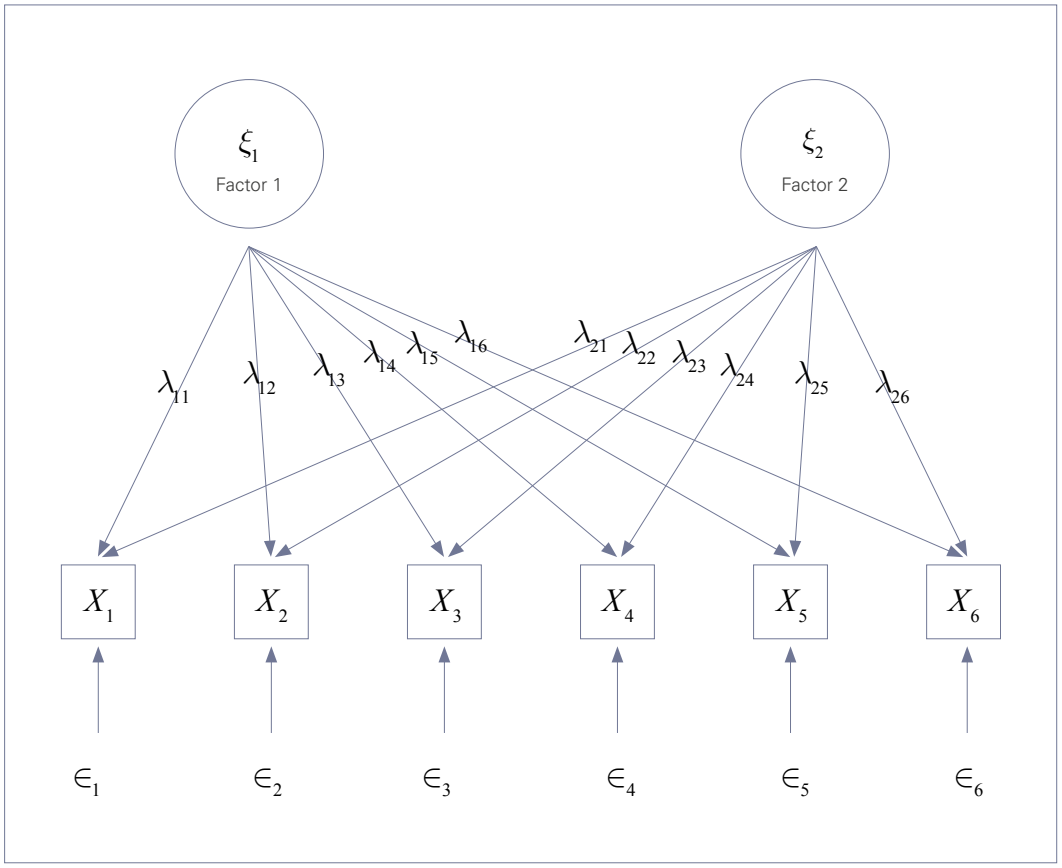
$$= \Lambda_x \Phi \Lambda_x' + \Theta_\in$$

Donde Σ es la matriz de covarianzas poblacional $q \times q, \Phi$ es la matriz $k \times k$ de varianzas y covarianzas de los factores y Θ_{ϵ} es una matriz diagonal q x q de varianzas únicas.

El modelo supone un número k de factores, por lo tanto, para identificar el valor apropiado en un AFE se suele utilizar un gráfico de sedimentación (Cattell, 1966), en éste se grafican ordenados los autovalores de la matriz de correlación. Se considera como número óptimo de factores el número de autovalores mayores a uno (Thompson, 2004). También se toma en cuenta el porcentaje de varianza explicada acumulada conforme se aumenta el número de factores, un 60% es satisfactorio (Hair *et al.*, 2009). El porcentaje de varianza explicada en cada factor se calcula como el porcentaje que cada autovalor aporta a la suma de autovalores.

En la figura 1 se muestra gráficamente el modelo empleado en el análisis factorial exploratorio para el caso de dos factores. Una vez que se ha identificado el número de factores se reconoce qué reactivos se encuentran asociados a cada factor. Para ello se analizan las cargas factoriales de la relación del reactivo con cada factor. Hair *et al.* (2009) recomiendan conservar aquellas relaciones con carga factorial mayor o igual a 0.5 y cargas factoriales entre 0.3 y 0.4 que son mínimamente aceptables.

Figura 1. Modelo factorial exploratorio con dos factores y seis reactivos.



Adaptado de Thompson (2004, p. 38).

Cuando no es sencilla la interpretación de la asociación de los reactivos con los factores, Hair *et al*. (2009) sugieren emplear un método de rotación para lograr soluciones de factor simples y teóricamente con mayor interpretación; el efecto final de la rotación de la matriz de factores es redistribuir la varianza de los factores anteriores (no rotados) a los posteriores (rotados). En otras palabras, la rotación de factor implica mover los ejes de medición del factor de las ubicaciones de las variables medidas en el espacio factorial, de modo que la naturaleza de los constructos subyacentes se hace más obvia para el investigador; cabe mencionar que la rotación no es posible cuando un solo factor es extraído (Thompson, 2004).

En el caso de las escalas de los cuestionarios A y B de PLANEA 2015 no se hizo uso de alguna rotación porque el objetivo era probar la unidimensionalidad de las escalas, es decir, el comportamiento de un factor.

En la literatura existen diferentes métodos de rotación entre los que destacan rotaciones ortogonales[2] y oblicuas[3] (Hair *et al.*, 2009; Thompson, 2004; Raykov y Marcoulides, 2011; Asparouhov y Muthén, 2009). En general, no hay una regla específica para utilizar algún tipo de rotación (Hair *et al.*, 2009); los autores Asparouhov y Muthén (2009, p. 429) argumentan que no hay una razón estadística para preferir un criterio de rotación sobre otro, la decisión e interpretación de los resultados está enteramente en las manos del investigador. Sin embargo, las rotaciones ortogonales son las más utilizadas cuando el propósito es una estructura simple;[4] si el propósito de la investigación es indagar sobre las correlaciones entre las constructos vinculados con los factores, se suele usar rotaciones oblicuas (Thompson, 2004). Es por esta razón que Raykov y Marcoulides (2011) indican que una rotación oblicua tiende a producir resultados más valiosos en investigaciones de comportamiento y sociales.

Harrington (2009, p. 10) recomienda que si una nueva medida de constructo está siendo desarrollada con un fuerte marco teórico, entonces se puede omitir el AFE y continuar con el AFC.

En el caso de las escalas construidas con los cuestionarios de PLANEA 2015, el análisis exploratorio se elaboró usando el software Stata (StataCorp, 2011) y Mplus (Muthén y Muthén, 2015) para calcular las estimaciones del modelo antes descrito y verificar que se cumplieran las condiciones antes mencionadas.

Análisis factorial confirmatorio

Después de verificar, en el AFE, que la estructura de los reactivos estuviera acorde con las escalas, se elaboró un AFC. Este análisis fue realizado únicamente para las escalas a fin de saber si se cumple el supuesto de unidimensionalidad. Previo a ello, los datos se prepararon de manera que las categorías estén en sentido creciente hacia la escala, que no haya

[2] Algunos ejemplos de rotaciones ortogonales son: Varimax, Quartimax, Equimax.

[3] Algunos ejemplos de rotaciones oblicuas son: Geomin, Oblimin, Promax.

[4] Una sola carga factorial alta para cada variable en sólo un factor (Hair *et al.*, 2009).

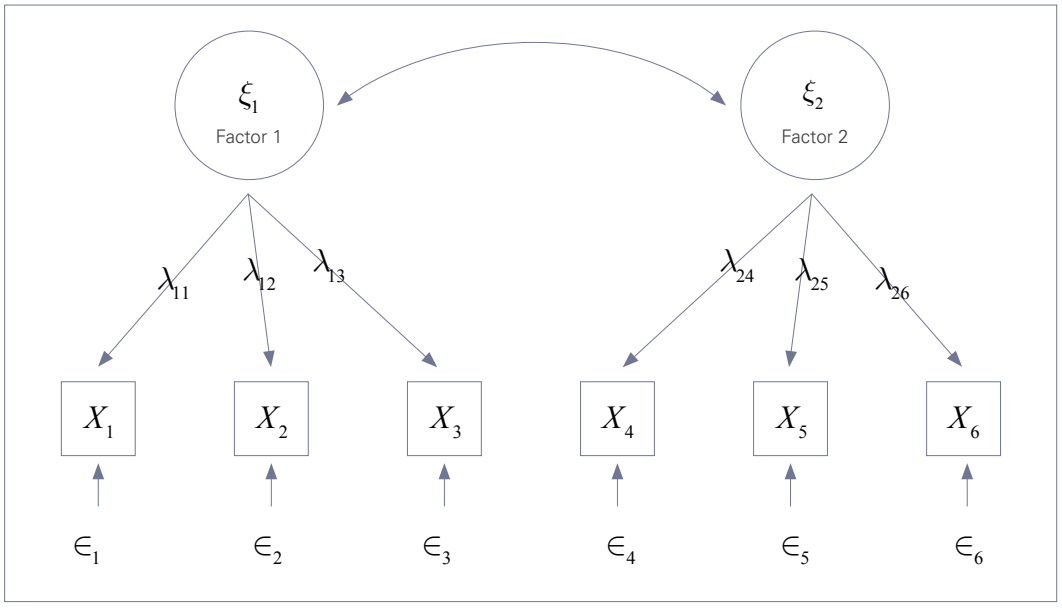
frecuencias menores al 5% en alguna categoría, y si las hay, estas categorías se colapsan con la categoría aledaña.

El modelo matemático empleado en este análisis es el mismo que el anterior (MLAF), salvo que existen restricciones en la matriz Λ_x (debido que el número y las restricciones impuestas en la matriz asociada al modelo factorial son más de K^2). Las restricciones impuestas en los elementos de Λ_x reflejan una hipótesis *a priori* de estructura simple dada por el investigador (Crocker, 2008). En la figura 2 se pueden apreciar de forma gráfica las restricciones, esto es, se restringe a que cada reactivo esté relacionado a un solo factor, teniendo carga factorial cero el resto de las relaciones. Lo anterior significa que el modelo, para cada escala, puede ser planteado como MLAF de un solo factor. Cabe señalar que no es posible rotar el modelo restringido, a diferencia del análisis exploratorio, porque destruiría el posicionamiento de las restricciones y, en consecuencia, la hipótesis bajo estudio (Kaplan, 2000). En relación con lo anterior, Brown (2015, p. 37) argumenta que en un AFC no es necesario rotar porque un reactivo está especificado para un solo factor.

Las escalas obtenidas de PLANEA 2015 fueron evaluadas calculando las estimaciones del modelo teórico y se verificó que cumplieran con las recomendaciones de varios autores en la literatura sobre el tema. A continuación, se describen dichas recomendaciones e interpretaciones para las estimaciones del modelo.

El software empleado para obtener los resultados del modelo suele mostrar las estimaciones de las cargas factoriales y las correlaciones, así como las estimaciones estandarizadas de las mismas; estas últimas corresponden a las estimaciones obtenidas tomando en cuenta tanto los valores estandarizados de la escala (factor) como de los reactivos, es decir, reescala los valores de la escala (factor) y del reactivo para que su varianza sea igual a uno (Wang y Wang, 2012, p. 39).

Figura 2. Modelo factorial confirmatorio con dos factores, cada uno de ellos con tres reactivos



Adaptado de Thompson (2004, p. 38).

La estimación estandarizada de las cargas factoriales representa la relación del reactivo con la variable latente; por ejemplo, si la carga factorial de un reactivo tuviera valor de 0.8 para una escala, un incremento estandarizado en el valor de la escala está asociado a un incremento estandarizado de 0.8 en el reactivo. El cuadrado de la estimación estandarizada de la carga factorial es la proporción de varianza del reactivo que es explicada por la escala, también suele llamarse a este valor la comunalidad del reactivo (Brown, 2015). Siguiendo con el ejemplo, una carga factorial del 0.8 representa el 64% (0.8^2=0.64) de la variabilidad del reactivo que está siendo explicada por la escala. Wang y Wang (2012) escriben que convencionalmente una carga factorial con valor de 0.3 es considerada como un valor de corte, aunque también suele usarse el valor de 0.4. Por otro lado, Hair *et al.* (2009) sugieren como regla de dedo que las cargas factoriales estandarizadas estimadas deberían ser mayor a 0.5 e idealmente mayor a 0.7.

Los residuos de correlación (o de covarianza) son la diferencia entre la matriz de correlación (o covarianza) implicada por el modelo y la matriz de correlación observada. Kline (2005) menciona como regla de dedo que un residuo de correlación mayor a 0.10 sugiere que el modelo estimado no explica adecuadamente la correlación entre los elementos del modelo; cuando el modelo es adecuado los residuos de correlación tendrán una distribución normal. Por su parte, Hoyle (1995) menciona que si el promedio de los valores absolutos de los residuos de correlación es de 0.02, el modelo es "bueno", esto se debe a que el promedio de error en las correlaciones es de 0.02 (aun existiendo correlaciones altas y bajas). Hoyle (1995) también argumenta que, si el valor más grande en los residuales de correlación es pequeño, es decir 0.10, el modelo es sólo marginalmente inadecuado para algunas variables; si la mayor discrepancia entre la correlación implicada por el modelo y la matriz de correlación observada es grande, es decir 0.4, claramente el modelo no explica algunas de las correlaciones.

Una vez que se obtienen las estimaciones del modelo en un análisis factorial confirmatorio, se evalúa el modelo, dicho de otra manera, se verifica la bondad de ajuste del modelo. Al efecto se han desarrollado índices que reflejen la discrepancia entre la matriz de covarianzas muestral y la matriz de covarianzas implicada por el modelo con las cargas factoriales y correlaciones estimadas (Lei y Wu, 2007). El primer índice en ser desarrollado fue la estadística X^2 (Brown, 2015), sin embargo este índice es raramente usado debido a su sensibilidad respecto al tamaño de muestra (Brown, 2015; Wang y Wang, 2012; Lei y Wu, 2007). Enseguida se enlistan los índices alternativos, usados para evaluar la bondad de ajuste del modelo.

Índices incrementales o comparativos. Índice Tucker-Lewis (TLI, por sus siglas en inglés) o índice de ajuste no normado (NNFI), índice de ajuste comparativo (CFI, *Comparative Fit Index*) y el índice de ajuste normado (NFI, *Normed Fit Index*) compara el modelo estimado con el modelo nulo.[5] El rango de los índices está entre cero y uno, valores grandes indican mejora del modelo estimado respecto al modelo nulo; valores mayores o igual a 0.9 son generalmente aceptados como indicadores de un buen ajuste, se ha sugerido incrementar el punto de corte a 0.95 (Wang y Wang, 2012; Lei y Wu, 2007; Brown, 2015).

[5] Modelo en el cual las covarianzas entre todos los indicadores se fijan como cero (Brown, 2015).

Índices de ajuste absoluto. Raíz cuadrada media residual estandarizada (SRMR, *standardized root mean square residual*), puede ser visto como el promedio entre la discrepancia de las correlaciones observadas y las correlaciones predichas por el modelo. Un criterio para la indicación de un buen ajuste del modelo es SRMR menor o igual que 0.08, valores entre 0.08 y 0.10 son aceptables (Lei y Wu, 2007; Wang y Wang, 2012; Brown, 2015). Media cuadrática del error de aproximación6 (RMSEA, *root mean square error of approximation*), es un índice de "error de aproximación" porque evalúa el grado con el que el modelo se ajusta razonablemente bien a la población. Valores de RMSEA menores o iguales a 0.06 indican un buen ajuste del modelo (Brown, 2015); algunos autores citan otros rangos como Wang y Wang (2012) y Kaplan (2000): igual a cero indica ajuste perfecto; menor a 0.05, ajuste cercano; entre 0.05 y 0.08, ajuste razonable; de 0.08 a 0.10, ajuste mediocre, y mayor a 0.10, mal ajuste.

Índices de criterio de información. Son criterios estadísticos empleados para la comparación de modelos, esto es porque en ocasiones se desea un modelo sobre un modelo alternativo que pudiera tener más o menos correlaciones o cargas factoriales a estimar. Los criterios están en función del modelo de máxima verosimilitud (valores de verosimilitud cercanos a uno implican valor del criterio bajo, valores de verosimilitud cercanos a cero implican valor del criterio alto) más una penalización en función del número de observaciones y del número de correlaciones y cargas factoriales a estimar (Wang y Wang, 2012). Criterio de información de Akaike (AIC, *Akaike's information criterion*), criterio de información Bayesiana (BIC, *Bayesian information criterion*) y criterio de información Bayesiana ajustado al tamaño de muestra (ABIC, *Adjusted Bayesian information criterion*). El modelo con menor medida del criterio de información tiene un mejor ajuste a los datos en relación a los modelos alternativos (Wang y Wang, 2012; Brown, 2015).

Si un modelo no cumple adecuadamente con las recomendaciones de un buen ajuste, éste puede modificarse añadiendo o removiendo correlaciones o cargas factoriales considerando el marco teórico y la interpretación de las modificaciones (Wang y Wang, 2012; Harrington, 2009; Brown, 2015).

Los índices de modificación (MI, *Modification index*), también llamados multiplicadores de Lagrange, son cantidades calculadas para cada carga factorial y correlación fija del modelo, es decir, que no fueron estimados dentro del modelo debido a que se fija su valor a cero. Los MI reflejan una aproximación de que tanto la estadística X^2 decrecerá si dicha carga factorial o correlación fuera estimada y no fija como en el modelo original (Brown, 2015; Hair *et al*., 2009).

6 Brown (2015, p. 71) lo clasifica como un índice de corrección de parsimonia, los cuales incorporan una función de penalización por modelo pobre en parsimonia.

Brown (2015) indica que el MI puede conceptualizarse con una estadística X^2 con 1 grado de libertad, por lo tanto, valores mayores o iguales a 3.84[7] sugieren que el estimar libremente la carga factorial o correlación mejorará significativamente el ajuste del modelo. Sin embargo, el MI es sensible al tamaño de muestra (Brown, 2015; Hair *et al.*, 2009), por lo cual se sugiere usar estos índices en conjunción con los parámetros esperados de cambio (EPC, *expected parameter change*) debido a que estos últimos son útiles para muestras grandes (Brown, 2015; Wang y Wang, 2012; Kaplan, 1989; Millsap y Olivera-Aguilar, 2012; Whittaker, 2012).

Los EPC son valores que indican el valor esperado de la estimación de la carga factorial o correlación, si ésta hubiera sido estimada, análogo a estos valores existen los *parámetros esperados de cambio estandarizados* (SEPC, por sus siglas en inglés). En Brown (2015) se cita el trabajo de Jöreskog (1993) quien sugiere que se estime libremente la carga factorial o la correlación, es decir, que se haga la modificación sobre aquella cuyo MI es el más grande (y el EPC) si la estimación puede ser interpretada sustancialmente; en caso de no existir base sustancial se considera el siguiente más grande en MI y EPC para darle el mismo tratamiento.

En el caso de las escalas de PLANEA 2015, el análisis fue realizado con el programa EQS (Bentler y Wu, 2014) tomando en consideración los criterios antes mencionados.

Análisis de la consistencia interna de las escalas mediante el coeficiente de confiabilidad Alpha de Chronbach

La confiabilidad de las escalas se midió con el coeficiente Alpha de Chronbach. Este coeficiente se estima mediante la expresión:

$$\hat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_i^k = \sigma_i^2}{\sigma_x^2}\right)$$

Donde k es el número de reactivos en la prueba, σ_i^2 es la varianza del reactivo i y σ_x^2 es la varianza total de la prueba (Crocker, 2008).

Este coeficiente es cercano a uno cuando la suma de las varianzas de los reactivos es más pequeña respecto a la varianza total de la prueba. En *The Cambridge Dictionary of Statistics* (Everitt y Skrondal, 2010) la interpretación sugerida para este coeficiente es: menor a 0.60 inaceptable, (0.60, 0.65) indeseable, (0.65, 0.70) mínimamente aceptable, (0.70, 0.80) bueno, (0.80, 0.90) muy bueno y mayor a 0.90 considere acortar la escala reduciendo el número de reactivos (Dunn, 2004). En general, el uso que se le da al instrumento determina los valores aceptables para este coeficiente.

[7] Una distribución X^2 con un grado de libertad es igual a la distribución normal estandarizada al cuadrado, por lo tanto, el punto $1.96^2 = 3.84$ es el punto crítico con un $\alpha = 0.05$.

La confiabilidad también se calcula eliminando un reactivo a la vez. Si el valor de alfa aumenta considerablemente al eliminar un reactivo, se considera no incluir tal reactivo en la escala final.

Además, se calcula la correlación reactivo-total que es el coeficiente de correlación de Pearson de un reactivo individual con el total de la escala calculada sumando los reactivos restantes. Una regla de dedo indica que los reactivos que correlacionan abajo de 0.20 deben ser descartados (Streiner y Norman, 2010; Everitt y Skrondal, 2010). Los reactivos con correlaciones negativas podrían estar en sentido inverso a la escala. Se deben reacomodar sus categorías de manera que, a mayor valor de la categoría de respuesta, indique mayor cantidad del rasgo latente en el alumno. Es decir, que los puntajes de las categorías de respuesta aumenten en el sentido de la escala. Todos los cálculos anteriores se realizaron con el programa SPSS Statistics® 23.

Análisis de la equidad de las escalas mediante el estudio del comportamiento diferencial de reactivos

La equidad es una característica deseable en los instrumentos de medición cuando se realiza evaluación a gran escala. El análisis de Comportamiento Diferencial de Reactivos (DIF, por sus siglas en inglés) sirve para confirmar la equidad de un instrumento de medición; se basa en estadísticos asociados con la cantidad de constructo de un reactivo. En preguntas de contexto, un reactivo con DIF se puede dar por problemas del reactivo como: ambigüedad, mala comprensión, o por elementos particulares de la población: culturales, sociales, económicos, etcétera.

A continuación, se presenta el mecanismo cuantitativo para el diagnóstico de DIF que está basado en un Modelo de Teoría de Respuesta al Ítem: Crédito Parcial con Facetas (MCPF). Este modelo permite agregar términos (facetas) que modifican la localización del reactivo y que indican diferentes factores que pueden tener influencia sobre la localización de éste (Wu y *et al.*, 2007).

Definición

De acuerdo con Angoff (1993), un reactivo con DIF muestra diferentes propiedades estadísticas para distintos grupos luego de controlar las diferencias en las localizaciones de los grupos. Por otro lado, Wilson (2005) da una definición más intuitiva: un reactivo presenta DIF cuando sujetos en distintos grupos. pero con cantidad de constructo similar, tienden a contestar el reactivo de forma diferente. La idea central es que la cantidad de constructo se deja fija. Sin embargo, el primer autor no hace referencia al análisis cualitativo del reactivo. En la práctica, expertos en los fundamentos teóricos del constructo analizan los reactivos diagnosticados con DIF. Esto se hace para tomar la decisión de incluir o descartar tal reactivo para la construcción de las escalas.

Grupos

Los grupos comparados se definieron por nivel educativo, género e indigenismo. Esto llevó a cinco comparaciones. La comparación por nivel educativo: primaria y secundaria, se realizó debido a que se aplicó el mismo cuestionario en los dos niveles educativos. Luego, se contrastó por género: hombres-mujeres y, por último, los grupos de indigenismo: no indígena-indígena. Esto último debido a que la población indígena se considera un grupo vulnerable.

Estimación

El análisis de DIF se realizó con el programa ACER ConQuest® de acuerdo con lo que recomienda el manual de operación en la sección de detección de DIF (ACER Conquest® [Wu, Adams y Haldane, 2007). El MCPF modela la probabilidad en categorías adyacentes de que el alumno n en el grupo m con localización en la escala del rasgo latente θ_n conteste la opción k del reactivo i en lugar de la opción $k-1$:

$$log_e\left(\frac{P_{nimk}}{P_{nim(k-1)}}\right)=\theta_n-(\delta_i-\rho_m-\gamma_m-\gamma_k)\quad con\quad K=1,\ldots,m_i$$

Donde:

P_{nimk} es la probabilidad de que el alumno n tenga el puntaje k en el grupo m del reactivo i,

$P_{nimk(k-1,}$ es la probabilidad de que el alumno n tenga el puntaje $k-1$ en el grupo m del reactivo i,

m_i es el número de opciones del reactivo i.

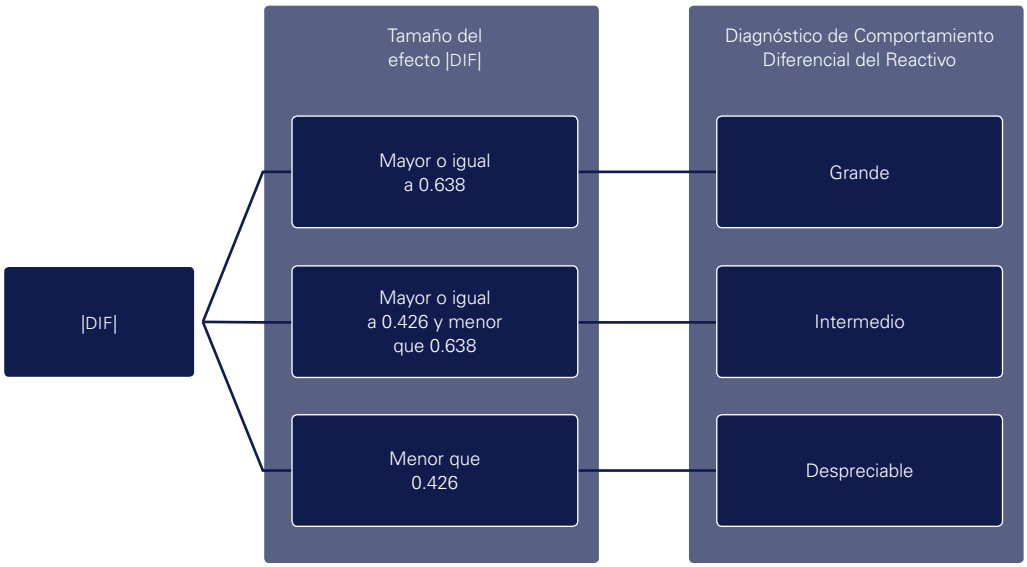
El MCPF tiene cuatro parámetros que interactúan con el parámetro de localización del alumno. Dos términos de efectos principales: reactivo δ_i y grupo ρ_m, los cuales estiman la localización para cada reactivo y la media para cada grupo. Un término de interacción: reactivo*grupo γ_m, que modela la variación en la localización del reactivo entre los grupos. El cuarto término: reactivo*grupo*step, modela la diferencia debida a cada paso de reactivo y cada grupo γ_k.

El estadístico para el diagnóstico se etiquetó como |DIF|. Este estadístico es la diferencia aritmética en valor absoluto de los parámetros estimados de interacción reactivo*grupo γ_m o de ubicación media para cada grupo. En particular, para dos grupos, en la parametrización de Conquest® basta con tomar el doble del valor del parámetro estimado.

Criterio para el diagnóstico

Cuando el tamaño de muestra es grande, los datos tienen suficiente potencia para detectar diferencias muy pequeñas, de modo que las pruebas estadísticas tradicionales frecuentemente muestran significancia. La prueba tradicional consiste en tomar como estadístico de prueba la diferencia entre las localizaciones de los reactivos para los dos grupos (o la diferencia entre los parámetros de interacción reactivo*grupo) y verificar si son más grandes de lo esperado dividiendo esta diferencia por su error estándar, como en una prueba tipo z. De esta manera, todos los reactivos serían diagnosticados con DIF.

Por lo anterior, se consideró el tamaño del efecto, que es la diferencia en valor absoluto entre los parámetros estimados para la interacción reactivo*grupo, la cual se denominó |DIF|. El criterio de clasificación está basado en Wilson (2005) y ETS (s. f.):



Los reactivos con diagnóstico de DIF grande deben tomarse con precaución. El tratamiento para un reactivo diagnosticado con DIF grande puede variar. Se recomienda consultar la opinión de expertos en los fundamentos teóricos del constructo. En ciertos casos, se decidió eliminar el reactivo de la escala. En otros casos, se conservó el reactivo y se reportó para que se tomaran los resultados con cautela.

Cabe resaltar que el análisis de DIF es un procedimiento posterior a la verificación de la bondad de ajuste de los reactivos. Los reactivos con ajuste deficiente pueden dar un diagnóstico espurio.

Recomendaciones que describen las características psicométricas de los reactivos dentro del modelo de medición de crédito parcial

Varios autores han desarrollado recomendaciones que describen las características psicométricas del reactivo bajo el modelo de Rasch de Crédito Parcial (MCP). Estas recomendaciones están basadas en los valores que toman las estimaciones de los parámetros del modelo y en las estadísticas numéricas de ajuste del mismo. Además, se pueden explorar las representaciones gráficas de la relación entre la localización del reactivo y del alumno o las curvas características de probabilidad, de probabilidad acumulada o de puntaje esperado del modelo. Cabe destacar que las recomendaciones apoyadas en resultados gráficos son limitadas en comparación con las basadas en rangos y criterios cuantitativos.

Recomendaciones paramétricas y de ajuste del modelo

Los lineamientos que se presentan a continuación están basados en Dimitrov (2012), Wu y Adams (2007) y Wilson (2005), considerando adaptaciones para el MCP.

- Al menos 10 observaciones en cada categoría

Debe haber al menos 10 alumnos en cada categoría de los reactivos. Cuando la frecuencia en la categoría es baja, los umbrales se estiman de manera pobre y potencialmente inestable.

- Distribución regular de las observaciones

La distribución de las observaciones en las categorías dentro de cada reactivo puede ser de tipo unimodal, que tiene un pico central o con categorías extremas o bimodal que tiene picos en las categorías extremas. Por lo tanto, se debe verificar la distribución empírica de las observaciones para cada reactivo en las categorías.

- El promedio de la localización de los alumnos debe ser monótono con las categorías

El *promedio de la localización de los alumnos en cada categoría* sirve para identificar si existe un desorden en las categorías, se esperaría que valores altos de categoría tengan *promedio de la localización de los alumnos en cada categoría* alto.

- La correlación punto biserial es creciente con las categorías

Según Wu y Adams (2007), se esperaría que a mayor categoría, mayor puntaje total. Esto implica que la correlación punto biserial debería incrementarse conforme se incrementa la categoría. Sin embargo, en Wu *et al.* (2007) se reporta que algunas veces los puntos biseriales no están ordenados cuando una proporción muy pequeña o muy grande de las respuestas al reactivo está en una categoría.

- Estadísticos de ajuste basados en residuales

El *infit* es un estadístico que indica que la pendiente de la curva observada es la misma que la esperada o teórica. Se espera tener valores cercanos a uno. Wilson (2005) plantea como intervalo aceptable para el *infit* (0.75, 1.33).

El *outfit* es un estadístico de ajuste que de acuerdo con Dimitrov (2012) permite detectar desorden en las categorías (de cada reactivo). Esta medida puede indicar el grado con el cual las categorías tienen terminología ambigua o si la cantidad de categorías es excesiva. Cuando el valor del *outfit* es cercano a 1 se obtiene un buen ajuste, y a lo más es recomendable que sea igual a 2 (Linacre, 2002).

- Los umbrales $\tau's$ avanzan monótonamente con las categorías

La probabilidad de obtener una calificación alta debería incrementarse con el aumento en la localización θ del alumno (Dimitrov, 2012). Por lo cual se debe verificar que los $\tau's$ de las categorías se incrementan conforme aumenta la categoría.

Si los umbrales no tienen avance monótono entre categorías adyacentes se dice que hay desorden de umbrales, desorden de calibraciones de paso o, simplemente, desorden de paso (Linacre, 2002). Las causas pueden ser, según Andrich, DeJong y Sheridan (1997), que no es posible definir un rango modal para alguna categoría, que no se utilicen todas las categorías o por un desorden empírico de las categorías.

- Amplitud de la categoría

Cuando los parámetros $\tau_1,\tau_2,\ldots,\tau_{mi-1}$ están ordenados respecto a las m_i categorías, la amplitud de la categoría intermedia $k(k=2,3,\ldots,m_i)$ se define como la distancia entre los umbrales para esa categoría. El intervalo (τ_k,τ_{k-1}) es el intervalo modal de la categoría, esto es, el intervalo en la escala *logit* donde la categoría k es más probable de ser observada que otra categoría. Dimitrov (2012) basado en resultados de Linacre (2002), sugiere que una amplitud adecuada es de 1.4 a 5.0 *logits*. Linacre (2002) notó que "cuando la distancia es mayor a 5.0 *logits* la información dada en el centro del reactivo es menor que la mitad de lo que daría una simple dicotomía". Por otro lado, un intervalo modal angosto (menor a 1.4 *logits*) puede indicar que la categoría representa un segmento muy angosto de la variable latente o corresponde a un concepto que está definido de manera deficiente en la mente de los que responden.

- Qué hacer cuando no se cumplen algunas recomendaciones

Si los umbrales $\tau's$ no están ordenados o si otros estadísticos de ajuste muestran valores fuera de rango en los datos y el modelo, entonces puede ser que colapsar categorías revele el número efectivo y el ordenamiento de categorías (Wright, 1994). Esto se puede hacer mediante un estudio de seguimiento usando el nuevo formato de categorías.

Recomendaciones gráficas del modelo

Por otro lado, la literatura hace referencia a gráficos que son de utilidad para la interpretación del modelo, sin embargo, es limitada la información que se extrae del análisis visual de dichos gráficos (Dimitrov, 2012). Es por esta razón que a continuación se plantean dos notas a observar, tomadas de Wu y Adams (2007), y Bond y Fox (2001). En estas notas se revisa el ajuste del modelo y la orientación de los resultados de los alumnos, de una forma visual.

- Curvas características de probabilidad acumulada de los reactivos. Son útiles para observar el comportamiento de los reactivos, en ellas se grafica por cada categoría la probabilidad de obtener dicha categoría o una mayor dependiendo de la localización (parámetro θ del modelo). En la misma gráfica se muestran dos curvas: la curva dada por el modelo ajustado (curva teórica) y la curva dada por el modelo de predicción (curva empírica). Wu y Adams (2007) mencionan que ambas curvas características del reactivo deben ser cercanas.
- Comparación de las distribuciones de alumnos y reactivos (Mapa de Wright). Estos valores se muestran gráficamente en el mapa de distribución latente o Mapa de Wright, del lado izquierdo las localizaciones estimadas de los alumnos y del lado derecho los umbrales de Thurstone de los reactivos.

El umbral de Thurstone para el reactivo i y la categoría k es la localización sobre la escala del rasgo latente en la cual la probabilidad de alcanzar la categoría k o una mayor alcanza 0.5 (Wu y Adams, 2007). Debido a que es una medida que acumula probabilidades, estos umbrales siempre están ordenados respecto a las categorías del reactivo y permiten dar una mejor medida de cantidad de constructo de las categorías de los reactivos que los parámetros del MCP δ_{ik} que no necesariamente están ordenados respecto a las categorías.

En la parte superior del lado izquierdo del Mapa de Wright se representan los alumnos con mayor cantidad de constructo y en la parte inferior se encuentran los alumnos con menor cantidad de constructo. Análogamente, en la parte superior del lado derecho están los reactivos con mayor cantidad de constructo y en la parte inferior se encuentran los reactivos con menor cantidad de constructo. Cabe mencionar que el rango del mapa está en escala *logit*, un rango aceptable es de -3 a +3 (Bond y Fox, 2001).

La medida se considera adecuada cuando hay un traslape entre los rangos de valores para ambas distribuciones (Dimitrov, 2012; Bond y Fox, 2001). En evaluaciones de logro, cuando la distribución de localizaciones de los alumnos se concentra en un rango de valores menores a las localizaciones de los reactivos, se dice que el instrumento es relativamente difícil; en el caso donde la distribución de localizaciones de los alumnos se encuentra en un rango de valores mayores a las localizaciones de los reactivos, el instrumento es relativamente fácil (Bond y Fox, 2001). En evaluaciones de contexto diríamos que la cantidad de constructo del instrumento es mayor a la de la población en el primer caso o que la cantidad de constructo en la población es mayor a la del instrumento en el segundo caso.

Coordinación
Andrés Sánchez Moguel y Mariana Zúñiga García

Redacción e integración del manual técnico
Sandra Patricia Reyes Lüscher

Diseño, desarrollo y validación de las pruebas PLANEA
María Cristina Aguilar Ibarra, Oscar Barrera Sánchez, Luis Manuel Cabrera Chim, Cecilia Kissy Guzmán Tinajero, Sara Rivera López y María Margarita Tlachy Anell

Diseño y análisis de los cuestionarios de contexto de PLANEA
Mariana Zúñiga García, Carolina Cárdenas Camacho, Juan Bosco Mendoza Vega, Jannet Valtierra Jiménez y Enrique Daniel Paredes Ocaranza

Diseño de muestras, procesamiento y análisis de datos
Laura Delgado Maldonado, Edgar Andrade Muñoz, Marisela García Pacheco, José Gustavo Rodríguez Jiménez, Enrique Estrada Cruz, Glenda Patricia Guevara Hernández, Violeta de la Huerta Contreras, Claudia Nila Luevano, Irma Rocío Zavala Sierra y Román Aguirre Pérez

Revisión técnica, revisión de estilo y diseño gráfico de los instrumentos
José Manuel Silva Cabrera, Sandra Fabiola Medina Santoyo, Elsa Yunuhen Nambo Peñaloza, Reynaldo Agustín Villafuerte Aguilar, Jaime Díaz Pliego, Elsa Mendieta Parra, Verónica Pérez Martínez y Nayelli Vilchis de la Concha

Levantamiento y procesamiento de datos
Oswaldo Palma Coca, Felipe Mendoza Lara, María de la Luz Ortiz González, Salvador Castro Tinoco y Sergio Sánchez Ortega

DIRECTORIO

Junta de Gobierno

Teresa Bracho González
Consejera presidenta

Bernardo Naranjo Piñera
Consejero

Sylvia Schmelkes del Valle
Consejera

Patricia Vázquez del Mercado
Consejera

Unidades Administrativas

Miguel Ángel de Jesús López Reyes
Unidad de Administración

Jorge Antonio Hernández Uralde
Unidad de Evaluación del Sistema Educativo Nacional

Rolando Erick Magaña Rodríguez (encargado)
Unidad de Información y Fomento de la Cultura de la Evaluación

Francisco Miranda López
Unidad de Normatividad y Política Educativa

José Roberto Cubas Carlín
coordinación de direcciones del inee en las entidades federativas

Tomislav Lendo Fuentes
Coordinación Ejecutiva de la Junta de gobierno

José de la Luz Dávalos (encargado)
Órgano Interno de Control

Dirección General de Difusión y Fomento de la Cultura de la Evaluación
José Luis Gutiérrez Espíndola

Dirección de Difusión y Publicaciones
Blanca Gayosso Sánchez

INEE
Instituto Nacional para la Evaluación de la Educación
México