INFORME 3: FINAL (REVISADO)

SDP 005/2016 ESTUDIO "ANÁLISIS DE METODOLOGÍAS PARA EVALUAR EL ALINEAMIENTO DE PRUEBAS ESTANDARIZADAS CON ESTÁNDARES DE CONTENIDO Y ESTÁNDARES DE DESEMPEÑO"

Presentado por:
Gilbert A. Valverde, Ph.D.
María José Ramírez, Ph.D.
Elisa de Padua Nájera, M.Sc.

De parte del:
Research Foundation for the State University of New York
Albany, New York, EEUU
22 de marzo, 2017

Índice

1. INTRODUCCIÓN

El propósito de este Informe 3 y final es dar cuenta en forma integral del estudio realizado por el equipo de investigación en el proyecto "*Análisis de metodologías para evaluar el alineamiento de pruebas estandarizadas con estándares de contenido y estándares de desempeño*". El alineamiento es fundamental para poder afirmar que tanto los estándares de desempeño como las pruebas están referidos a los estándares de contenido (el curriculum nacional), y así poder interpretar sus resultados como logro de los aprendizajes curriculares. En este estudio se analizaron tres dimensiones de alineamiento:

- Dimensión 1: Alineamiento entre el curriculum y los estándares de desempeño
- Dimensión 2: Alineamiento entre el curriculum y las pruebas
- Dimensión 3: Alineamiento entre los estándares de desempeño y las pruebas

El objetivo general que guía a este estudio es "Caracterizar las metodologías tradicionales y actuales, tanto cuantitativas como cualitativas, usadas para evaluar y mantener el alineamiento entre pruebas estandarizadas con estándares de contenido y estándares de desempeño en los países participantes de la prueba PISA 2012 más Ecuador". Este objetivo general se traduce en los siguientes objetivos específicos (Términos de Referencia, p. 34):

- "Caracterizar y describir los procesos que realizan distintos sistemas educativos para evaluar periódicamente el alineamiento que existe entre las pruebas estandarizadas y los estándares de desempeño establecidos.

- Identificar, describir y analizar metodologías usadas para evaluar y mantener el alineamiento entre los estándares de contenido, los estándares de desempeño y las pruebas estandarizadas establecidas.

- Identificar, en base a criterios, metodologías o elementos dentro de éstas, usados para evaluar y mantener el alineamiento entre los estándares de contenido y los estándares de desempeño con las pruebas estandarizadas que puedan ser transferidos o replicados en el sistema educacional chileno actual.

- Generar recomendaciones para el sistema educacional chileno respecto de estrategias o metodologías a implementar para mantener el alineamiento entre los estándares de contenido y los estándares de desempeño con las pruebas estandarizadas."

Para responder a estos objetivos, se usó una metodología de trabajo consistente en: (a) Una revisión documental sobre el curriculum, estándares de desempeño y pruebas estandarizadas, en los países participantes en PISA 2012, más Ecuador; (b) Un análisis en profundidad para tres sistemas educativos; y (c) Un diagnóstico o línea de base sobre la metodología actualmente utilizada en el Ministerio de Educación (MINEDUC) y la Agencia de Calidad de Chile para procurar el alineamiento del curriculum, estándares de desempeño, y pruebas estandarizadas. Los tres sistemas educativos seleccionados por el MINEDUC para los casos en profundidad fueron: NAEP (*National Assessment of Educational Progress*) de Estados Unidos (EEUU); el Programa de Evaluación de la Provincia de Ontario, Canadá; y PLANEA (*Plan Nacional de Evaluación de los Aprendizajes*) de México.

Los resultados de este estudio indican que en sólo unos pocos países hay información técnica disponible que trate sobre alineamiento con suficiente profundidad (15% de los 52 países revisados). La mayoría de los países (60%) no tiene información disponible sobre el tema. En los países que sí tienen información, esta usualmente se refiere a cuestiones generales de alineamiento entre el curriculum y las pruebas (25%). En cambio, la literatura académica sobre alineamiento es abundante, y ofrece variedad de modelos, criterios y procedimientos para realizar este tipo de estudios.

Los resultados muestran que en los tres programas de evaluación revisados en profundidad se realizan procedimientos sistemáticos para resguardar el alineamiento en las tres dimensiones arriba mencionadas. Sin embargo, en ninguno de estos programas existe un documento formal para hacer estudios o revisiones integrales y periódicas de alineamiento.

Dado que en el MINEDUC tampoco hay modelo formal para evaluar y revisar el alineamiento, el equipo de investigación recomienda la adopción de uno. El modelo propuesto sería de utilidad para que el MINEDUC realice estudios integrales y periódicos de alineamiento entre el curriculum, los estándares de desempeño, y las pruebas SIMCE. La realización de estos estudios contribuiría a asegurar mayores grados de validez y credibilidad de las pruebas SIMCE para medir los aprendizajes descritos en el curriculum nacional, y para reportar resultados por niveles de desempeño.

2. ANTECEDENTES SOBRE EL TEMA

2.1 DESARROLLOS TEÓRICOS Y PRÁCTICOS EN TORNO A PROCEDIMIENTOS DE ALINEAMIENTO

Este estudio tiene por objetivo conocer qué procesos o metodologías se emplean en los sistemas educativos de otros países para evaluar y mantener el alineamiento de las pruebas con los estándares de contenido (curriculum) y estándares de desempeño. Para ello, se realizó un estudio descriptivo sobre los procesos y metodologías para evaluar y mantener el alineamiento de pruebas estandarizadas con estándares de contenido y estándares de desempeño en países participantes de la prueba PISA 2012 más Ecuador y, además, se realizó una revisión bibliográfica acerca de metodologías de alineamiento propuestas en publicaciones técnicas y académicas. Esta revisión bibliográfica se presenta a continuación en este apartado.

Los sistemas educativos más efectivos suelen tener programas de evaluación que están estrechamente alineados con otros componentes del sistema. Esto es, sus programas de evaluación son coherentes con las expectativas de aprendizaje establecidas en el curriculum, con las prácticas pedagógicas en el aula, con la formación de profesores, entre otros. Esto permite hacer sinergias que movilizan a la mejora (Clarke, 2012). El alineamiento entre el curriculum, los estándares de desempeño, y las pruebas estandarizadas es, entonces, un aspecto fundamental de la coherencia global del sistema educativo.

Es importante señalar que, en evaluación educativa, el alineamiento es un término fundamentalmente ligado a la evidencia relativa al contenido de las pruebas[1]. Esto es, al grado en que las pruebas miden efectivamente los contenidos (ej. curriculum, estándares curriculares, dominios de evaluación, áreas disciplinarias) que pretenden medir. El alineamiento es condición necesaria para la correcta interpretación de los resultados de las pruebas (ej. puntajes, niveles de desempeño) en términos de logro de los aprendizajes asociados a los contenidos evaluados. En otras palabras, el alineamiento es necesario para hacer inferencias válidas sobre los resultados de las pruebas.

En los Estándares para Evaluación Educativa y Psicológica, la American Educational Research Association/American Psychological Association/National Council on Measurement in Education

[1] La evidencia relativa al contenido de las pruebas es lo que tradicionalmente se ha entendido como validez de contenido. Sin embargo, este término está en desuso dado que se refiere a un aspecto de validez en términos absolutos. Hoy en día, se enfatiza un concepto de validez integral, y relativo a las inferencias que se hacen de las evaluaciones.

(AERA/APA/NCME, 2014) definen el alineamiento como "el grado en que los contenidos y demanda cognitiva de las preguntas de las pruebas calzan con los contenidos y demanda cognitiva descritos en las especificaciones de las pruebas" (p. 216). Los siguientes estándares deben resguardarse para asegurar el alineamiento de las pruebas:

- Para cada propósito de las pruebas, las especificaciones de las pruebas deben decir claramente cómo los puntajes de las evaluaciones deben ser interpretados y utilizados (Estándar 1.11 y 4.1)

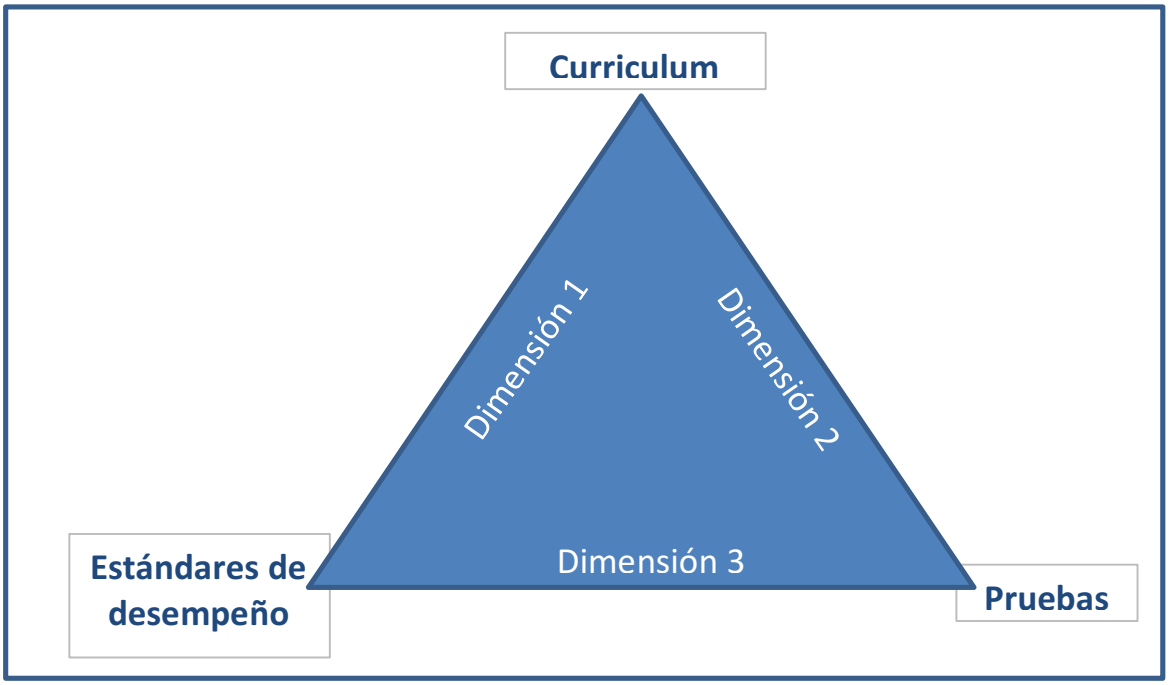
- Las especificaciones de la prueba deben incluir los propósitos de la prueba, la definición del constructo o dominio a evaluar, el largo de la prueba, formato de preguntas, características psicométricas deseadas de cada ítems y del conjunto de la prueba, y el orden de los ítems y secciones (Estándar 4.1, 4.2)

- Quienes desarrollan las pruebas deben documentar el grado en que el contenido y los procesos cognitivos evaluados en las pruebas representan los contenidos y procesos cognitivos y demanda cognitiva definidos en las especificaciones de la prueba (Estándar 4.12), y el grado en que éstos representan el constructo o dominio evaluado (12.4)

En el marco de este proyecto, el alineamiento entre el curriculum nacional (estándares de contenido), los estándares de desempeño (estándares de aprendizaje), y las pruebas SIMCE son necesarios para afirmar que las pruebas miden efectivamente los aprendizajes esperados del curriculum, que los estándares de desempeño reflejan adecuadamente el curriculum, y que las evaluaciones permiten dar cuenta del logro de los estándares de desempeño.

La Figura 1 muestra las tres dimensiones de alineamiento que surgen a partir de la relación entre el curriculum, los estándares de desempeño, y las pruebas.

Figura 1. Triángulo de alineamiento entre curriculum, estándares de desempeño, y pruebas.



La primera dimensión del triángulo se refiere al alineamiento entre el curriculum y los estándares de desempeño. Preguntas claves relativas a esta dimensión son: ¿Reflejan los estándares de desempeño los aprendizajes claves establecidos en el curriculum? ¿Describen los estándares de desempeño la progresión de aprendizajes necesaria para el logro de los objetivos de aprendizaje? ¿Existe una brecha entre el curriculum y los estándares de desempeño? ¿Cómo se justifica y qué medidas se toman para enfrentar esta brecha?

La segunda dimensión del triángulo se refiere al alineamiento entre el curriculum y las pruebas. Preguntas claves relativas a esta dimensión son: ¿Hay tablas de especificaciones basadas en el curriculum? ¿Las tablas de especificaciones describen los contenidos y habilidades que deben ser evaluados en las pruebas? ¿Los contenidos y habilidades evaluados en las pruebas reflejan lo estipulado en las tablas de especificaciones? ¿Existe una brecha entre el curriculum y las pruebas? ¿Cómo se justifica y qué medidas se toman para enfrentar esta brecha?

La tercera dimensión del triángulo se refiere al alineamiento entre estándares de desempeño y las pruebas. Preguntas claves relativas a esta dimensión son: ¿Reflejan los estándares de desempeño los contenidos y habilidades medidos en las pruebas? ¿Cubren los estándares el espectro de la escala de puntajes de las pruebas? ¿Es adecuada la cantidad y ubicación de los puntos de corte de los estándares de desempeño? ¿Permiten los estándares de desempeño diferenciar o discriminar entre estudiantes que quedan clasificados en distintos niveles de

desempeño? ¿Permiten las pruebas dar cuenta de los desempeños descritos en los estándares? ¿Existe una brecha entre los estándares de desempeño y las pruebas? ¿Cómo se justifica y qué medidas se toman para enfrentar esta brecha?

2.1.1 Modelos de alineamiento

Hay una vasta literatura académica sobre alineamiento entre el curriculum, estándares de desempeño, y pruebas (Baker, 2005; CCSSO, 2009; Case & Zucker, 2005; Davis-Becker 2013; Martone & Sireci, 2009; Näsström & Henrrikson, 2008; Roach, Niebling & Kurz, 2008; Vockley & Lang, 2009; Webb, 1997, 2007; entre otros). La literatura se refiere a distintos marcos conceptuales, metodologías, criterios, y procedimientos para realizar estudios de alineamiento, ya sea con propósitos de revisión o de evaluación. La dimensión más abordada en la literatura es la del alineamiento entre el curriculum y las pruebas. Esto se debe a la importancia de asegurar que las pruebas midan efectivamente el curriculum, condición necesaria para hacer inferencias válidas e interpretar los resultados de las pruebas en términos de logro de aprendizajes curriculares. Otra dimensión también abordada de manera importante es el alineamiento con prácticas pedagógicas, pero esta dimensión no es considerada en este estudio, aun cuando podría ser relevante que el MINEDUC la considerara en otros estudios.

A continuación se revisan tres modelos de alineamiento ampliamente reconocidos en EEUU, que están bien documentados, y que pueden ser de utilidad para Chile. Estos son el método Webb (2007), el modelo Achieve (Roach, 2008), y el modelo de estudios de alineamiento propuesto por Davis-Becker and Buckendahl (2013).

El método Webb (2007) es utilizado en varios estados de EEUU y fue desarrollado teniendo en consideración la necesidad de los estados de documentar el grado de alineamiento existente entre sus evaluaciones y los curriculum estatales (estándares de contenidos). Este método también fue una respuesta a la demanda de los estados por alinear sus evaluaciones con marcos de evaluación más innovadores (ej. Common Core State Standards, Partnership for 21st Century Skills).

Webb (1997) define el alineamiento como “el grado en el que una expectativa [los estándares de contenido] y la evaluación son coherentes y sirven de manera conjunta a que los estudiantes aprendan aquello que se espera que sepan y sean capaces de hacer” (p. 4). Otras definiciones de alineamiento ponen más énfasis en la relación entre curriculum y pedagogía, o evaluación y pedagogía.

El método Webb cuantifica el grado de alineamiento entre el curriculum y las pruebas (foco en evidencia relativa al contenido de las pruebas). Algunos criterios utilizados son:

a. Alineamiento de los ítems (preguntas, problemas, o tareas) con los ejes temáticos del curriculum. Para cada uno de los ítems de las pruebas, panelistas verifican si calzan en los contenidos y habilidades especificados en el curriculum.

b. Alineamiento de la demanda cognitiva o nivel de dificultad de los ítems versus el curriculum. Para cada uno de los ítems, panelistas verifican que los ítems elaborados sean adecuados en su nivel de dificultad para el grado evaluado.

c. Alineamiento global del conjunto de ítems con el curriculum. Para el conjunto de ítems de la prueba, panelistas verifican que los contenidos y habilidades especificados en el curriculum estén representados en una proporción adecuada en las pruebas.

Para cada criterio evaluado, se calcula un índice que es luego contrastado con un valor mínimo utilizado para juzgar si hay alineamiento o no. Si bien este método no especifica criterios para evaluar el alineamiento entre el curriculum y los estándares de desempeño, esto se podría hacer usando una variación del criterio de dificultad (ej. juzgando en qué medida los ítems de las pruebas reflejan la dificultad del estándar de desempeño).

El modelo Achieve también analiza el grado de coherencia entre el curriculum y las pruebas (foco en evidencia relativa al contenido de las pruebas). Consiste en un protocolo de alineamiento que se aplica en dos etapas: la primera considera un análisis de cada una de las preguntas o ítems que componen las pruebas en relación con las especificaciones de la prueba; la segunda etapa consiste en un juicio holístico del calce entre la evaluación (sus ítems y especificaciones) y los estándares de contenido (curriculum) a los que refiere. Esta metodología hace tanto un análisis cuantitativo como cualitativo de alineamiento (Rothman, et. al., 2002; Rothman, 2003).

El modelo Achieve se ha utilizado en EEUU en estudios de alineamiento principalmente entre estándares de contenido estatales y estándares de contenido interestatales (ej., Common Cores State Standards), y en estudios de alineamiento entre las evaluaciones estandarizadas estatales y los estándares estatales e interestatales. En la Tabla 1 se identifican los estados que han usado esta metodología.

Tabla 1: Estados de EEUU que han utilizado metodología de alineamiento Achieve

Estados que han usado Achieve en estudios de alineamiento entre estándares estatales e interestatales	Estados que han usado Achieve para determinar alineamiento entre estándares y evaluaciones
- California - Delaware - Indiana - Louisiana - Maryland - Massachusetts - Michigan - Minnesota - New Jersey - Ohio - Oklahoma - Oregon - Pennsylvania - Tennessee - Texas - Washington	- Hawai - Illinois - Indiana - Maryland - Massachusetts - Michigan - Minnesota - New Jersey - Oklahoma - Oregon - Pennsylvania - Texas - Washington

Fuente: http://www.achieve.org/who-we-are/history/state-reviews

Una diferencia relevante entre el método Webb y el modelo Achieve, es que este último considera las especificaciones de la prueba como un elemento central del análisis de alineamiento. Otra diferencia con el método de Webb es que no se establece un valor mínimo para indicar si la evaluación está o no alineada a los estándares de contenido. Sólo se basa en un juicio experto que determina el grado de alineamiento.

Davis-Becker y Buckendahl (2013) proponen un marco conceptual para evaluar y revisar el alineamiento. Para estos autores, el alineamiento debe ser juzgado sistemáticamente en función de la evidencia de validez disponible para afirmar que los resultados de las evaluaciones se pueden interpretar en función del dominio evaluado (ej. un área disciplinaria del curriculum).

El modelo identifica cuatro fuentes de evidencia de validez: validez de uso, de procedimientos, validez interna, y validez externa. La validez de uso (término similar a la validez de contenido) se refiere a evidencia que justifique la interpretación de los resultados de las pruebas como midiendo distintos niveles de aprendizaje de un dominio determinado (en el caso del SIMCE, de distintas áreas disciplinarias del curriculum nacional). La validez de procesos se refiere a la calidad técnica del diseño del estudio de alineamiento (métodos, procedimientos, calificaciones

de los panelistas). La validez interna se refiere la consistencia de juicios y respuestas de los panelistas de un mismo estudio de alineamiento. La validez externa se refiere a la consistencia de resultados entre distintos estudios de alineamiento.

A la literatura académica se suma una incipiente documentación sobre métodos de alineamiento utilizados en programas nacionales de evaluación en distintos países del mundo. Esta documentación suele centrarse en el alineamiento entre el curriculum y las pruebas, y aborda temas relativos a los contenidos y habilidades evaluados por el curriculum, cómo estos se reflejan en una tabla de especificaciones, y cómo ésta se utiliza para elaborar ítems, preguntas, o tareas que son luego incluidos en las pruebas.

La documentación técnica sobre alineamiento de programas nacionales de evaluación está en general poco sistematizada. A diferencia de la literatura académica, usualmente no incluye un claro marco conceptual, metodologías, criterios ni procedimientos formales para revisar y evaluar el alineamiento del curriculum, estándares de desempeño, y pruebas.

En las tres metodologías aquí presentadas se integran elementos cuantitativos y cualitativos para revisar y mantener el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas estandarizadas. No hay una separación drástica entre lo cuantitativo y lo cualitativo. Esta integración entre lo cuantitativo y lo cualitativo permite un análisis más global y coherente del alineamiento. De hecho, la práctica en alineamiento contemporáneo no reconoce diferencias paradigmáticas entre los llamados métodos cuantitativos y cualitativos: en ambos métodos se sigue una misma lógica variando únicamente el tipo de evidencia empírica sobre la cual se basa el trabajo. Todo método y procedimiento en alineamiento se sustenta en uso de evidencia empírica, siguiendo procedimientos rigorosamente documentados con evaluaciones públicas de las limitaciones de la evidencia en términos de confiabilidad y validez.

En la literatura especializada sobre alineamiento, la distinción entre el componente cualitativo de los estándares de desempeño (descripciones) y el cuantitativo (puntos de corte) no se traduce en el uso de metodologías cualitativas para el uno y cuantitativas para el otro. Más bien se integran elementos cuantitativos y cualitativos para velar por el alineamiento de ambos componentes. Por ejemplo, para juzgar el alineamiento de las descripciones, chequeando que cada contenido y habilidad descrito tenga un correlato en los ítems de las pruebas (elemento cuantitativo), y haciendo un juicio de valor respecto de qué tan bien la descripción refleja el conjunto de ítems evaluados (elemento cualitativo).

Finalmente, es necesario reirterar que el desarrollo técnico sobre alineamiento es fecundo al abordar la relación entre pruebas y currículum, pero más escaso cuando se reemplaza uno de estos componentes por estándares de desempeño, especialmente en cuanto a su componente

cualitativo, ya que el componente cuantitativo es abordado por el desarrollo de diferentes metodologías para establecer puntajes de corte .

2.1.2 Implementación de estudios de alineamiento

La literatura en medición educativa y los estándares de evaluación de AERA/APA/NCME (2014) establecen que el alineamiento entre pruebas estandarizadas y el curriculum (estándares de contenidos), forman parte de la validación de constructo que se debe establecer para cada iteración de cada prueba, por tanto, forma parte del diseño y construcción de las pruebas cada vez que se administran.

Hay acuerdo en que estos estudios deben realizarse regularmente y por una entidad externa, aun cuando el curriculum no haya cambiado. Se recomienda hacer un estudio de aliniemiento en profundidad cada vez que haya revisiones sustanciales al curriculum existente, o se apruebe un nuevo curriculum.

La realización de estudios de alineamiento es un fenómeno relativamente nuevo, que comenzó aproximadamente en el año 2000 en EEUU. En este país, los estudios de alineamiento entre las pruebas y el curriculum suelen ser licitados a contratistas externos. A nivel estatal, Achieve ha tenido un rol importante en la realización de estudios de alineamiento entre el curriculum estatal, las pruebas estatales, y otros marcos de evaluación interestatatales (ej., Common Core). A nivel federal, AIR (American Institutes for Research[2]) ha jugado un rol clave en la realización de estudios comparativos y de alineamiento entre el NAEP, las evaluaciones internacionales de PISA y TIMSS, y nuevos marcos de evaluación interestatatales (ej. Common Core, 21st Century Skills, y Next Generation Science Standards)[3].

Los estudios de alineamiento entre el NAEP y otras evaluaciones se llevan a cabo a distintos niveles: (1) comparaciones entre marcos de evaluación de distintas pruebas, (2) comparaciones entre ítems de las distintas pruebas, (3) alineamiento entre los ítems de una prueba y el marco de evaluación de la otra. Así, por ejemplo, se han realizado estudios comparando los ítems NAEP y PISA, y estudios clasificando los ítems de NAEP en el marco de evaluación de PISA. Estos estudios han permitido interpretar mejor las diferencias en los resultados entre las distintas

[2] http://www.air.org/topic/p-12-education-and-social-development/international-comparisons-education
[3] Ver por ejemplo: http://www.air.org/news/press-release/new-study-examines-alignment-between-naep-and-common-core-state-standards-4th-8th

evaluaciones. También han informado decisiones sobre el qué y cómo evaluar distintas áreas disciplinarias[4].

2.2 DIAGNÓSTICO DE PROCEDIMIENTOS DE ALINEAMIENTO EN CHILE

2.2.1 Contexto

La evaluación nacional SIMCE se rige por los lineamientos del Sistema de Aseguramiento de la Calidad de la Educación Parvularia, Básica y Media, el cual fue creado mediante la promulgación de la Ley 20529 de agosto de 2011 (Ley SAC). El SAC establece como un deber del Estado propender a asegurar una educación de calidad y equidad, entendiendo por esta última el que todos los alumnos tengan las mismas oportunidades de recibir una educación de calidad.

La Ley SAC crea dos nuevos organismos: la Superintendencia de Educación y la Agencia de Calidad. El SIMCE pasó a estar a cargo de la Agencia de Calidad de la Educación, y a interactuar con cuatro instituciones claves: Ministerio de Educación (MINEDUC), Superintendencia de Educación, Consejo Nacional de Educación y Agencia de Calidad de la Educación. Esta última está a cargo de aplicar el Plan de Evaluaciones Nacionales e Internacionales, el cual es diseñado por el MINEDUC y aprobado por el Consejo Nacional de Educación.

Antes de la promulgación de la Ley SAC, la Unidad de Curriculum y Evaluación (UCE) del MINEDUC tenía a su cargo la evaluación nacional (pruebas SIMCE) y el desarrollo del curriculum. En ese entonces los estándares de desempeño (o Estándares de Aprendizaje) asociados a las pruebas SIMCE eran desarrollados por el equipo de evaluación de la UCE, mientras que los estándares de contenido (el curriculum nacional), estaban a cargo del equipo de curriculum de la UCE.

En este nuevo contexto normativo e institucional, el desarrollo de estándares de desempeño continúa a cargo de la UCE, siendo esta unidad la responsable de definir tanto su componente cuantitativo como el cualitativo. Los estándares de desempeño son presentados por el MINEDUC al Consejo Nacional de Educación y, luego de que este los aprueba, la Agencia de Calidad de la Educación es la encargada de verificar su cumplimiento en el sistema escolar.

[4] Para más información sobre estudios de alineamiento entre NAEP y las evaluaciones internacionales, ver: https://nces.ed.gov/nationsreportcard/about/international.aspx

La UCE es también responsable de asegurar el alineamiento entre el curriculum nacional, los estándares de desempeño, y las pruebas SIMCE. Este alineamiento en condición necesaria para que los resultados de las pruebas SIMCE puedan ser interpretados como reflejando distintos niveles de logro de los aprendizajes esperados, según lo indicado en el curriculum nacional.

Las pruebas SIMCE se aplican según el Plan de Evaluaciones Nacionales e Internacionales. Las pruebas SIMCE abordan gran parte de las áreas del currículum vigente. Sus cuestionarios indagan sobre información del entorno del estudiante, relevante para comprender los niveles de aprendizaje observados.

Los cuestionarios SIMCE indagan sobre información del entorno del estudiante, relevante para comprender los niveles de aprendizaje observados. Los cuestionarios recogen información sobre los "otros indicadores de calidad educativa", tales como la autoestima académica y la motivación escolar, el clima de convivencia, la participación y formación ciudadana, hábitos de vida saludable, asistencia y retención escolar, equidad de género y titulación técnico-profesional. (Agencia de Calidad de la Educación, 2015).

De acuerdo al artículo 37 de la Ley General de Educación (Nº 20370), la medición debe verificar el grado de cumplimiento de los objetivos generales del curriculum a través de la medición de Estándares de Aprendizaje (o estándares de desempeño). Es así que los resultados de las pruebas son reportados en cuanto al cumplimiento de esos estándares, los que reflejan una descripción de lo que los estudiantes deben saber y poder hacer para demostrar el cumplimiento de los objetivos de aprendizaje estipulados en el curriculum vigente. Al mismo tiempo, la Ley SAC indica que los estándares de desempeño tienen una vigencia de 6 años desde su publicación.

Las pruebas censales tienen consecuencias sobre las instituciones escolares de acuerdo a lo que se estipula en la Ley SAC. De acuerdo a esta ley, las escuelas y liceos son clasificados en cuatro categorías de desempeño: Alto, Medio, Medio-bajo o Insuficiente. La clasificación se basa en la distribución de los estudiantes en los estándares de desempeño, y en otros diez indicadores complementarios (como, por ejemplo, clima de convivencia escolar, hábitos de vida saludable o tendencia en el tiempo del puntaje Simce). Los estándares de desempeño reciben la mayor ponderación (67% de la clasificación) para efectos de la clasificación. La clasificación también toma en consideración las características socioeconómicas de los estudiantes.

La clasificación de las instituciones escolares tiene consecuencias fuertes. La clasificación está asociada a reconocimientos y sanciones que pueden llegar hasta el cierre del establecimiento. Por ejemplo, tal como se estipula en artículo 31 de Ley SAC, si un establecimiento educacional subvencionado o que reciba aportes del Estado luego de cuatro años se mantiene en la

categoría de Desempeño Insuficiente -considerando como único factor el grado de cumplimiento de los estándares de aprendizaje- el establecimiento educacional perderá el reconocimiento oficial al término del respectivo año escolar.

2.2.2 Metodología y procedimientos para asegurar el alineamiento entre curriculum nacional, estándares de desempeño y pruebas SIMCE.

En primer lugar, es necesario notar es que en Chile no existe un documento oficial que formalice los procedimientos de alineamiento entre el curriculum nacional, los estándares de desempeño y las pruebas SIMCE. La información que se presenta a continuación fue obtenida a través de: (a) entrevista a equipo de estándares de aprendizaje de la UCE, (b) revisión de documentos técnicos que respaldan el desarrollo de los estándares de aprendizaje, y (c) revisión de documentos técnicos que respaldan el desarrollo de las pruebas SIMCE.

A continuación, se revisan la metodología y procedimientos actualmente utilizados en Chile para asegurar el alineamiento entre:

- Dimensión 1: Alineamiento entre el curriculum y los estándares de desempeño
- Dimensión 2: Alineamiento entre el curriculum y las pruebas
- Dimensión 3: Alineamiento entre los estándares de desempeño y las pruebas

Dimensión 1: Alineamiento entre el curriculum nacional y los estándares de desempeño

De acuerdo al artículo 3 de la Ley SAC, los estándares de desempeño (o de aprendizaje) deben estar referidos al curriculum nacional vigente. Tal como se señala en los fundamentos de los estándares de II medio (MINEDUC, 2014), los estándares de desempeño se elaboran basándose en el curriculum vigente para el período evaluado por el SIMCE en el grado correspondiente. En aquellas situaciones en las que los alumnos han sido expuestos a dos curriculum durante el período a evaluar, se utilizan ambos para elaborar los estándares de desempeño. De esta manera los estándares de desempeño pueden continuar vigentes una vez que se produzca el cambio curricular. En estos casos, los estándares de desempeño incluyen las habilidades y conocimientos comunes a los curriculum que han estado vigentes en el período cursado por los estudiantes. Además, cuando resulta necesario, se agregan las habilidades y conocimientos que se introducen en el nuevo curriculum, explicitando que pasarán a formar parte de los estándares de desempeño una vez que las pruebas SIMCE comiencen a evaluarl el nuevo curriculum.

Según lo señalado en los fundamentos que respaldan el desarrollo de los estándares de II medio, los estándares son desarrollados de acuerdo a un enfoque mixto, que combina tanto la expectativa teórica como la evidencia empírica respecto de los logros demostrados por los estudiantes chilenos. “La adopción de este enfoque implica que el trabajo sea realizado en dos etapas sucesivas, en las cuales se delega la tarea de fijar la exigencia de los estándares a un grupo de especialistas. En una primera etapa, el grupo de especialistas define un “deber ser” a priori, o exigencia teórica, que corresponde a una expectativa de lo que debiera establecerse para cada Nivel de Aprendizaje basada en lo estipulado en el curriculum vigente. Luego, en una segunda etapa, los especialistas contrastan dicho “deber ser” con evidencia empírica de lo que los estudiantes realmente saben y son capaces de hacer, de manera de establecer Estándares cuya exigencia sea realista. De esta forma, se busca que los Estándares de Aprendizaje sean elaborados considerando tanto lo estipulado en el curriculum vigente como los aprendizajes reales de los estudiantes en términos de la dificultad real o empírica asociada a los aprendizajes exigidos” (p. 48).

Un primer paso para la elaboración de los estándares de desempeño consiste en describir los requisitos mínimos teóricos para alcanzar cada Nivel de Aprendizaje. Estos requisitos mínimos se traducen en un listado de indicadores con una lógica de progresión. Es decir, si un estudiante logra un nivel superior se supone que también ha alcanzado la exigencia para los niveles anteriores. Estos requisitos mínimos incluyen aquellos conocimientos y habilidades considerados “aprendizajes terminales” en cada asignatura y grado evaluado. De este modo, los requisitos mínimos asociados a cada nivel no constituyen un listado exhaustivo de todo lo que el curriculum explicita que un estudiante debe saber y poder hacer, sino lo esencial que se exige para alcanzar un determinado Nivel de Aprendizaje.

Los estándares de desempeño tienen una función de responsabilización de las escuelas por los aprendizajes de sus estudiantes. En este sentido, deben describir un “mínimo irrenunciable” que se establece como expectativa nacional. Esto tiene como consecuencia que, desde la perspectiva del MINEDUC, no deben quedar fuera de los estándares de desempeño elementos centrales del curriculum nacional, aún cuando éstos no sean evaluados de manera sistemática por las pruebas nacionales.

Los requisitos mínimos teóricos incluyen todos aquellos conocimientos y habilidades esenciales que pueden ser medidos por pruebas censales que involucran preguntas de respuesta abierta o cerrada, y que pueden ser contestadas en una prueba de lápiz y papel como el SIMCE. También incluyen conocimientos y habilidades que pueden ser medidos utilizando grabaciones, regla, calculadora u otros elementos, aunque éstos no sean actualmente medidos en las pruebas SIMCE. Por ejemplo, los requisitos mínimos de matemática de 4to grado se refieren a

conocimientos y habilidades que requieren del uso de la regla, aunque las pruebas SIMCE actualmente no incluyen preguntas que requieran usar regla. Sólo se excluyen del listado de requisitos mínimos aquellos contenidos, habilidades y actitudes que serían muy difíciles de evaluar en una prueba censal (por ejemplo, cálculo mental).

Los requisitos mínimos teóricos son definidos por el equipo de Estándares de Aprendizaje en conjunto con el equipo elaborador del curriculum, basándose en el curriculum vigente. Luego, estos requisitos mínimos son contrastados con evidencia empírica del nivel de aprendizaje de los estudiantes, tanto nacional como internacional. De esta manera se elaboran descripciones de estándares de desempeño que sean efectivamente alcanzables por los estudiantes chilenos, sin dejar de ser desafiantes y sin dejar de estar alineadas a las exigencias del curriculum nacional. Las descripciones así obtenidas son contrastadas con diversos especialistas y docentes, obteniéndose así una versión final.

De acuerdo a la entrevista realizada a profesionales del equipo de la UCE, el alineamiento entre los estándares de desempeño y el curriculum se da naturalmente. Esto ocurre dado que los mismos profesionales a cargo del desarrollo del curriculum nacional están a cargo de elaborar las descripciones cualitativas de los estándares de desempeño. Este alineamiento se entiende como logrado si las descripciones de los estándares son capaces de dar cuenta de los pasos o etapas previas (ya sean contenidos o habilidades a dominar) que se deben cumplir para alcanzar los objetivos del curriculum correspondientes al grado que será evaluado.

Desde 2010 se realizan procedimientos de validación con externos a la UCE (docentes y especialistas en la disciplina) donde se consulta acerca de la relación entre las descripciones de los estándares de desempeño y el curriculum nacional. Si bien estas validaciones no tienen por propósito verificar el alineamiento curricular, permiten constatar que los estándares de desempeño son comprensibles y que efectivamente corresponden a pasos previos para el cumplimiento de los objetivos del curriculum. Profesionales de la UCE participan de estas validaciones, velando por que los cambios sugeridos estén alineados al curriculum.

Finalmente, el Consejo Nacional de Educación debe evaluar y aprobar los estándares de desempeño antes de su publicación. Esta evaluación la realiza a través de la contratación de especialistas en las disciplinas e incluye, entre otros criterios, la verificación de alineamiento entre estándares de desempeño y el curriculum[5]:

[5] Extraído del Acuerdo 075/2012 del Consejo Nacional de Educación, disponible en: http://www.cned.cl/public/Secciones/SeccionEducacionEscolar/acuerdos/Acuerdo_075_2012.pdf

> "*1.1. Los estándares son coherentes con el marco normativo chileno referido a aprendizajes. Este criterio considera aspectos tales como: los estándares de aprendizaje son consistentes con los principios y objetivos generales señalados en la LGE (LGE. Artículo 37, párrafo 1) y permiten verificar el grado de cumplimiento de ellos; los estándares propuestos están alineados con y cubren adecuadamente los objetivos de las Bases Curriculares para cada una de las asignaturas o disciplinas y niveles que serán evaluados. Son completos, sin llegar a incluir detalles innecesarios ni omitir dimensiones esenciales en el aprendizaje de la disciplina; los estándares cubren los Objetivos de Aprendizaje y/o Ejes Temáticos valorados en las Bases Curriculares, sin exclusiones; la eventual exclusión de asignaturas o áreas curriculares es justificada consistentemente en la propuesta o se compromete un plazo para incluirlas; el nivel de exigencia de los estándares de aprendizaje es consistente con los objetivos de aprendizaje que, según las bases curriculares, deben ser cubiertos en el ciclo o nivel evaluado.*"

Para cumplir con este criterio, los resultados del alineamiento entre los estándares de desempeño y el curriculum se presentan al Consejo Nacional de Educación a través de una tabla que señala la relación de elementos de los estándares de desempeño con los objetivos del curriculum. Del mismo modo, se entrega una tabla donde se justifican aquellos elementos del curriculum nacional que no se incluyen en los estándares de de desempeño.

Dimensión 2: Alineamiento entre el curriculum nacional y las pruebas SIMCE

De acuerdo a la entrevista con el equipo de estándares de aprendizaje de la UCE, el trabajo coordinado entre el equipo a cargo de la elaboración de las pruebas SIMCE y el equipo a cargo del desarrollo del curriculum nacional es fundamental para velar por el alineamiento entre el curriculum nacional y las pruebas. El instrumento central para velar por el alineamiento entre estos dos componentes son las especificaciones técnicas de las pruebas. Éstas usan como principal insumo del curriculum vigente para los grados que corresponde evaluar.

En el informe técnico de SIMCE 2013 (Agencia de Calidad de la Educación, 2015) se señala que es fundamental que las pruebas sean representativas del curriculum nacional vigente al momento de la aplicación. Para ello, el Departamento de Construcción de Pruebas (DCP) de la Agencia de Calidad procura que las pruebas cubran, en extensión y profundidad, los objetivos y contenidos curriculares de cada uno de los grados evaluados, trabajando de manera coordinada con la UCE del MINEDUC. Un primer paso es delimitar aquellos objetivos y contenidos que son susceptibles de evaluar en las pruebas SIMCE.

Una vez delimitado el curriculum a evaluar, el equipo DCP elabora las especificaciones técnicas de las pruebas, donde la representatividad y cobertura curricular quedan plasmadas a partir de la operacionalización de los contenidos en la forma de Objetivos de Evaluación (OE) y su correspondiente matriz de evaluación teórica para cada nivel y área a evaluar.

Entre las actividades que se describen en el informe técnico del SIMCE 2013 para asegurar alineamiento entre las pruebas y el curriculum nacional están la formación de equipos disciplinares por área disciplinaria, quienes realizan un análisis curricular y definen objetivos de evaluación a medir. A partir de este análisis se genera un documento preliminar por área, el que es aprobado por la jefatura del DCP y luego por la UCE del MINEDUC. La UCE actúa como contraparte externa experta, retroalimentando esta etapa del proceso. Así, se generan las versiones finales de cada documento de especificaciones técnicas para la elaboración de ítems por área y grado.

Las especificaciones técnicas de cada prueba incluyen los siguientes contenidos: (a) una introducción que explicita los referentes curriculares, así como una breve descripción de los ejes temáticos y dominios cognitivos que serán evaluados; (b) las matrices de evaluación que indican los pesos relativos atribuibles a cada uno de los ejes temáticos; y (c) los objetivos de evaluación con su respectivo eje de contenido y de habilidad.

Es importante considerar que, tal como se señala en el informe técnico de SIMCE 2013, existen constructos que, por el formato de la prueba, las características de los objetivos y contenidos de aprendizaje, y el proceso de implementación curricular, resultan complejos o imposibles de medir, siendo estos un desafío permanente en el desarrollo de las pruebas nacionales.

Dimensión 3: Alineamiento entre estándares de desempeño y las pruebas SIMCE

De acuerdo a los antecedentes aportados en la entrevista con los profesionales del equipo de elaboración de estándares de la UCE, no existe un procedimiento específico para verificar el alineamiento entre los estándares de desempeño y la prueba. Este alineamiento se da por cumplido al estar la prueba basada en el curriculum nacional, al igual que los estándares de desempeño.

Desde una perspectiva psicométrica, el alineamiento entre los estándares de desempeño y la prueba se resguarda a través de los procedimientos que se emplean para establecer los puntajes de corte. Esto es así dado que la descripción de los estándares de desempeño se traduce en un puntaje en la prueba, el que se establece a través de un mecanismo válido y confiable (dependiendo de la prueba que se trate, se emplea método Bookmark o Angoff).

De acuerdo con lo señalado en la entrevista con los profesionales de la UCE, hasta ahora los estándares de desempeño no se emplean como insumo para la elaboración de la prueba. Esto es algo que se está intentando cambiar desde la UCE, para lo cual se requiere evidencia de sistemas de evaluación de otros países donde esto sí se haga. No obstante, según la Ley SAC, la prueba debe evaluar el cumplimiento de los estándares de desempeño y no el curriculum nacional. Pareciera entonces relevante incluir los estándares de desempeño como un insumo más a la hora de diseñar especificaciones e ítems de las pruebas.

Tampoco se cuenta con un marco de evaluación de las pruebas que alimente la elaboración de los estándares de desempeño. Esto es, un documento que describa en términos generales qué evalúan las pruebas, y cómo (formato de pruebas, tipos de preguntas). ítems

Al respecto, en el informe técnico del SIMCE 2013 se señala que "las pruebas SIMCE, sin excepción para el proceso 2013, deben contar con la información más completa posible respecto de los puntajes de corte, de modo de incrementar la precisión con la cual se entrega la clasificación para cada nivel evaluado" (p. 7). Sin embargo, no fue posible encontrar información que describa cómo se logra esto de manera detallada.

Un tema que concita el interés del equipo de estándares de la UCE es indagar sobre la estabilidad de los puntajes de corte en el tiempo. ¿En qué medida los puntajes de corte son reflejo de los estándares de desempeño si, en el tiempo, el comportamiento de los ítems puede cambiar de manera significativa? ¿Bajo qué premisas se podrían cambiar los puntajes de corte sin necesariamente cambiar la descripción de los estándares de desempeño?

2.2.3 Preguntas para orientar el estudio de alineamiento

A partir de la revisión de la metodología y procedimientos para asegurar el alineamiento entre el curriculum nacional, estándares de desempeño, y pruebas SIMCE, los siguientes puntos resultan de especial interés para orientar este estudio:

- Alineamiento entre estándares de desempeño y las pruebas SIMCE. Al analizar la evidencia disponible acerca de los procedimientos de alineamiento entre curriculum, estándares de desempeño, y pruebas SIMCE, pareciera que una de las principales áreas en donde se requiere afinar y formalizar los procesos, es respecto del alineamiento entre estándares de desempeño y las pruebas. En las otras áreas pareciera que el nivel de formalización y desarrollo de los procesos no genera controversia y se encuentra detallado en documentos oficiales.

- Roles de las distintas instituciones involucradas en los procedimientos de alineamientos. El alineamiento entre estándares de desempeño y las pruebas SIMCE se verifica inmediatamente en la evaluación realizada por el Consejo Nacional. En este sentido ¿qué institución debe ser la principal garante del alineamiento? ¿Se deben fijar criterios para juzgar el alineamiento?

- Estabilidad de los puntajes de corte en el tiempo. Los cambios de comportamiento de los ítems en el tiempo llevan a preguntarse por la validez de los puntos de corte para diferenciar entre estudiantes que alcanzan un nivel y los que no. Es entonces importante indagar sobre los criterios que justifican revisar y actualizar los puntos de corte asociados a los estándares de desempeño.

- Utilización de los estándares de desempeño como insumo para la elaboración de las pruebas. Esto es, como referentes de contenidos y habilidades, y como referentes para fortalecer las pruebas alrededor de los puntajes de corte.

- Mecanismos de coordinación para el trabajo entre especialistas a cargo de la elaboración de estándares y quienes están a cargo de la elaboración de las pruebas.

3. METODOLOGÍA DE TRABAJO LLEVADA A CABO

Se llevó a cabo un estudio descriptivo sobre los procesos y metodologías para evaluar y mantener el alineamiento de pruebas estandarizadas con el curriculum (estándares de contenido) y con los estándares de desempeño, en los países participantes de la prueba PISA 2012, más Ecuador.

El estudio incluyó cuatro componentes:

(1) Una revisión de publicaciones académicas sobre metodologías de alineamiento.

(2) Una revisión documental sobre el curriculum, estándares de desempeño y pruebas estandarizadas, en los países participantes en PISA 2012, más Ecuador.

(3) Un análisis en profundidad para tres sistemas educativos.

(4) Un diagnóstico o línea de base sobre la metodología actualmente utilizada en el MINEDUC para procurar el alineamiento del curriculum, estándares de desempeño, y pruebas estandarizadas.

A continuación, se especifica cada componente.

3.1 REVISIÓN DE PUBLICACIÓNES ACADÉMICAS SOBRE ALINEAMIENTO

Para este componente se emplearon diversos recursos o motores de búsqueda de publicaciones académicas referidas a procesos de alineamiento entre curriculum, estándares de desempeño y pruebas estandarizadas, tales como Google Scholar, AERA (American Educational Research Association) Catalogue[6], y la base de datos ERIC[7].

Luego de identificar las publicaciones fundantes y aquellas más actualizadas en el área, se procedió a revisarlas para seleccionar aquella información que permitiera generar una descripción de lo métodos o procedimientos de alineamiento más empleados o que pudieran orientar recomendaciones prácticas para el sistema educacional chileno.

El producto de este componente puede ser revisado en el Capítulo de “Antecedentes sobre el tema” de este informe.

[6] AERA 2016 Catálogo:
[7] Base de datos ERIC: http://eric.ed.gov/

3.2 REVISIÓN DOCUMENTAL SOBRE CURRICULUM, ESTÁNDARES DE DESEMPEÑO Y PRUEBAS ESTANDARIZADAS PARA DESARROLLO DE CATASTRO

Para este componente, se realizó una revisión documental preliminar sobre desarrollo de estándares de desempeño y pruebas nacionales o estatales. El objetivo de esta tarea fue conocer si existen estos elementos mínimos para seleccionar a los países donde sería posible indagar las metodologías de alineamiento utilizadas entre curriculum, estándares de desempeño y pruebas estandarizadas.

Esta revisión preliminar se hizo en base a los sitios web oficiales de los sistemas educacionales de los países de interés (países PISA 2012 más Ecuador), de las bases de datos o reportes de Eurydice, el *Center on International Education Benchmarking* del NCEE, las OECD Reviews of Evaluation and Assessment in Education, el *Catalogue of Learning Assessments* del Instituto de Estadísticas de UNESCO, SABER-Student Assessment del Banco Mundial, y de los siguientes documentos: Comisión Europea/EACEA/Eurydice (2015); Phelps (2014); Cox, C., Meckes, L. & De Padua, E. (2013); Parveva, De Coster & Noorani (2009).

A partir de la revisión documental descrita anteriormente, se realizó una selección de países viables de incluir en el estudio de acuerdo a criterios como:

- si tienen un sistema nacional o estatal de evaluación
- si tienen estándares de desempeño asociados a dicha evaluación
- si tienen información sistematizada y accesible en español, francés, inglés, o portugués (idiomas que maneja el equipo de investigación), u otros idiomas a través de los cuales el equipo pueda acceder a la información (ej., vía asistente o con traducciones).

Como resultado de esta revisión documental, se seleccionaron 52 países viables de incluir en el estudio de alineamiento. También se seleccionaron aquí aquellos países, para los cuales no se tenía información suficiente sobre sus sistemas de evaluación[8].

[8] En el informe 1 de este estudio se incluyó una lista de 49 sistemas educativos, contando al Reino Unido como un sistema. A partir del Informe 2 se consideran 52 sistemas educativos, ya que se cuentan por separado a Escocia, Gales e Irlanda del Norte.

Tabla 2: Listado de países viables de incluir en el estudio de alineamiento.

1. Alemania	19. Eslovaquia	37. Japón
2. Argentina	20. Eslovenia	38. Jordania
3. Australia	21. España	39. Luxemburgo
4. Austria	22. Estados Unidos	40. Malasia
5. Bélgica Flamenca	23. Estonia	41. México
6. Brasil	24. Finlandia	42. Noruega
7. Bulgaria	25. Francia	43. Nueva Zelanda
8. Canadá	26. Gales	44. Perú
9. Catar	27. Grecia	45. Polonia
10. China (Shanghai)	28. Holanda	46. Portugal
11. Colombia	29. Hong Kong, China	47. República Checa
12. Corea del Sur	30. Hungría	48. Rusia
13. Costa Rica	31. Indonesia	49. Singapur
14. Croacia	32. Inglaterra	50. Turquía
15. Dinamarca	33. Irlanda, Rep.	51. Uruguay
16. Ecuador	34. Irlanda del Norte	52. Vietnam
17. Emiratos Árabes	35. Islandia	
18. Escocia	36. Italia	

En conjunto con lo anterior, se excluyeron "a priori" del estudio a 15 países, ya sea por dificultades para acceder a la información (por disponibilidad o por idioma), o porque no contaban con una evaluación asociada a estándares de desempeño (Tabla 3).

Tabla 3: Lista de países excluidos "a priori" del estudio de alineamiento.

País	**Sistema de evaluación**	**Estándares de desempeño**	**Información en inglés o español**
1. Albania	Sin información	Sin información	Sin información
2. Chipre	No	No	Si
3. Israel	No	Sin información	Sin información
4. Kazajistán	Si	Sin información	Sin información
5. Liechtenstein	No	No	Sin información
6. Lituania	Si	Sin información	Sin información
7. Macao, China	No	No	No
8. Montenegro	Sin información	Sin información	Sin información

9. Rumania	Sin información	Sin información	Sin información
10. Serbia	Si	Sin información	Sin información
11. Suecia	Si	No	Si
12. Suiza	No	No	No
13. Tailandia	Si	No	No
14. Taiwán	Sin información	Sin información	No
15. Túnez	Si	Sin información	Sin información

Con el listado de 52 países viables de incluir en el estudio de alineamiento se hizo una revisión documental más exhaustiva de sitios web institucionales u otros medios con información sobre implementación de procedimientos para resguardar el alineamiento entre curriculum nacional, estándares de desempeño y pruebas nacionales o estatales.

Junto con esta revisión, se contactó a profesionales de las instituciones correspondientes para recolección de documentación adicional sobre procedimientos para asegurar el alineamiento entre curriculum, estándares de desempeño y pruebas nacionales o estatales. Para ello, se elaboró un "email tipo" (ver Anexo 1), que fue enviado a informantes claves de 37 países para los que fue posible conseguir un email de contacto. También se establecieron contactos informales (por teléfono o en persona) con algunos informantes claves. Se pudo obtener información de en torno a la mitad de los países contactados.

Finalmente, se sistematizó la información recolectada a en un catastro, en donde se clasificó a los países según sus métodos y procedimientos de alineamiento. Se utilizaron las siguientes categorías:

- Contexto: características generales de la evaluación
  - Nombre programa de evaluación
  - La evaluación reporta resultados usando estándares de desempeño (ej. Niveles Avanzado, Básico, etc.)
  - Institución(es) a cargo de la evaluación
  - Propósitos, Usos, o Consecuencias
  - Grados evaluados
  - Censo/muestra
  - Áreas disciplinarias evaluadas
  - Características de las pruebas (formato, escala puntajes, etc.)
  - Resultados (ej. ¿a nivel escuela? ¿públicos?)
- Alineamiento: metodologías, procesos, y aspectos transferibles a Chile
  - Hay una matriz de referencia/ marco de evaluación que especifique qué se evalúa?

- Alineamiento entre curriculum y las pruebas
- Alineamiento entre las pruebas y estándares de desempeño
- Alineamiento entre curriculum y estándares de desempeño
- Gobierno y responsabilidades en alineamiento
- Otros elementos que podrían resultar interesantes para el sistema chileno

- Recursos: dirección de sitios web de programas de evaluación.

Finalmente, de acuerdo a la cantidad y calidad de la información encontrada , en el catastro se clasificó al país o sistemaeducativo en una de las siguientes categorías::

(1) Con información técnica suficiente: países en los que se encontró información técnica más elaborada sobre alineamiento. Usualmente incluye información de alineamiento entre curriculum y las pruebas, curriculum y estándares de desempeño, estándares de desempeño y las pruebas, o procedimientos generales.

(2) Con información mínima: países en los que se encontró mínima información sobre alineamiento. Usualmente esta información esta referida al alineamiento entre el curriculum y pruebas (ej., tablas de especificaciones).

(3) Sin información: países que no tienen estándares de desempeño asociados a las pruebas o en los que no se encontró información sobre alineamiento.

Los resultados de este componente se presentan en archivo Excel <A_Catastro.xls> adjunto a este informe.

3.3 ANÁLISIS EN PROFUNDIDAD PARA TRES SISTEMAS EDUCATIVOS

Además de la recolección de información a través de la revisión documental descrita en el apartado anterior, se realizaron tres casos de estudio en profundidad:

- Caso en profundidad 1: National Assessment of Educational Progress - NAEP (Evaluación Nacional del Progreso Educativo), EEUU.
- Caso en profundidad 2: Ontario Provincial Assessment Program (Programa de Evaluación de la Provincia de Ontario), Canadá.
- Caso en profundidad 3: Plan Nacional de Evaluación de los Aprendizajes (PLANEA), México.

El objetivo de este análisis es contar con información en profundidad de tres sistemas educativos sobre la metodología y procedimientos de alineamiento entre curriculum, estándares de desempeño, y pruebas.

Los casos de estudio fueron seleccionados a partir de la revisión realizada en la fase de estudio documental, velando por la relevancia, aplicabilidad y similitud de los casos con Chile. Se veló por cumplir con los siguientes criterios de selección:

- Programas que cuenten con definiciones de estándares de contenido (curriculum) y estándares de desempeño o similares.
- Programas que tengan procedimientos bien establecidos de alineamiento.
- Programas de evaluación con altas consecuencias para los establecimientos escolares.
- Programas que cuenten con información en inglés o español.

Para cada caso en profundidad se realizaron las siguientes tareas:

- Revisión de informes técnicos y sitios web institucionales de sus programas de evaluación de aprendizaje.
- Entrevistas vía email, teléfono, y videoconferencia con personas que vinculadas a cada programa de evaluación. Los nombres y cargo de estas personas pueden consultarse en cada uno de los Casos en Profundidad.

Además, en el caso de NAEP (EEUU), se realizaron múltiples entrevistas en persona con una profesional que ha liderado estudios comparativos y de alineamiento para este programa de evaluación.

Los resultados de este componente se presentan en tres casos en profundidad (ver sección "Descripción y análisis de tres casos en profundidad"). La información de cada caso fue organizada según las siguientes categorías:

- Características generales del programa de evaluación: áreas disciplinarias y niveles considerados en las evaluaciones; características de las evaluaciones (ej., formato, escalas de puntaje, niveles de desempeño, muestra/censo); consecuencias asociadas a las evaluaciones.
- Organismos responsables: (a) del curriculum nacional o estatal, (b) de las pruebas asociadas y (c) del alineamiento entre las diferentes dimensiones estudiadas.
- Alineamiento entre el curriculum y los estándares de desempeño (dimensión 1).
- Alineamiento entre el curriculum y las pruebas (dimensión 2).
- Alineamiento entre los estándares de desempeño y las pruebas (dimensión 3).

3.4 DIAGNÓSTICO O LÍNEA DE BASE

El objetivo de este componente fue elaborar un diagnóstico o "línea de base" describiendo la metodología y procedimientos actualmente utilizados en Chile para alinear las pruebas con estándares de contenido y estándares de desempeño. Los resultados de este diagnóstico orientaron el trabajo de investigación. También orientaron la elaboración de recomendaciones que sean pertinentes para mejorar la forma en que actualmente se hace el alineamiento en Chile.

Para realizar este diagnóstico, el equipo de investigación se entrevistó con la contraparte del MINEDUC, revisó la documentación facilitada por ésta sobre alineamiento (Fundamentos Estándares de Aprendizaje Matemática, Lenguaje y Comunicación: Lectura, II Medio, Anexo 4, y Fundamentos Estándares de Aprendizaje 4º y 8º Básico Anexo 4) y revisó la documentación disponible sobre el tema en el website de la Agencia de Calidad.

Lamentablemente, no fue posible concretar una reunión con la Agencia del Calidad.

Los resultados de este componente se presentan en el Capítulo de "Antededentes sobre el Tema", sección "Diagnóstico de procedimientos de alineamiento en Chile".

4. CATASTRO SOBRE EVALUACIÓN DE ALINEAMIENTO EN DIFERENTES SISTEMAS EDUCATIVOS Y DESCRIPCIÓN Y ANÁLISIS DE TRES CASOS EN PROFUNDIDAD.

4.1 CATASTRO SOBRE EVALUACIÓN DE ALINEAMIENTO EN DIFERENTES SISTEMAS EDUCATIVOS

La Tabla 4 muestra los resultados de la revisión de metodologías de alineamiento hecha en base a la selección de 52 países viables de ser incluidos en este estudio y presentados anteriormente en la sección metodológica de este informe.

La tabla indica que en solo 8 países (15%) se encontró información técnica que trate sobre este tema con suficiente profundidad. En 31 países (60%) no hay información disponible sobre alineamiento. En otros 13 países (25%) la información sobre alineamiento es mínima, y suele estar circunscrita a la tabla de especificaciones de las pruebas. Si Chile tuviera que ser clasificado en esta tabla según la información de alineamiento actualmente disponible online, quedaría clasificado como “con información mínima”.

Tabla 4. Resultados de la revisión de metodologías de alineamiento en 52 países o sistemas educativos.

Clasificación de países o sistemas educativos según información de alineamiento	**Países o sistema educativo**	**Cantidad y porcentaje de países**
Sin información. Países que no tienen estándares de desempeño asociados a las pruebas, o en los que no se encontró información sobre alineamiento.	Argentina*, Bélgica (Fl.)*, Bulgaria*, China (Rep. Popular)*, Costa Rica*, Croacia, Ecuador*, Emiratos Árabes*, Eslovaquia, Eslovenia*, Estonia, Finlandia*, Grecia, Holanda*, Hungría, Hong Kong*, Indonesia*, Irlanda (Rep.), Irlanda del Norte, Islandia*, Italia*, Japón*, Jordania*, Luxemburgo*, Noruega*, Polonia, República Checa, Rusia*, Turquía*, Uruguay*, y Vietnam*	31 (60%)
Con información mínima. Países en los que se encontró mínima información sobre alineamiento. Usualmente esta información estaba referida al alineamiento entre el curriculum y las pruebas (ej., tablas de especificaciones).	Alemania*, Austria*, Brasil*, Catar*, Colombia*, Corea del Sur*, Dinamarca*, Escocia, España*, Francia*, Malasia*, Portugal*, y Singapur	13 (25%)
Con información técnica suficiente. Países en los que se encontró información técnica más elaborada sobre alineamiento. Esta información usualmente incluía información de alineamiento entre (1) curriculum y las pruebas, (2) curriculum y los estándares de desempeño, (3) estándares de desempeño y las pruebas, o (4) procedimientos generales.	Australia, **Canadá (Ontario)***, **EEUU (NAEP)***, Gales, Inglaterra*, **México***, Nueva Zelanda, Perú*	8 (15%)
TOTAL		**52**

* Países contactados (a través de email de contacto inicial, conversaciones telefónicas o en persona) para revisión de metodologías de alineamiento.

En **negrita y subrayado**: Países seleccionados para estudio en profundidad.

Fuente: Excel Catastro de países.

El Catastro de Países (documento Excel <A_Catastro.xls>) presenta información detallada sobre la revisión de alineamiento para los 52 países viables de ser incluidos en este estudio.

4.2 DESCRIPCIÓN Y ANÁLISIS DE TRES CASOS EN PROFUNDIDAD

A continuación, se presentan los tres casos en profundidad seleccionados por el MINEDUC:

- Caso en profundidad 1: National Assessment of Educational Progress - NAEP (Evaluación Nacional del Progreso Educativo), EEUU.
- Caso en profundidad 2: Ontario Provincial Assessment Program (Programa de Evaluación de la Provincia de Ontario), Canadá.
- Caso en profundidad 3: Plan Nacional de Evaluación de los Aprendizajes (PLANEA), México.

Para cada caso, se presentan las características generales del programa de evaluación, los organismos responsables, y los criterios y procedimiento para asegurar el alineamiento de cada dimensión: Curriculum y Estándares de Desempeño (Dimensión 1), Curriculum y Pruebas (Dimensión 2), y Estándares de Desempeño y Pruebas (Dimensión3.).

ALINEAMIENTO ENTRE CURRICULUM, ESTÁNDARES DE DESEMPEÑO, Y PRUEBAS
CASO EN PROFUNDIDAD # 1
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)
Evaluación Nacional del Progreso Educativo, EEUU

1. CARACTERÍSTICAS GENERALES DEL PROGRAMA DE EVALUACIÓN

El *National Assessment of Educational Progress* (NAEP)[9] – Evaluación Nacional del Progreso Educativo) es también conocido como el *Nation's Report Card* ("Reporte Nacional de Notas"). El propósito principal del NAEP es monitorear lo que los estudiantes saben y pueden hacer en distintas áreas disciplinarias. Los "Report Cards" comunican los hallazgos del NAEP y comparan los resultados de la escolarización entre estados, distritos escolares, escuelas públicas y privadas, y distintos estratos socio demográficos; e informan sobre tendencias en el aprendizaje desde 1969.

El NAEP realiza evaluaciones periódicas en 12 áreas disciplinarias, incluyendo matemáticas, lectura, escritura, geografía, historia de los EEUU, educación cívica, economía y las artes. En 2014 administró por primera vez una evaluación en Tecnología y "Engineering Literacy" (Ingeniería). Recientemente ha adoptado un marco de evaluación para futuras evaluaciones en idiomas extranjeros.

Las evaluaciones principales de NAEP se llevan a cabo en 4º, 8º y 12º grado, aunque no se evalúa necesariamente todas las áreas disciplinares en todos los grados. El estudio de tendencias del NAEP – que se basa en marcos de evaluación más antiguos e informa sobre tendencias desde el año escolar de 1969-70, evalúa estudiantes de 9, 13 y 17 años de edad.

Los resultados del NAEP se reportan en tres niveles de desempeño ("achievement levels"). Para cada nivel de desempeño existe una definición de general del nivel (que el National Assessment Governing Board [NAGB][10] denomina "policy definition"), una descripción de los desempeños que involucra, un conjunto de preguntas de evaluación ilustrativas, y un punto de corte. Las definiciones generales para cada estándar de desempeño son:

- Básico: Denota dominio parcial de conocimientos y desempeños que son prerrequisitos y fundamentos para el desempeño competente en el grado y área disciplinar evaluado.
- Competente: Representa desempeño académico solido en cada grado y área disciplinar evaluado. Estudiantes en este nivel han demostrado competencia sobre contenidos y desempeños desafiantes.
- Avanzado: Significa desempeño superior

[9] http://nces.ed.gov/nationsreportcard/

10 El NAGB es responsable del NAEP (ver descripción en la sección 2)

Hay un cuarto nivel por "*default*" que se llama "inferior a básico" que simplemente significa que el estudiante no presenta suficiente evidencia para ser categorizado en el nivel básico.

Los puntajes de las pruebas NAEP se calculan usando Teoría de Respuesta al Ítem (TRI) en una escala usualmente con un rango de 0-500, con Desviación Estándar de 35.

2. ORGANISMOS RESPONSABLES

El National Assessment Governing Board (NAGB, Junta de Gobierno de la Evaluación Nacional) es responsable de establecer las políticas del NAEP. Decide qué áreas disciplinares serán evaluadas, aprueba marcos de evaluación y niveles de desempeño, establece y ejecuta la comunicación de resultados al Congreso y al público. El NAGB tiene la autoridad final acerca de todas las decisiones de diseño del NAEP, inclusive la decisión final acerca de cada uno de los ítems que forman parte de las pruebas.

El NCES (National Center for Education Statistics) es, en la práctica, el organismo responsable de implementar el NAEP. Esto lo hace a través de la ejecución directa y de la subcontratación de tareas claves de la evaluación, tales como diseño de las pruebas y de desarrollo de ítems, puntuación, elaboración de informes, entre otros.

Contratistas externos tienen a cargo las tareas técnicas relativas a la implementación del NAEP. Contratistas "históricos" han sido ETS[11], AIR[12], y Westat[13]. ETS ha tenido a su cargo tareas de escalamiento, equating, procedimientos de estándares de desempeño, y análisis y reporte de resultados. AIR ha estado a cargo del desarrollo de marcos de evaluación y especificaciones técnicas, elaboración y revisión de ítems, y estudios comparativos y de alineamiento, entre otros. Westat ha estado a cargo del muestreo y operación de campo.

3. ALINEAMIENTO ENTRE EL CURRICULUM Y LOS ESTÁNDARES DE DESEMPEÑO (DIMENSIÓN 1)

El referente orientador del NAEP no es un curriculum en específico, sino que son los marcos de evaluación de cada área evaluada, los que son desarrollados por el propio NAEP (NAEP Frameworks)[14]. Esto es así dado que en EEUU no hay un curriculum nacional. Los marcos de evaluación del NAEP procuran ser representativos (en general) del curriculum de cada uno de los estados. Dan cuenta del enfoque conceptual con el que se evaluará la disciplina y de los lineamientos para la evaluación (el "qué y el cómo"). Hay un marco de evaluación por cada área disciplinaria evaluada. Además de los marcos de evaluación, hay especificaciones técnicas para guiar el desarrollo de las pruebas e ítems (ver Sección 4).

[11] https://www.ets.org/k12/naep
[12] http://www.air.org/topic/p-12-education-and-social-development/national-assessment-educational-progress-naep
[13] https://www.westat.com/projects/assessing-american-children%E2%80%99s-educational-progress-naep
[14] https://nces.ed.gov/nationsreportcard/frameworks.aspx

Los marcos de evaluación incluyen los estándares de desempeño ("achievement levels"), los que están alineados con las expectativas del marco de cada área disciplinaria y grado evaluado. Los estándares de desempeño son referentes de lo que los estudiantes deben saber y ser capaces de hacer en las pruebas NAEP en los grados 4, 8, y 12. Los tres niveles de desempeño de NAEP (Básico, Competente, y Avanzado) dan cuenta de "how good is good enough" en cada grado.

Los estándares de desempeño se desarrollan utilizando, primero, un enfoque esencialmente referido a criterio, y luego, integrando elementos normativos. La primera etapa de desarrollo comienza antes de la administración de las pruebas, y vela por el alineamiento de los estándares de desempeño con el marco de evaluación para cada área disciplinaria y grado evaluado. El NAGB elabora definiciones generales de los niveles de desempeño ("policy definitions"), tomando como único insumo las expectativas del marco de evaluación.

Luego, comisiones ad hoc elaboran descripciones preliminares de los estándares de desempeño para orientar el desarrollo de ítems. Estas descripciones preliminares se desarrollan tomando como insumo las definiciones generales de los estándares de desempeño, el marco de evaluación, y las especificaciones técnicas de cada área disciplinaria y grado. Las descripciones preliminares son fundamentales para asegurar el alineamiento de las pruebas con los estándares de desempeño (ver Sección 5).

La segunda etapa de desarrollo de los estándares de desempeño comienza después de administradas las pruebas. Los estándares de desempeño se fijan velando por que reflejen las descripciones preliminares (elemento criterial), y estén a una desviación estándar de distancia aproximadamente en la escala de puntajes (elemento normativo). Sin embargo, no se fijan las distancias entre los niveles, por lo que hay variaciones en los grados y áreas disciplinares evaluadas.

Las descripciones definitivas asociadas a los puntos de corte de los estándares de desempeño dan cuenta de las expectativas señaladas en los marcos de evaluación para el área disciplinaria y grado evaluado, así como de la evidencia empírica recogida en las pruebas. Esta intersección permite asegurar que los estándares de desempeño están alineados con los marcos de evaluación.

Comisiones ad hoc participan en el establecimiento de los estándares de desempeño. Las descripciones definitivas se adoptan después de consultas públicas y luego se convocan paneles de expertos para revisarlos, para considerar la información de las consultas públicas, y para hacer recomendaciones finales al NAGB. En estos paneles, los miembros revisan además datos de pruebas definitivas y pilotos, y comparaciones con los Marcos del NAEP en un proceso estructurado llevado a cabo por una empresa contratista.

El NAGB es el encargo de aprobar los estándares de desempeño (puntos de corte y descripciones definitivas). Los estándares de desempeño son relativamente estables, y tienen

una vigencia de 10 años aproximadamente. Los estándares de desempeño sólo se modifican cuando cambian los marcos de evaluación.

Los estándares de desempeño tienen que equilibrar estabilidad y cambio. La estabilidad es necesaria para medir tendencias en el tiempo usando "la misma vara". El cambio es necesario que se mantengan vigentes, dados los cambios en las áreas disciplinarias evaluadas, curriculum y pedagogía, y marcos de evaluación, entre otros. Para compatibilizar estabilidad y cambio, el NAEP realiza dos evaluaciones:

1. El NAEP con resultados de largo plazo[15] (NAEP Long-Term Trend Assessments). Reporta resultados históricos desde 1971 a la fecha, utilizando las mismas metodologías de evaluación y los mismos estándares de desempeño utilizados en el primer año.

2. El NAEP principal (Main NAEP)[16]. Reporta comparaciones interanuales dentro del período de vigencia de cada marco de evaluación.

Para mayor información sobre la metodología NAEP para desarrollar los estándares de desempeño, y como se vela por su alineamiento con el curriculum, ver Bourque (2009) y NAGB (2008), Apéndice B.

4. ALINEAMIENTO ENTRE EL CURRICULUM Y LAS PRUEBAS (DIMENSIÓN 2)

Como se dijo en la sección anterior, el NAEP no usa como referente un curriculum nacional, sino que usa marcos de evaluación que procuran ser representativos del curriculum de cada uno de los estados.

Los marcos de evaluación dan las directrices generales para el desarrollo de las pruebas e ítemsítems en cada área disciplinaria. Incluyen tablas de especificaciones con categorías para clasificar los ítemsítems de las pruebas en distintas dimensiones, y con la ponderación correspondiente a cada categoría en la prueba. Así, por ejemplo, en matemáticas, las tablas especifican el porcentaje de ítemsítems por contenido a evaluar (ej. propiedades de los números y operaciones, geometría, álgebra); el porcentaje de tiempo de la evaluación que debe ser dedicado a ítems de nivel de complejidad alto, medio y bajo; el porcentaje de tiempo de la evaluación para responder preguntas de selección múltiple versus preguntas abiertas; y contextos (matemáticas puras o problemas de la vida real) que presentan los ítems (NAGB, 2012). El marco de matemáticas no especifica las habilidades a medir, tales como comprender conceptos, ejecutar procedimientos, razonar, y resolver problemas. Más bien, se espera que, en cada nivel de complejidad, haya ítems que midan estas distintas habilidades.

[15] https://nces.ed.gov/nationsreportcard/ltt/
[16] https://nces.ed.gov/nationsreportcard/subjectareas.aspx

Además de los marcos de evaluación, hay especificaciones técnicas por área disciplinaria con directrices específicas para guiar el desarrollo de las pruebas e ítems. Las especificaciones operacionalizan el marco de evaluación de cada área disciplinaria, sirviendo como guía para los contratistas que desarrollan las pruebas. Así, por ejemplo, las especificaciones de ciencias se llaman "Science Assessment and Item Specifications for the 2009 NAEP[17]" (el año alude al marco de evaluación vigente). Para cada área disciplinaria a evaluar, las especificaciones dan los lineamientos para el desarrollo de pruebas e ítems. Identifican los contenidos y habilidades a evaluar, niveles de complejidad, contextos, formatos de los ítems, y criterios de corrección, entre otros. En la medida que las pruebas cumplan con estas especificaciones, se puede afirmar que están alineadas con los marcos de evaluación.

Para verificar que las pruebas respetan las especificaciones derivadas del marco de evaluación, el NAEP realiza revisiones de alineamiento a dos niveles. Estos niveles de revisión son:

1. Revisión de alineamiento entre el marco de evaluación y cada uno de los ítems de las pruebas. Para cada uno de los ítems desarrollados por el NAEP (usualmente a través de contratistas), panelistas externos (que no participaron en la elaboración de ítems) revisan que cada ítem corresponda a una de las categorías de las tablas de especificaciones. En matemáticas, por ejemplo, se revisa la correspondencia de cada ítem con los contenidos, niveles de complejidad, formatos, y contextos, según lo indica el marco de evaluación de esta prueba. Cada ítem es revisado por múltiples panelistas, lo que permite evaluar el nivel de acuerdo entre ellos. En función de esta revisión de alineamiento y de otros criterios (ej. relevancia, claridad), cada ítem es rechazado, modificado, o seleccionado directamente para la siguiente etapa (ej. pilotaje). De no haber acuerdo entre los panelistas, el ítem es modificado o eliminado del conjunto de ítems a pilotear. De ser modificado, el ítem pasa nuevamente por una revisión con panelistas externos.

2. Revisión de alineamiento entre el marco de evaluación y el conjunto de ítems de las pruebas. Una vez piloteados y seleccionados los ítems que se alinean con las tablas de especificaciones, se seleccionan los ítems para la prueba definitiva. Para ello, se utilizan tanto ítems nuevos (provenientes del piloto) como ítems del banco de ítems NAEP. El conjunto de ítems a incluir en la prueba definitiva es seleccionado velando por que respete los porcentajes indicados en las tablas de especificaciones (aceptándose variaciones leves). De este modo se vela por que las pruebas estén debidamente balanceadas, y cubran todo el dominio de evaluación. Esta revisión global es inicialmente llevada a cabo por contratistas; luego el NCES hace una verificación final.

La revisión de alineamiento se lleva a cabo en cada proceso de desarrollo de pruebas e ítems del NAEP. Esta revisión opera en base al juicio profesional, sin utilizar una métrica para cuantificar el grado de alineamiento. Todos los marcos de evaluación y las especificaciones técnicas son de acceso público.

[17] https://www.nagb.org/publications/frameworks/science/2009-science-specification.html

Los marcos de evaluación se desarrollan con la participación y consulta pública a educadores en ejercicio y oficiales de sistemas educativos. Se hacen revisiones por comités conformados por decidores de política, educadores, miembros del público en general, supervisores de las distintas asignaturas en agencias educativas (de estados, distritos escolares, etc.). También se hacen audiencias públicas requeridas por la ley que establece el NAEP, y revisiones de académicos y expertos del National Center for Education Statistics (Centro Nacional para las Estadísticas Educativas) y por un comité asesor en políticas para el NAEP.

Los marcos de evaluación del NAEP tienen que equilibrar estabilidad y cambios. Los marcos están diseñados para medir cambios en el tiempo, y por lo tanto deben ser relativamente estables. De hecho, se espera que tengan una vigencia de unos 10 años, aproximadamente. Al mismo tiempo, los marco deben responder a los cambios conceptuales en la disciplina, curriculum y estándares de desempeño, por lo que deben ser ajustados en el tiempo (Jago, 2009). Las últimas revisiones a los marcos de evaluación NAEP se realizaron en: Escritura, en 2011; matemáticas para 4o y 8o grado en 2005, y en matemáticas para 12o grado en 2013; en lectura y en ciencias, en 2009.

5. ALINEAMIENTO ENTRE LOS ESTÁNDARES DE DESEMPEÑO Y LAS PRUEBAS (DIMENSIÓN 3)

El alineamiento entre los estándares de desempeño y las pruebas NAEP se asegura por medio de dos procedimientos principales: (1) elaborando ítems que toman como insumo las descripciones preliminares de los estándares de desempeño, y (2) elaborando descripciones definitivas de los estándares de desempeño que están estrictamente referidas a los ítems incluidos en las pruebas.

Respecto del primer procedimiento (y tal como ya se dijo en la Sección 3), en el NAEP se desarrollan descripciones preliminares de los estándares de desempeño con el propósito de guiar el desarrollo de ítems. Estas descripciones preliminares son eminentemente criteriales, y están alineadas a las expectativas del marco de evaluación de cada área disciplinaria y grado evaluado. Las descripciones preliminares incluyen ejemplos de lo que los estudiantes en distintos niveles de desempeño deberían saber y poder hacer en las pruebas.

El uso de descripciones preliminares para guiar el desarrollo de ítems tiene varias ventajas. Permite asegurar que los ítems desarrollados cubran los aprendizajes terminales del marco de evaluación, permite cubrir todo el rango de niveles de dificultad (desde ítems básicos hasta los más avanzados), permite asegurar una cantidad de ítems suficiente que ancle cerca de los puntos de corte. De hecho, se vela por desarrollar una mayor cantidad de ítems con un nivel de dificultad cercano a los puntos de corte esperados. Una vez administradas las pruebas, estos ítems proveen de evidencia clave para elaborar las descripciones definitivas de los estándares de desempeño.

Respecto del segundo procedimiento, el alineamiento entre los estándares de desempeño y las pruebas NAEP se asegura elaborando descripciones definitivas que están estrictamente referidas a los ítems incluidos en las pruebas. Una vez administradas las pruebas, paneles de amplia representación participan en el establecimiento de estándares de desempeño para fijar los puntos de corte. Los puntos de corte se fijan tomando en consideración las expectativas del marco curricular (elemento criterial) y velando por que estén a una desviación estándar aproximadamente en la escala de puntajes (elemento normativo).

Los paneles elaboran las descripciones definitivas que se utilizan para reportar los resultados del NAEP. Esto se hace ajustando las descripciones preliminares según la evidencia empírica entrega por las pruebas. Aspectos no evaluados del marco NAEP no pueden ser parte de las descripciones asociadas a los estándares de desempeño. Esto implica que se eliminan contenidos y habilidades que pudieran estar en las descripciones preliminares, pero que no fueron evaluados en las pruebas. Así, las descripciones definidas están estrictamente alineadas con las pruebas.

6. Otra información relevante

No hay.

7. Referencias

Bourque, M. L. 2009. A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989-2009. Washington DC: National Assessment Governing Board. https://www.nagb.org/content/nagb/assets/documents/who-we-are/20-anniversary/bourque-achievement-levels-formatted.pdf

Jago, C. 2009. A History of NAEP Assessment Frameworks. Washington DC: National Assessment Governing Board. https://www.nagb.org/content/nagb/assets/documents/who-we-are/20-anniversary/jago-frameworks-formatted.pdf

NAGB, U.S. Department of Education. 2008. Science Framework for the 2009 National Assessment of Educational Progress. Washington DC: NAGB.

NAGB, U.S. Department of Education. 2012. Mathematics Framework for the 2013 National Assessment of Educational Progress. Washington DC: NAGB. https://www.nagb.org/content/nagb/assets/documents/publications/frameworks/mathematics/2013-mathematics-framework.pdf

8. Contacto

Teresa Neidorf Smith
AIR -- American Institute of Research

Ha sido la investigadora principal en varios estudios de validez y alineamiento del NAEP, realizados por AIR en contrato con el NCES.
tneidorf@air.org

ALINEAMIENTO ENTRE CURRICULUM, ESTÁNDARES DE DESEMPEÑO, Y PRUEBAS
CASO EN PROFUNDIDAD # 2
Ontario Provincial Assessment Program
Programa de Evaluación de la Provincia de Ontario, Canadá

1. CARACTERÍSTICAS GENERALES DEL PROGRAMA DE EVALUACIÓN

El programa de evaluación de Ontario incluye cuatro subcomponentes que se administran a todos los estudiantes de la provincia en las cuatro etapas consideradas clave en la educación provincial:

- Al final de la educación primaria inferior: Primary Division Assessments (3º Grado)
- Al final de la educación primaria superior: Junior Division Assessments (6º Grado)
- El primer año de la educación secundaria: Grade 9 Assessment of Mathematics (9º Grado)
- Un examen de certificación como requisito de graduación: Ontario Secondary School Literacy Test (10º Grado)

El propósito del programa de evaluación de Ontario es servir como auditoría independiente del nivel de dominio que alcanzan estudiantes en el sistema público y privado con respecto a las expectativas del curriculum provincial. Es parte de las políticas de *accountability* y control de calidad sistémicos. Los resultados de las pruebas de primaria se usan también para proporcionar información pedagógicamente útil a docentes, estudiantes y sus familias.

El compromiso del programa de evaluación de la provincia de Ontario es evaluar, con propósitos de *accountability,* el logro de los objetivos del curriculum de la provincia. El curriculum de Ontario describe los objetivos en conocimientos y habilidades que deben ser alcanzados por parte de los estudiantes en cada materia escolar y grado. Las evaluaciones provinciales examinan únicamente una selección de materias (lectura, escritura, matemáticas) en 3º, 6º, 9º, y 10º grado.

Las cuatro pruebas son censales y se publican los resultados provinciales, por subsistema ((*District School Boards*) y por escuela en informes especializados, además de estar disponibles en la página de internet del Ministerio de Educación de la provincia.

Cada escuela también recibe un informe de los resultados de sus estudiantes, y los estudiantes y sus familias reciben un informe individualizado. Los docentes reciben un informe acerca del desempeño de los estudiantes en las aulas evaluadas a su cargo. Aunque el Ministerio no publica *ranking*s de escuelas o jurisdicciones, otras organizaciones pueden preparar rankings a partir de la información que está disponible al público a través de la página de internet.

2. ORGANISMOS RESPONSABLES

Un organismo ministerial está a cargo del curriculum, y otro organismo semiautónomo está a cargo de la evaluación. La División de Aprendizaje y Curriculum (*Division of Learning and Curriculum*) del Ministerio de Educación de la Provincia de Ontario es responsable directo por la política curricular de la provincia y es autor de *The Ontario Curriculum* (el Curriculum de Ontario).

El programa de evaluación está a cargo de una agencia *arm´s length* (semiautónoma) de calidad. Las agencias *arm's length* en Canadá son corporaciones sin fines de lucro que ofrecen servicios contratados mediante un memorándum de entendimiento por una autoridad gubernamental. En este caso específico, la Oficina de Calidad y Accountability de Ontario (*Education Quality and Accountabiltiy Office - EQAO*) realiza todas las evaluaciones censales provinciales y organiza la participación de Ontario en evaluaciones muestrales nacionales e internacionales regido por un Memorándum de Entendimiento con el Ministerio de Educación y el gobierno de la provincia. El mandato de EQAO es servir de auditor independiente del logro de las expectativas del curriculum de Ontario en el sistema educativo de la provincia.

La EQAO es responsable del alineamiento entre curriculum, estándares de desempeño, y las pruebas. Esta responsabilidad es parte de su acuerdo contractual en el memorándum de entendimiento con el Ministerio de Educación.

3. ALINEAMIENTO ENTRE EL CURRICULUM Y LOS ESTÁNDARES DE DESEMPEÑO (DIMENSIÓN 1)

Las cuatro pruebas siguen los mismos procedimientos de alineamiento. Los esfuerzos para asegurar convergencia o alineamiento entre el curriculum y los estándares de desempeño es el resultado principalmente de tres años de trabajo llevado a cabo entre 2007 y 2009, con el EQAO y el Ministerio de Educación trabajando en colaboración: El Ministerio de Educación como encargado y experto en el curriculum y el EQAO como encargado y experto en la medición de estándares de desempeño.

El trabajo luego fue revisado en un proceso de consulta con distintos grupos de docentes de la provincia (*Teacher Moderation Process)* para asegurar la compatibilidad de los estándares de desempeño con el curriculum de 2009 a 2010. Desde entonces, no se han llevado a cabo revisiones de los estándares de desempeño. No existe un calendario acordado para revisiones periódicas, pero se anticipa que la Junta de Asesores Externos del EQAO y las auditorías externas, resulten en recomendaciones al respecto. Existe un compromiso de evaluar y determinar un proceso de revisión de los estándares cada vez que se lleve a cabo una revisión o cambio en el curriculum provincial. En esos casos se reconoce que los objetivos de mantener la comparabilidad interanual de las evaluaciones y por consiguiente la comparabilidad interanual

de la categorización de estudiantes por niveles de desempeño, no es plenamente combatible con la necesidad de alinear los estándares con revisiones al curriculum.

Todos los procedimientos están documentos, y todos los años se publican además informes técnicos correspondientes a cada administración de las pruebas, con información acerca del alineamiento, como un componente de las evidencias de validez. Todos los informes técnicos están disponibles al público en la página de internet de EQAO (http://www.eqao.com/) y el proceso de consulta a docentes (incluyendo manuales, protocolos, videos de entrenamiento, etc.) se encuentran también disponibles a través del Ministerio de Educación[18].

Se utilizan dos procedimientos para buscar alineamiento entre el curriculum y los estándares de desempeño. Primero, se establece un diseño de prueba basado en los objetivos de contenido y habilidades del curriculum y después se sigue un procedimiento de operacionalización de los niveles de logro que se establecen en el curriculum. Estas definiciones operacionales comienzan el procedimiento de establecimiento de estándares de desempeño.

El curriculum de Ontario se organiza en dos componentes principales: en un componente se establecen los objetivos en conocimientos y habilidades por asignatura y grado. Este ámbito del curriculum se llama *Specific Expectations* (expectativas específicas). El otro componente principal establece niveles de logro (*achievement levels*) organizados en un mapa de logros (*Achievement Chart*) correspondientes a las expectativas específicas.

En el proceso de diseño de las pruebas un panel de docentes de aula y expertos de medición del EQAO construyen un diseño de prueba *(test blueprint*) basado en las expectativas específicas. Este panel revisa el curriculum provincial y decide cuales de las expectativas pueden ser evaluadas en pruebas escritas estandarizadas a gran escala. Estas expectativas son luego presentadas en un documento que organiza los objetivos por bloques temáticos (*clusters*).

Los paneles preparan un documento que explica la relación entre el diseño de la prueba y el curriculum provincial. Este documento, llamado el Marco (*Framework*), forma parte del material de difusión que explica la prueba a docentes, padres y apoderados, y estudiantes. Éstos últimos pueden consultar el marco como parte de su preparación para la prueba. El marco especifica el alineamiento en el contenido de las evaluaciones con respecto a los objetivos en conocimiento y habilidades del curriculum de la provincia.

[18] Ver por ejemplo: Literacy and Numeracy Secretariat. (2016, November 3). *Teacher moderation: Collaborative assessment of student work* (Capacity Building Series: Secretariat Special Edition 2). Retrieved from (http://www.edu.gov.on.ca/eng/literacynumeracy/inspire/research/Teacher_Moderation.pdf); Literacy and Numeracy Secretariat. (2007, September 10*). Teacher moderation: Collaborative assessment of student work*. (Webcasts for Educators). [Video webcast]. Retrieved from (http://www.curriculum.org/secretariat/september10.shtml); Literacy and Numeracy Secretariat. (2007, October 15). *Developing inquiring minds: Moderation of student work* [Webcast]. Retrieved from (http://www.curriculum.org/secretariat/inquiring/moderation.shtml)

El curriculum de Ontario[19] presenta un Mapa de Logros (*Achievement Chart*) que establece cuatro niveles de logro (*levels of achievement*) y los diseños de prueba toman estos niveles como punto de partida en el proceso de desarrollo de ítems. El proceso de operacionalizar los niveles de logro que establece el curriculum en estándares de desempeño en las evaluaciones provinciales continúa con nuevos grupos de expertos que examinan evidencia sobre el desempeño real de los estudiantes de la provincia (este proceso precede al establecimiento de puntos de corte – ver sección 5). Grupos de 5 a 10 docentes de aula revisan una muestra de 20 cuadernillos de pruebas auténticos de estudiantes escogidos al azar de administraciones piloto y/o administraciones reales de pruebas basadas en los diseños de prueba, sin embargo, los cuadernillos solo tienen las contestaciones de los estudiantes, sin corregir.

Los docentes reciben un entrenamiento en procedimientos para distinguir diferencias cualitativas de logro en las contestaciones de estudiantes. Después del entrenamiento, cada panelista, en forma individual, reparte los cuadernillos en 4 grupos[20], que estima corresponden a los cuatro niveles de logro que propone el curriculum de Ontario. Estos son basados en distinciones cualitativas únicamente, los cuadernillos no están corregidos y no se cuenta con información acerca de los puntajes en este punto del proceso. Estos grupos representan su definición operacional individual de los niveles de logro. Seguidamente, se pasa por procesos de debate y consulta y finalmente se escogen los cuadernillos que con mayor frecuencia se asignaron, por parte de cada panelista, a los mismos niveles de logro.

En base a estos cuadernillos comunes, cada panelista describe en forma individual, la calidad de los desempeños que representan las contestaciones de los estudiantes en esos cuadernillos. El objetivo no es evaluar las contestaciones, sino describir el nivel de dominio demostrado por parte de los estudiantes en sus respuestas. Este proceso se repite con cada uno de los cuatro grupos de cuadernillos correspondientes a cada nivel de logro. El resultado son descripciones – definiciones operacionales – que se denominan estándares de desempeño (*performance standards*), aun cuando no se han establecido puntos de corte. El procedimiento resulta en un documento que relaciona los estándares de desempeño de la prueba con los niveles de logro propuestos en el curriculum. Estos estándares de desempeño describen tres ámbitos de desempeño de los estudiantes: grado de exactitud (*degree of accuracy*), grado de complejidad (*degree of complexity*), y profundidad de análisis (*depth of analysis*). Los puntos de corte se establecen en un siguiente paso, que detallamos en la sección 5, abajo.

[19] El proceso de desarrollo del curriculum de Ontario incluyó el uso se resultados de pruebas provinciales, nacionales e internacionales para determinar expectativas específicas y los niveles de logro.

[20] En el caso de la prueba de 10º grado, que es un requisito de graduación, solo se distribuyen los cuadernillos en dos grupos (aprobado y reprobado).

4. ALINEAMIENTO ENTRE EL CURRICULUM Y LAS PRUEBAS (DIMENSIÓN 2)

Como se puede apreciar en el apartado 3, el procedimiento de asegurar alineamiento entre el curriculum y los estándares de desempeño es a su vez el mecanismo que se sigue para asegurar el alineamiento del curriculum con las pruebas.

El Marco (*Framework*) descrito arriba, que representa la selección de expectativas específicas del curriculum que serán medidos, es la base del diseño *(Blueprint*) de la prueba. El Marco es difundido ampliamente a distintas audiencias como insumo para ayudar a los docentes, estudiantes, padres y apoderados, a comprender lo que miden las pruebas, y los ámbitos de las asignaturas a los que deben prestar atención como preparación para las mismas. Los *blueprint* se encuentran en los informes técnicos que se difunden después de concluidas las pruebas

El Marco es también el punto de partida para el diseño de ítems en la prueba – con respecto a cada ámbito del marco se construyen especificaciones de ítem que guían el trabajo de los escritores de ítems. Estos ítems son piloteados cada año, como un grupo de ítems que no cuenta para los resultados, repartidos aleatoriamente en las pruebas.

Los ítems se escriben de acuerdo al *blueprint* de la prueba, que incluye especificaciones acerca del número y tipo de ítems para cada bloque temático de la prueba. Los ítems son juzgados por paneles de docentes que evalúan su grado de alineamiento con el *blueprint*.

5. ALINEAMIENTO ENTRE LOS ESTÁNDARES DE DESEMPEÑO Y LAS PRUEBAS (DIMENSIÓN 3)

Los estándares de desempeño se definen en referencia a las pruebas. En el apartado 3, describimos un procedimiento que culmina en las descripciones de estándares de desempeño referidos a los niveles de logro correspondientes al curriculum, cuya definición deriva del análisis de muestras de trabajo de estudiantes en cuadernillos de pruebas.

Seguidamente, un nuevo panel independiente, de aproximadamente 20 docentes y directores de escuelas participan, con asesoría de psicometristas del EQAO, en el establecimiento de los puntos de corte que separan los niveles en los estándares de desempeño. En este proceso, los panelistas se enfocan especialmente en las descripciones de niveles de desempeño adyacentes (por ejemplo, Nivel 1 y Nivel 2), procurando entender las diferencias cualitativas entre un nivel y otro.

En las evaluaciones de 3º, 6º y 9º grado, las pruebas incluyen ítems de selección múltiple y de respuesta abierta. Los niveles se asignan usando el θ (theta) calculado en la calibración (se usan 3PL -*Three Parameter Logistic* - para ítems de escogencia y GPC - *Generalized Partial Credit*- para ítems de respuesta correcta). Cuatro θ establecen los puntos de corte entre los cinco niveles, el ancho de cada nivel de desempeño se calcula tomando el rango completo de thetas y dividiendo por cinco. θ es el parámetro de "proficiencia" que estima la habilidad latente de los estudiantes para responder a cada ítem.

No se siguen estrictamente criterios algorítmicos, la división de la distribución de puntajes en 5 niveles no se hace para establecer un rango semejante de nivel en nivel. Lo que se hace es que el panel de docentes y directores, que no son especialistas en psicometría, leen y resuelven una muestra de cuadernillos de prueba. Luego leen y discuten los estándares de desempeño que estableció el panel de jueces anterior, y luego comienzan a asignar cuadernillos a los niveles de esos estándares. Se discute las formas de asignar cuadernillos a niveles, se procura llegar a relativa uniformidad de criterios, y luego en forma individual, trabajando con los expertos, cada panelista fija sus propios puntos de corte para distinguir un nivel de otro. Típicamente, estos niveles de corte son diferentes unos de otros, y los psicometristas asesoran al grupo mostrándoles datos acerca del impacto de sus puntos de corte (datos acerca del porcentaje de estudiantes que estarían categorizados en cada nivel, según los niveles de corte y muestra de los cuadernillos correspondientes a los distintos niveles en nuevas muestras aleatorias de cuadernillos). El panel comienza a trabajar en forma conjunta, consensuando criterios y viendo nuevos datos acerca del impacto de nuevos puntos de corte consensuados. Generalmente toma 2 o 3 iteraciones de trabajo común llegar a puntos de corte con altos niveles de consenso entre panelistas.

Combinando el análisis de ítems comunes de anclaje entre las pruebas, y los parámetros de los ítems nuevos a partir de su pilotaje en años anteriores, se clasifican los ítems de acuerdo a sus probabilidades (P) en las distintas bandas correspondientes a los estándares de desempeño. Se ha acordado que el Nivel 3 indica que se han alcanzado las expectativas curriculares correspondientes a los grados evaluados.

El establecimiento de los puntos de corte para cada banda es un proceso iterativo en donde los panelistas asignan puntos de corte y los especialistas en psicometría les presentan datos acerca del impacto de sus decisiones – principalmente en la forma de porcentajes de estudiantes que serian categorizados en cada nivel de desempeño según sus decisiones. El resultado final de este proceso es el establecimiento de los puntos de corte.

La equiparación (equating) de pruebas tiene como objetivo principal asegurar la comparabilidad interanual de los puntos de corte y la estabilidad de las categorizaciones de los estudiantes por nivel de desempeño. En los informes técnicos de cada año, se reportan los procedimientos y resultados de los procesos de escalamiento y calibración como la evidencia de la comparabilidad interanual de los estándares de desempeño.

6. OTRA INFORMACIÓN RELEVANTE

El EQAO considera que los instrumentos principales para garantizar el alineamiento de las pruebas con las expectativas específicas del curriculum, y el alineamiento de los estándares de desempeño de las pruebas con los niveles de logro del curriculum, son el Marco (Framework) y los Diseños (Blueprint) que hemos descrito arriba. Sin embargo, todos los años, en los informes técnicos de las pruebas, se presentan evidencias de alineamiento principalmente asociados a los procedimientos para garantizar adherencia a los Marcos y Diseños, el uso de docentes con experticia en el Curriculum de Ontario y con experiencia reciente en su implementación en el aula, y sus procedimientos de entrenamiento y control de calidad.

7. REFERENCIAS

Education Quality and Accountability Office. 2009. Framework: Grade 9 Assessment of Mathematics. Toronto: The Queens Printer for Ontario.

Education Quality and Accountability Office. 2015. EQAO's Technical Report for the 2013-14 Assessments. Toronto: EQAO

Education Quality and Accountability Office. 2012. Understanding Levels of Performance. Toronto. The Queen's Printer for Ontario.

Ministry of Education of Ontario. 2007. The Ontario Curriculum Grades 11 and 12 Mathematics (Revised). Toronto. The Queen's Printer for Ontario.

8. CONTACTO

Richard G. Wolfe, Professor Emeritus, Ontario Institute for Studies in Education – University of Toronto: wolferg@gmail.com

ALINEAMIENTO ENTRE CURRICULUM, ESTÁNDARES DE DESEMPEÑO, Y PRUEBAS
CASO EN PROFUNDIDAD # 3
Plan Nacional de Evaluación de los Aprendizajes (PLANEA)
Instituto Nacional para la Evaluación de la Educación, México

1. CARACTERÍSTICAS GENERALES DEL PROGRAMA DE EVALUACIÓN

Los exámenes de Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) son pruebas elaboradas por el Instituto Nacional para la Evaluación de la Educación (INEE), para evaluar el nivel de dominio que los estudiantes mexicanos alcanzan de los planes y programas de estudio del curriculum nacional. También tienen la misión de identificar los factores asociados a las diferencias entre los *niveles de logro* o *niveles de competencia*[21] (estándares de desempeño). Existen dos regímenes de pruebas PLANEA:

1. PLANEA–ELSEN (Evaluaciones de Logro del Sistema Educativo) elaborados y administrados por el INEE a una muestra representativa del sistema educativo nacional, y de los distintos estados. Se lleva a cabo cada 4 años.
2. PLANEA–ELCE (Evaluación de Logro de los Centros Escolares) elaborados por INEE, y administrados en forma conjunta por el INEE y la Secretaria de Educación Publica (SEP) en forma censal a todos los centros escolares de México. La prueba ELCE es una prueba previamente administrada como prueba ELSEN, liberada para su uso en ELCE. Las pruebas ELCE son coordinados por cada centro educativo siguiendo normas de la SEP y del INEE.

Los PLANEA no tienen consecuencias para escuelas, estudiantes o docentes; tampoco se consideran en la evaluación de docentes, que tiene sus propios instrumentos. La prueba muestral (ELSEN) se considera la prueba base, de mayor calidad de información, y es para informar acerca de los resultados de la escolarización a nivel nacional y estatal. El propósito de la prueba censal (ELCE) es que maestros, directores y supervisores cuenten con una herramienta de evaluación que les permita obtener información acerca del logro alcanzado por los alumnos de cada centro escolar, permitiendo que el colectivo reflexione durante las sesiones del Consejo Técnico Escolar (CTE) acerca de estos resultados. De esta manera, se pretende facilitar la identificación de áreas, temas o contenidos que requieren mayor atención y, con ello, facilitar la intervención pedagógica establecida en la Ruta de Mejora de las escuelas.

Los estándares de desempeño se utilizan principalmente en el marco del cumplimiento del objetivo, establecido en la Ley del Instituto Nacional para la Evaluación de la Educación de "difundir información que contribuya a evaluar los componentes, procesos y resultados del Sistema Educativo Nacional[22]". El INEE y su sitio de internet difunden todo el material

[21] Los materiales del INEE usan tanto "niveles de logro" como "niveles de competencia" al referirse a los estándares de desempeño.
[22] Artículo 12, Inciso IV

necesario para entender el uso de los Estándares de Desempeño en los informes del INEE, en el Banco de Indicadores Educativos (un sistema nacional de indicadores educativos disponible desde el ciclo 2004-2005) y en el informe nacional anual, denominado "Panorama Educativo de México".

Las áreas disciplinares y grados evaluados por PLANEA se presentan en la Tabla 1

Tabla 1 México: Niveles, grados y áreas disciplinares de los PLANEA.

Nivel	Grado	Áreas Disciplinares
Preescolar	3º[23]	Razonamiento Numérico
		Razonamiento Verbal
Primaria	3º	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales
	6º	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales
Secundaria	3º[24]	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales

-

2. ORGANISMOS RESPONSABLES

PLANEA es responsabilidad del Instituto Nacional para la Evaluación de la Educación (INEE). El INEE funciona desde el año 2002, primero por decreto presidencial, y después de 2013 ha adquirido su actual carácter como organismo público autónomo, con personalidad jurídica y patrimonio propio. El INEE depende directamente del Senado de la República de México y es gobernado por una Junta de Gobierno que consiste en cuatro consejeros y una consejera presidente. Su misión es evaluar la calidad, el desempeño y los resultados del Sistema Educativo Nacional de México en la educación preescolar, primaria, secundaria y media superior.

[23] El 3º grado de preescolar es el grado terminal de ese nivel. Los alumnos ingresan al 1º grado a los 3 años de edad. La edad normativa de 3º de preescolar es 5 años.

[24] Es el grado terminal de secundaria, y la edad normativa es 14 a 15 años. Existen cuatro modalidades: secundaria general, telesecundaria, secundaria técnica industrial y secundaria federal.

Los exámenes de PLANEA son los sucesores de las pruebas llamadas Excale (Los primeros Excale se realizaron en el año lectivo 2004-2005) y el cambio de nombre se da como resultado de la reforma educativa del 2013. Las evaluaciones son las mismas de Excale, manteniendo la comparabilidad interanual y demás características. La diferencia es que el INEE es responsable único de la evaluación muestral PLANEA ELSEN (Evaluaciones de Logro del Sistema Educativo Nacional) y es responsable por el desarrollo de una versión censal de la misma prueba en donde las normas de aplicación son establecidas por el INEE y la Secretaría de Educación Pública, pero que esta a cargo de los directivos y los docentes de los propios centros educativos evaluados.[25]. Los procedimientos de desarrollo de estándares establecidos originalmente en Excale se siguen de igual forma en cada prueba PLANEA, y los procedimientos de equiparación (equating) y calibración se siguen para asegurar comparabilidad interanual.

3. ALINEAMIENTO ENTRE EL CURRICULUM Y LOS ESTÁNDARES DE DESEMPEÑO (DIMENSIÓN 1)

En México no se considera un solo documento como la especificación única del curriculum nacional. Por tal motivo los documentos de referencias de PLANEA son: los planes y programas de estudio, los libros de texto oficiales del estudiante y del docente, fichas de trabajo y distintos materiales instruccionales oficiales. La mayoría de estos documentos provienen de la Subsecretaria de Desarrollo Curricular de la Subsecretaria de Educación Básica de la Secretaría de Educación Pública, o de las distintas Direcciones Generales correspondientes a los distintos planes de estudio en la Subsecretaria de Educación Media.

Los estándares de desempeño no forman parte del curriculum nacional de México. Como hemos explicado, el curriculum nacional se considera como un conjunto amplio de documentos oficiales, que especifican distintos ámbitos del curriculum formal y que deben ser analizados (en un procedimiento que describimos abajo en la Sección 4) para el diseño de las pruebas.

Los estándares de desempeño, denominados en México niveles de logro, son herramientas creadas por el INEE como categorías para el análisis y la comunicación de resultados de sus pruebas.

Fueron establecidos en un seminario académico, convocado por el INEE con participación de expertos nacionales e internacionales en donde se consideraron distintos criterios para el establecimiento de estándares de desempeño, se propusieron y consideraron distintos criterios de calidad para los estándares, y se emitieron recomendaciones. La Dirección de Pruebas y Medición del INEE, apoyada por las recomendaciones del seminario académico, estableció una categorización y definición base de cuatro niveles de logro que se presentan en la Tabla 2.

[25] Previo al 2013, la Secretaría de Educación Pública tenía su propia prueba censal, sobre la cual no se difunde información técnica alguna, y acerca de la cual había muchas dudas y criticas técnicas, llamada ENLACE.

Tabla 2 México: Categorías y definición de base de los niveles de logro PLANEA.

Nivel	Definición base
Por debajo del nivel básico	Indica carencias importantes en el dominio curricular de los conocimientos, habilidades y destrezas escolares que expresan una limitación para poder seguir progresando satisfactoriamente en la materia.
Básico	Indica el dominio imprescindible suficiente, mínimo, esencial, fundamental, o elemental de conocimientos, habilidades y destrezas escolares necesarias para poder seguir progresando satisfactoriamente en la materia
Medio	Indica un dominio sustancial (adecuado, apropiado, correcto o considerable) de conocimientos, habilidades y destrezas escolares, que pone de manifiesto un buen aprovechamiento de lo previsto en el curriculum.
Avanzado	Indica un dominio muy elevado (intenso, inmejorable, óptimo o superior) de conocimientos, habilidades y destrezas escolares que refleja el aprovechamiento máximo de lo previsto en el curriculum.

El alineamiento de estos estándares de desempeño con el curriculum se da a través del juicio de los panelistas como parte de la metodología de diseño de las pruebas PLANEA. El diseño de la prueba especifica una relación analítica documentada entre el (a) el curriculum y los ítems de la prueba (que describimos en la sección 4), y (b) el uso de los ítems de las pruebas para establecer los puntos de corte entre los niveles de logro para operacionalizar estos últimos (en un proceso que describimos en la sección 5). Los estándares de desempeño se operacioinalizan en las evaluaciones, y no están presentes en el curriculum. En sentido estricto lo que se alinea es la interpretación del juicio experto con respecto a los estándares de desempeño implícitos en el curriculum nacional.

4. ALINEAMIENTO ENTRE EL CURRICULUM Y LAS PRUEBAS (DIMENSIÓN 2)

Como hemos observado en la sección anterior, en México el curriculum nacional no es un solo documento a cargo de un solo organismo. Más bien el curriculum intencional de México se considera formado por planes y programas de estudio, libros de texto, libros del profesor, ficheros, el Curso Nacional de Actualización del Magisterio, y otros documentos de política curricular oficial emitidos por distintas direcciones de las Secretarias y Subsecretarias de Educación, correspondientes a los niveles educativos, áreas disciplinares, y planes de estudio[26].

[26] En secundaria, por ejemplo, existen cuatro planes de estudio con distintos programas específicos en cada área disciplinar: secundaria general, telesecundaria, secundaria técnica industrial y secundaria federal.

Debido a esta naturaleza del curriculum intencional o formal en México, la primera labor que realiza el INEE para diseñar pruebas alineadas al curriculum, es conformar comités académicos cuyo propósito es analizar todos los documentos relevantes para determinar el universo completo de metas curriculares por área disciplinar y grado y luego, a partir de este análisis, seleccionar las áreas curriculares que se medirán en PLANEA.

Cada Comité Académico (CA) es conformado por aproximadamente una docena de participantes, y sus miembros son especialistas curriculares en las áreas disciplinares, académicos especialistas, autores de libros de texto, directivos de escuelas, profesores de instituciones de formación docente, docentes en servicio y asesores metodológicos del grupo técnico del INEE. Cada CA recibe capacitación en análisis curricular, diseño de pruebas y otros elementos considerados esenciales para llevar a cabo sus labores[27].

Luego el CA analiza un amplio espectro de documentos del curriculum formal de México, debido a que, en palabras del Manual Técnico:

> *"Como en la práctica ningún documento contiene todo lo que se debe enseñar o lo que es importante curricularmente, en esta primera etapa es necesario efectuar un análisis de contenido de diversas fuentes que definen el curriculum de la asignatura, tales como el plan y los programas de estudios, los libros de texto y del maestro, las fichas de trabajo, así como diversos materiales instruccionales. Esto con el propósito de hacer explícito el dominio de resultados de logro pretendidos por el curriculum (en la asignatura correspondiente) y determinar su alcance." (Manual Técnico, pg. 17)*

El análisis curricular efectuado por parte de cada CA resume los resultados importantes pretendidos por el curriculum del área disciplinar en una tabla de doble entrada llamada *Retícula*. Esta retícula además de resumir un análisis comprensivo del curriculum formal, toma algunas decisiones preliminares acerca de prioridades para la evaluación. Con el propósito de validar y complementar el análisis curricular llevado a cabo por los CA, se llevan a cabo consultas acerca de prioridades curriculares a muestras por conveniencia de docentes en servicio, y se comisionan estudios independientes por parte de especialistas.

La retícula es sujeta de análisis secundario, con el propósito de identificar prioridades de contenidos y expectativas de rendimiento en el curriculum formal, con el fin de determinar los elementos a ser considerados en los diseños de las pruebas. La caracterización del curriculum que se registra en la retícula es compleja, incluye no solo los elementos curriculares antes mencionados, sino también relaciones entre contenidos, entre contenidos y habilidades, aspectos epistemológicos y muchos otros. Hay un importante protocolo de análisis para determinar relevancia, cadenas de contenido, y otros elementos a partir de la evidencia empírica relevada en el análisis curricular y que se explica detalladamente en los materiales técnicos. Las

[27] Los materiales de entrenamiento, protocolos de trabajo, e instrumentos utilizados en la labor de los CA se encuentran en la página de internet del INEE. http://www.inee.edu.mx/index.php/proyectos/excale/excale-documentos-tecnicos

decisiones acerca de las expectativas curriculares a medir y el peso que tendrá cada expectativa en la prueba, se registra en Tablas de Contenidos, que son los documentos básicos del diseño de las pruebas. A partir de estas Tablas de Contenido se desarrollan los diseños de los ítems de las pruebas. El alineamiento se asegura verificando que cada ítem de la prueba tenga su correspondencia en la Tabla de Contenidos. También se asegura verificando que el conjunto de ítems de la prueba corresponda al peso esperado.

Las evidencias de la validez curricular de PLANEA – y por consiguiente de su alineamiento – es la documentación y los análisis que respaldan el análisis curricular que resulta en las retículas y las Tablas de Contenido. Hasta el momento, no hay estudios independientes ni evaluaciones post hoc por parte del INEE del alineamiento.

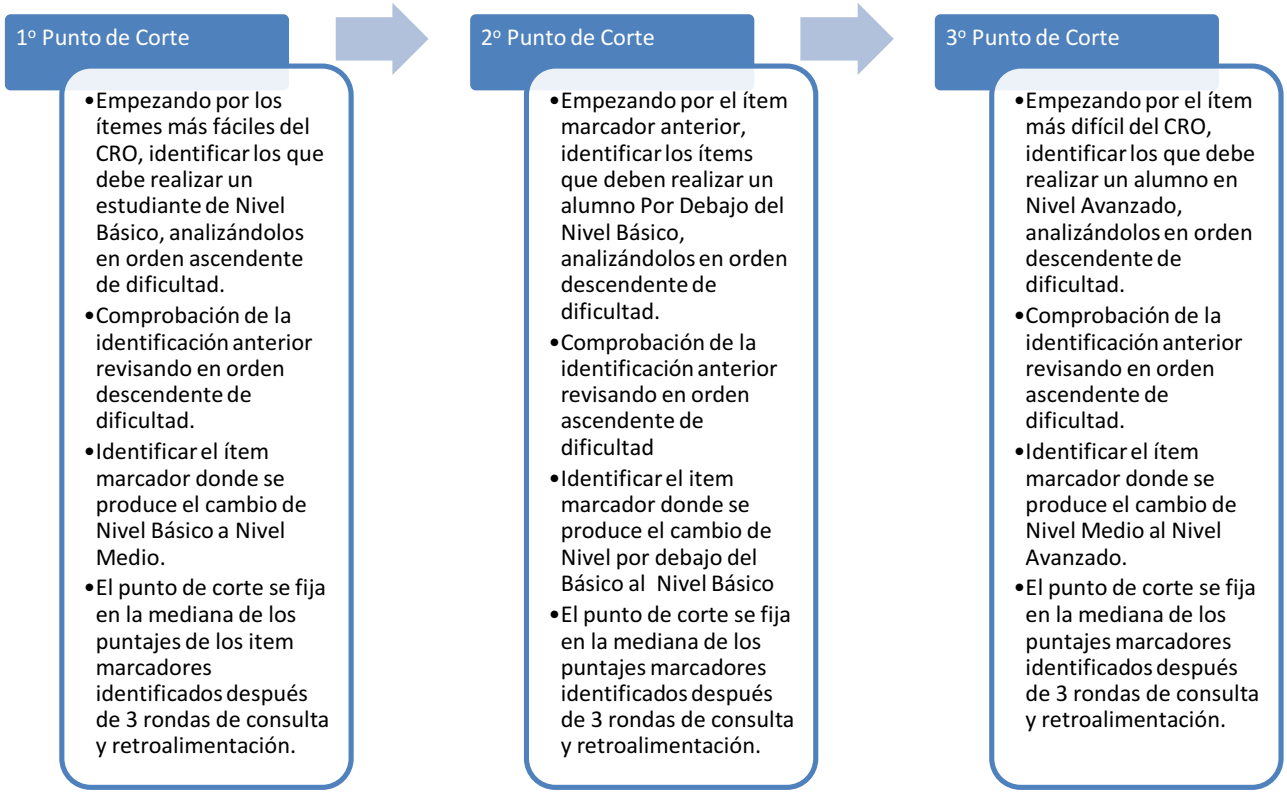
5. ALINEAMIENTO ENTRE LOS ESTÁNDARES DE DESEMPEÑO Y LAS PRUEBAS (DIMENSIÓN 3)

Los puntos de corte que operacionalizan los estándares de desempeño que se describen en la Sección 3 arriba, se definen en referencia a las pruebas, y no se usan como parte del curriculum intencional de México. Es decir, los puntos de corte solo tienen funcionalidad en la evaluación del sistema educativo, no se prescribe su uso en la instrucción en el aula. Come se definen en referencia estricta a la prueba, el alineamiento de estándares y pruebas es consecuencia de la metodología.

Para establecer los puntos de corte, se sigue una metodología Bookmark. Los puntos de corte se establecen según procedimientos de uso de juicio experto, con el propósito de identificar los tres puntos de inflexión que marcan la diferencia entre los cuatro niveles de logro (definición de base a partir de los referentes curriculares). Los puntajes correspondientes a los puntos de inflexión se establecen después de aproximadamente tres rondas de consulta en un comité de jueces expertos y se fijan en la mediana del puntaje TRI (Teoría de Respuesta al Ítem) correspondiente al nivel de habilidad correspondiente a los ítems identificados como "marcadores", es decir, en los puntos de inflexión de los niveles de logro.

En el establecimiento de los puntajes de corte el método es el siguiente: el Comité de Puntuaciones de Corte (CPC) recibe entrenamiento en la finalidad y los objetivos de la determinación de niveles de logro. Cada miembro del CPC trabaja sobre un cuadernillo de reactivos ordenados (CRO) por nivel de dificultad. En las sesiones de juicio se trabaja primero en forma individual, y luego en forma conjunta, para lograr consenso o congruencia entre jueces con respecto a los ítems que marcan los puntos de inflexión entre niveles. Estos ítems –llamados ítems marcadores– se identifican comenzando por los más fáciles y por el punto de inflexión entre los primeros dos niveles de logro. En la Ilustración 1 se presenta un esquema general de las sesiones de juicio para establecer los puntos de corte.

Ilustración 1 México: Esquema del proceso de establecimiento de puntos de corte PLANEA



6. Otra información relevante

Toda la documentación técnica de los procedimientos, incluyendo los manuales, protocolos, marcos de referencia y documentación de talleres de entrenamiento, se difunden a través de la página de internet del INEE (http://www.inee.edu.mx/index.php/proyectos/excale/excale-documentos-tecnicos).

7. Referencias

Instituto Nacional para la Evaluación de la Educación. 2009. *Manual Técnico*: *Diseño de Exámenes de la Calidad y el Logro Educativos: Excale.* México, D.F.: INEE

Instituto Nacional para la Evaluación de la Educación. 2006. *Manual Técnico: Establecimiento de niveles de competencia.* México, D.F.: INEE

Instituto Nacional para la Evaluación de la Educación. 2005. *Plan General de Evaluación del Aprendizaje* México, D.F.: INEE

8. CONTACTO

Felipe Martínez Rizo, ex–Director General del INEE (Investigador Honorifico del INEE): felipemartinez.rizo@gmail.com

Margarita Zorrilla Fiero, Consejera, Junta de Gobierno del INEE: margarita.zorrilla@gmail.com

Para más información:

Instituto Nacional para la Evaluación de la Educación
José María Velasco 101, Piso 5
Col. San José Insurgentes, C.P. 03900, México, D.F.
www.inee.edu.mx

5. ANÁLISIS DE LOS RESULTADOS ENCONTRADOS

La Tabla 5 presenta una síntesis de los resultados de los tres casos en profundidad seleccionados por el MINEDUC: (1) NAEP [National Assessment of Education Progress] de EEUU, (2) OPAP [Ontario Provincial Assessment Program] de Canadá, y (3) PLANEA [Plan Nacional para la Evaluación de los Aprendizajes] de México. La tabla presenta, además, los resultados para el SIMCE, obtenidos a través del análisis de la información del diagnóstico. Para cada caso, se analizaron las tres dimensiones del triángulo de alineamiento: (1) alineamiento entre el curriculum y los estándares de desempeño, (2) alineamiento entre el curriculum y las pruebas estandarizadas, y (3) alineamiento entre los estándares de desempeño y las pruebas estandarizadas.

Las dimensiones y criterios utilizados en esta tabla surgen de las orientaciones dadas por el equipo del MINEDUC, de la literatura especializada sobre alineamiento, y de la práctica profesional en programas de evaluación de distintos países. El modelo incorpora fuentes de evidencia de alineamiento propuestas por Davis-Becker (2013), criterios del método Webb y del modelo Achieve, así como criterios y procedimientos propios de la práctica profesional de programas de evaluación de distintos países.

Tabla 5. Resumen revisión de alineamiento entre curriculum, estándares de desempeño, y las pruebas SIMCE (Chile), NAEP (EEUU), OPAP (Ontario, Canadá), y PLANEA (México).

CRITERIO		Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
VALIDEZ DE USOS: *Se refiere a la correcta interpretación de los resultados de las evaluaciones como reflejo de los aprendizajes alcanzados en el currriculum nacional.*					
Dimensión 1: Alineamiento entre el curriculum y los estándares de desempeño					
1.1	Los estándares de desempeño se elaboran tomando como primera referencia el curriculum oficial.	SI	SI	NO	NO
1.2	Los estándares de desempeño describen los aprendizajes terminales o "punta de iceberg" del curriculum.	SI	SI	SI	SI
1.3	Todos los contenidos y habilidades descritos en los estándares de desempeño están especificados en el curriculum oficial.	SI	SI	SI	NO

Tabla 5 (cont.)

Dimensión 1: Alineamiento entre el curriculum y los estándares de desempeño
COMENTARIOS **1.1.** SIMCE: Un primer paso para la elaboración de los estándares de desempeño consiste en describir los requisitos mínimos teóricos para alcanzar el estándar ("expectativa teórica"). NAEP: En EEUU no hay un curriculum nacional. Los estándares de desempeño se elaboran tomando como primera referencia el marco de evaluación de la prueba. Se hacen descripciones preliminares de los estándares de desempeño (Achievement Levels) a partir del marco de evaluación (previo a la administración de las pruebas). OPAP: El curriculum incluye niveles de logro, que son la base para elaborar los estándares de desempeño. PLANEA: Los estandares de desempeño fueron establecidos por un panel de expertos sin referirse al curriculum nacional, luego se usa el curriculum nacional en el diseño de las pruebas y los items que se usan para operacionalizar los estándares.Se hacen "Definiciones de base de los Niveles de Logro", a partir del marco curricular (previo a la administración de las pruebas). **1.2.** SIMCE: Los requisitos mínimos "teóricos" de los estándares de desempeño incluyen aquellos conocimientos y habilidades considerados “aprendizajes terminales”. NAEP: Los estándares de desempeño incluyen estos aprendizajes terminales. OPAP considera los objetivos terminales del ciclo en el Curriculum de Ontario. PLANEA infiere aprendizajes terminales a partir del análisis del curriculum intencional. SIMCE: Los requisitos mínimos "teóricos" de los estándares de desempeño incluyen aquellos conocimientos y habilidades considerados “aprendizajes terminales” NAEP: Los estándares de desempeño incluyen estos aprendizajes terminales. OPAP considera los objetivos terminales del ciclo en el Curriculum de Ontario. PLANEA infiere aprendizajes terminales a partir del análisis del curriculum intencional. **1.3.** PLANEA: No existe un referente curricular único, y por tanto los estándares de desempeño se derivan de un estudio de los distintos instrumentos del curriculum formal, pero esos instrumentos no contienen los estándares de desempeño. NAEP: Todos los contenidos y habilidades descritos en los estándares de desempeño están especificados en el marco de evaluación (dado que no hay un curriculum nacional en EEUU).

Tabla 5 (cont.)

CRITERIO		Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
VALIDEZ DE USOS: *Se refiere a la correcta interpretación de los resultados de las evaluaciones como reflejo de los aprendizajes alcanzados en el currriculum nacional.*					
Dimensión 2: Alineamiento entre el curriculum y la prueba					
Alineamiento entre el curriculum nacional y el marco de evaluación de la prueba					
2.1	Hay un marco de evaluación que describe los propósitos de la evaluación, interpretación y usos esperados de los resultados, enfoque conceptual del área disciplinaria a evaluar, objetivos curriculares, contenidos, habilidades, contextos/situaciones a evaluar, formatos de la prueba y de las preguntas, especificaciones de la prueba, tiempos de evaluación, entre otros.	PARCIAL	SI+	SI+	SI
2.2	Hay un marco de evaluación que cubre los objetivos curriculares factibles de ser evaluados en una prueba de lápiz y papel.	SI	SI	SI	SI
2.3	Hay un marco de evaluación que describe los contenidos del curriculum a evaluar.	SI	SI+	SI+	SI
2.4	Hay un marco de evaluación que describe las habilidades del curriculum a evaluar.	SI	SI+	SI+	SI
2.5	Hay un marco de evaluación que describe las situaciones/contextos en el que los estudiantes deben demostar lo aprendido, según lo indicado en el curriculum.	NO	SI	SI	SI
2.6	Hay un marco de evaluación que describe el nivel de complejidad cognitiva o nivel de dificultad de los ítems de la prueba.	NO	SI	SI+	SI
2.7	Hay un marco de evaluación que cubre los "aprendizajes terminales" del área disciplinaria y grado evaluado, según lo indicado en el curriculum.	SI	SI	SI	SI
2.8	Hay un documento técnico que operacionaliza el marco de evaluación y que sirve para guiar el desarrollo de las pruebas e ítems.	SI	SI	SI+	SI

Tabla 5 (cont.)

Dimensión 2: Alineamiento entre el curriculum y la prueba
Alineamiento entre el curriculum nacional y el marco de evaluación de la prueba
COMENTARIOS **2.1-2.5.** SIMCE: Informes técnicos incluyen características generales de las pruebas, tablas o matrices de especificaciones con contenidos, habilidades a evaluar. Sin embargo, no hay lineamientos para la elaboración de los estándares de desempeño. NAEP: Los "NAEP frameworks" ofrecen información detallada, que cubren desde el enfoque conceptual hasta las tablas de especificaciones con contenidos, habilidades, contextos, y nivel de complejidad. OPAP: El Marco (Framework) representa la selección de expectativas del currículum que serán evaluados, y es la base del diseño (Blueprint) de la prueba. Especifica contenidos, habilidades, nivel de logro o expectativa de desempeño. PLANEA: Las "Retículas" especifican contenidos, habilidades, y expectativas de desempeño. **2.6.** SIMCE: El marco de evaluación no incluye lineamientos respecto del niveles de complejidad cognitiva, o de dificultad de los ítems. NAEP: El marco de evaluación incluye tablas de especificaciones por nivel de complejidad cognitiva. OPAP: El curriculum incluye niveles de logro, que especifican las expectativas de desempeño. Test blueprints se basan en estos niveles de logro. PLANEA: Las expectativas de rendimiento son parte de la retícula, e informan la elaboración de ítems. **2.7.** NAEP: Aprendizajes terminales son mencionados en las descripciones preliminares de niveles de desempeño, para guiar así la elaboración de ítems. OPAP usa los objetivos terminales mencionados en el Curriculum de Ontario. PLANEA: Induce los aprendizajes terminales como resultado del analisis del conjunto de documentos que constituyen el curriculum intencional de Mexico. **2.8.** SIMCE: Las tablas de especificaciones sirven este propósito. NAEP: Hay documentos de acceso público especialmente elaborados para este propósito ("Assessment and item specifications") OPAP: Test blueprints están basados en los niveles de logro del curriculum. PLANEA: Las "Tablas de Contenidos" son los documentos básicos del diseño de las pruebas.

Tabla 5 (cont.)

CRITERIO		Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
Dimensión 2: Alineamiento entre el curriculum y la prueba					
Alineamiento entre la tabla de especificaciones de la prueba y las preguntas/problemas/ítems de la prueba					
2.9	Se revisa que cada uno de los ítems de la prueba puede ser clasificado de acuerdo a los contenidos y habilidades identificadas en las especificaciones de las pruebas.	SI	SI	SI	SI
2.10	Se revisa que el conjunto de ítems de la prueba refleje las ponderaciones asignadas a cada uno de los ejes de contenidos de tabla de especificaciones.	SI	SI	SI	SI
2.11	Se revisa que hay una distribución balanceada de los ítems de la prueba en los distintos subejes de la tabla de especificaciones.	SI	SI	SI	SI
2.12	Se revisa por que el conjunto de ítems de la prueba refleje las ponderaciones asignadas a cada uno de los ejes de habilidades de tabla de especificaciones.	SI	SI	SI	SI
2.13	Se revisa que cada uno de los ítems de la prueba pueda ser clasificado en una de las situaciones/contextos descritos en el marco de evaluación.	No aplica	SI	SI	SI
2.14	Se revisa que haya un número suficiente de ítems por nivel de complejidad cognitiva o dificultad para el grado evaluado, según lo indicado en el marco de evaluación.	No aplica	SI	SI	SI

COMENTARIOS

2.9-2.12. TODOS: El alineamiento se asegura verificando que cada ítem de la prueba tenga su correspondencia en las especificaciones de la prueba (o equivalente), y verificando que el conjunto de ítems de la prueba corresponda al peso esperado.

2.13. SIMCE: Los contextos no están especificados.

NAEP, OPAP y PLANEA: Los contextos en los que los estudiantes deben demostrar lo que saben y pueden hacer forman parte de los marcos de evaluación y especificaciones de las pruebas.

2.14. SIMCE: El marco de evaluación no incluye lineamientos respecto del niveles de complejidad cognitiva, o de dificultad de los ítems.

NAEP: El marco de evaluación incluye tablas de especificaciones por nivel de complejidad cognitiva.

OPAP: El curriculum incluye niveles de logro, que especifican las expectativas de desempeño. El marco de evaluación se basa en estos nivele de logro.

PLANEA: Las expectativas de rendimiento son parte de la retícula, e informan la elaboración de ítems.

Tabla 5 (cont.)

	CRITERIO	Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
Dimensión 3: Alineamiento entre estándares de desempeño y la prueba					
3.1	Se elaboran descripciones preliminares de los estándares de desempeño, basadas en el curriculum solamente, con el fin de guiar el desarrollo de los ítems de las pruebas.	NO	SI	SI	SI
3.2	Los estándares de desempeño que se utilizan para reportar los resultados de las evaluaciones están estrictamente referidos a los contenidos y habilidades evaluados en las pruebas.	NO	SI	SI	SI
3.3	Se asegura que haya un número suficiente de ítems que anclan en torno a los puntos de corte de los niveles de desempeño.	PARCIAL	SI	SI	SI
3.4	Se asegura que haya un número suficiente de ítems que evalúen los contenidos, habilidades y contextos descritos para cada nivel de desempeño.	PARCIAL	SI	SI	SI
3.5	Se asegura que haya un número similar de ítems que anclan en cada nivel de desempeño.	PARCIAL	SI	SI	SI

COMENTARIOS

3.1. SIMCE: Se elaboran "requisitos mínimos teóricos" a partir del curriculum.
NAEP: Se hacen descripciones preliminares de los niveles de desempeño ("Achievement Levels") a partir del marco de evaluación (previo a la administración de las pruebas), con el propósito de guiar el desarrollo de ítems.
OPAP: Estas descripciones vienen dadas en los niveles de logro, los que no forman parte del curriculum. Planea formula las descripciones preliminares en forma independiente, en un seminario de especialistas.
3.2. NAEP, OPAP, PLANEA: Estándares de desempeño utilizados para reportar resultados están estrictamente basados en la evidencia aportada por las pruebas.
SIMCE: Estándares de desempeño también incluyen aspectos del curriculum no evaluados en las pruebas (ej. uso de la regla en matemáticas).
3.3-3.5. SIMCE: Los estándares de desempeño no informan la elaboración de ítems.
NAEP, OPAP, Planea: Se asegura "por diseño", solicitando la elaboración de ítems alineados con las descripciones preliminares (NAEP) o niveles de logro curriculares (OPAP), o niveles de logro (PLANEA).

Tabla 5 (cont.)

	CRITERIO	Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
VALIDEZ DE PROCESOS: Se *refiere a la calidad técnica del diseño estudio de alineamiento, al uso apropiado de métodos y procedimientos, y a las calificaciones de los panelistas.*					
Panelistas					
4.1	Los panelistas tienen conocimientos disciplinarios y pedagógicos adecuados para realizar la revisión de alineamiento.	SI	SI	SI	SI
4.2	Los panelistas tienen conocimiento de la población de estudiantes evaluadas (ej. nivel de aprendizajes, oportunidades de aprendizaje en la escuela).	SI	SI	SI	SI
4.3	Los paneles incluyen panelistas independientes del equipo que desarrolló el curriculum, las pruebas, y los estándares de desempeño (panelistas externos).	SI	SI	SI	SI
4.4	Los panelistas representan a distintos usuarios de la información de las evaluaciones (ej. docentes de aula, directivos, padres, funcionarios Ministerio de Educación).	PARCIAL	PARCIAL	SI	PARCIAL
4.5	Los panelistas representan la diversidad cultural del país/estado (ej. representantes de distintas regiones y grupos étnicos, de distintos niveles socioeconómicos, hombres y mujeres).	NO	SI	SI	NO
4.6	Los panelistas representan a distintos actores sociales (ej. docentes, líderes comunitarios, empresarios).	NO	NO	SI	NO

COMENTARIOS

4.1-4.6. SIMCE: Panelistas internos de perfil técnico mayoritariamente (ej. funcionarios UCE) hacen revisión de distintos aspectos de alineamiento. En ocasiones, los panelistas incluyen externos a la UCE (ej., docentes y especialistas en la disciplina). La validación oficial de los estándares la hace el Consejo Nacional de Educación, el que convoca a externos para evaluar la alineación, entre otros aspectos.
NAEP: Paneles técnico-políticos de amplia representación hacen revisión de alineamiento.
OPAP: Paneles de docentes que juzgan el grado de alineamiento de los ítems con el blueprint. Se hacen paneles distintos, y estándares de desempeño distintos para las comunidades angloparlantes y francoparlantes. Se representan comunidades aborgines en distintas partes del proceso (en Canadá "First Peoples").
PLANEA: Paneles mixtos de perfil técnico hacen revisión de alineamiento, siempre incluyen docentes y administradores de centros educativos, no usuarios externos.

Tabla 5 (cont.)

CRITERIO		Chile: SIMCE	Canadá (Ontario): OPAP	Estados Unidos: NAEP	México: PLANEA
VALIDEZ DE PROCESOS: Se *refiere a la calidad técnica del diseño estudio de alineamiento, al uso apropiado de métodos y procedimientos, y a las calificaciones de los panelistas.*					
Métodos y procedimientos					
4.7	Se realizan estudios independentes o externos para evaluar distintos aspectos del alineamiento (tipo auditoría externa).	NO	SI+	SI+	NO
4.8	El estudio de alineamiento tiene un diseño técnico adecuado.	PARCIAL	PARCIAL	PARCIAL	PARCIAL
Documentación					
4.9	Existe documentación técnica sobre estudios de alineamiento entre el curriculum, estándares de desempeño y pruebas (ej., metodología, resultados).	PARCIAL	SI	SI	SI
4.10	Existe un documento con un modelo formal de alineamiento entre el curriculum, estándares de desempeño y las pruebas.	NO	NO	NO	NO

COMENTARIOS

4.7. NAEP: Estos estudios se realizan regularmente, y son licitados a contratistas externos. OPAP: todos los procedimientos son replicados por contratistas externos con el propósito de auditar los procedimientos; se llevan a cabo en la rendición de cuentas acerca de la calidad del trabajo que se hace con el Ministerio de Educación. PLANEA: no hay auditoria externa formal o técnica, únicamente revision por parte de un comité técnico internacional.

4.8. TODOS: Los programas de evaluación no diseñan un estudio formal de alienamiento. Lo que hacen es implementar una serie de procedimientos para asegurar el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas.

4.9. SIMCE: Documentación parcial sobre alineamiento se encuentra en los informes técnicos. Esta información se refiere principalmente al alineamiento entre el curriculum y las pruebas, entre el curriculum y los estándares de desempeño. No hay información sobre evaluación de proceso. PLANEA: Se documentan todos los procedimientos de análisis curricular, resultados, y retículas, entre otros.

4.10. TODOS: Los programas de evaluación no tienen un modelo formal de alieamiento. Lo que tienen son procedimientos para velar por distintos aspectos del alineamiento entre el curriculum, los estándares de desempeño, y las pruebas.

A continuación, se comenta sobre algunos resultados de mayor interés:

En cuanto al alineamiento entre el curriculum y los estándares de desempeño (Dimensión 1), sólo en Chile y Ontario, los estándares se elaboran tomando como primera referencia el curriculum oficial (Criterio 1.1). Dado que en EEUU. no hay un curriculum oficial nacional, la referencia es el marco de evaluación NAEP. En México, los estándares fueron definidos inicialmente sin referirse al currículum nacional, en una reunión internacional de expertos. Luego, la operacionalización de esos estándares –el establecimiento de puntos de corte, por ejemplo– se hace en estricta relación con la prueba, cuyos ítems tienen su fundamentación en un análisis del currículum oficial.

El nivel de documentación y rigurosidad con el que se vela por el alineamiento en cada dimensión varía sustantivamente. Hay documentación más detallada y exhaustiva sobre los métodos y procedimientos utilizados para velar por el alineamiento entre el curriculum y los estándares de desempeño (Dimensión 1), y entre el curriculum y las pruebas (Dimensión 2). El alineamiento entre los estándares de desempeño y las pruebas (Dimensión 3) recibe menos atención. En Chile, la documentación sobre alineamiento entre el curriculum y las pruebas es parcial (Dimensión 2, Criterio 2.1), restringiéndose a las especificaciones de la prueba y otras tablas de correspondencia. La documentación sobre el alineamiento entre estándares de desempeño y las pruebas (Dimensión 3) es mínima.

Las revisiones de alineamiento entre el curriculum y las pruebas (Dimensión 2) reciben especial atención en todos los casos. Esto es razonable dada la importancia de asegurar que las pruebas midan el curriculum[28], y que por lo tanto sus resultados puedan ser interpretados válidamente en términos de logro de aprendizajes curriculares. Hay dos criterios de esta dimensión en los que Chile se diferencia de los otros casos. El primero se refiere a la falta de lineamientos sobre los contextos en los que los estudiantes deben demonstrar lo que saben y pueden hacer (Criterio 2.5). Por contexto se entiende el tipo de situaciones en las que el estudiante debe demostrar lo que sabe y puede hacer. En las pruebas NAEP de matemáticas, por ejemplo, el marco de evaluación indica si los problemas deben ser de tipo “matemáticas puras” o de tipo “aplicados”.

El segundo criterio en el que Chile se diferencia de los otros casos se refiere a la falta de lineamientos sobre la complejidad cognitiva de los ítems a incluir en la prueba (Criterio 2.6). Este criterio es distinto a las habilidades descritas en las especificaciones de las pruebas SIMCE, ya que toma en cuenta la demanda intelectual que requiere poner en práctica esa habilidad. Así, por ejemplo, dos preguntas que requieren “analizar información” (habilidad) pueden tener distintos niveles de complejidad cognitiva, dependiendo del contexto en el que se pide analizar

[28] Es, después de todo, la principal evidencia de validez de constructo de la prueba.

dicha información. En NAEP, la complejidad cognitiva refiere tanto a las habilidades que se ponen en juego para resolver un ítem, como a su nivel de dificultad. La complejidad cognitiva se usa como un criterio para guiar la elaboración de ítems, asegurando así que los ítems que se elaboren cubran toda la escala de dificultad de la prueba. Esto contribuye al alineamiento entre el marco de evaluación y la prueba. En Ontario y en México, el nivel de complejidad cognitiva también se toma como insumo para elaborar los ítems.

En cuanto al alineamiento entre los estándares de desempeño y las pruebas (Dimensión 3), Chile muestra variaciones importantes respecto de los otros países. A diferencia de Chile, en México y Ontario se usan versiones preliminares de los estándares de desempeño, basadas en el curriculum, para guiar la elaboración de ítems (Criterio 3.1). En EE.UU., esto se hace a partir del marco de evaluación. Estas descripciones preliminares aseguran el desarrollo de un número suficiente y balanceado de ítems que anclan en torno a los puntos de corte (Criterio 3.3), que evalúen los distintos contenidos, habilidades y contextos (Criterio 3.4), y que correspondan a los distintos niveles de desempeño (3.5).

Otra diferencia con Chile es que, en EEUU, Ontario, y México, los estándares de desempeño utilizados para reportar resultados de las evaluaciones están estrictamente referidos a los contenidos evaluados en las pruebas (Criterio 3.2). En Chile, en cambio, los estándares de desempeño también incorporan elementos no evaluados en las pruebas. Así, por ejemplo, en los estándares de matemática de 4o grado se incluye la capacidad de medir la longitud en centímetros, siendo que esto no es medido en la prueba (hacerlo requeriría el uso de una regla, siendo que el SIMCE es una prueba de lápiz y papel que no incluye manipulativos). En Chile, esto se hace así dado que, por ley, los estándares deben dar cuenta de lo que los estudiantes deben saber y poder hacer para demostrar el cumplimiento de los objetivos de aprendizaje estipulados en el currículo vigente.

Respecto a la validez de proceso en los estudios de alineamiento, EEUU se distingue por incorporar una mayor diversidad de panelistas (Criterios 4.1-4.6). Estos panelistas son externos al programa de evaluación, y representan a distintos actores sociales y usuarios de la información (desde grupos étnicos hasta empresarios). En Chile, México y Ontario, los panelistas tienen un perfil más técnico.

En los tres programas de evaluación revisados en profundidad se realizan procedimientos sistemáticos para resguardar el alineamiento en las tres dimensiones. Sin embargo, en ninguno de estos programas existe un documento o protocolo para hacer estudios o revisiones integrales de alineamiento (Criterio 4.10). Más bien, lo que hay son una serie de procedimientos instalados que velan por el alineamiento de cada dimensión. Usualmente estos

procedimientos forman parte del desarrollo de pruebas e ítems, o del desarrollo de estándares de desempeño.

En NAEP y OPAP, las revisiones de alineamiento se realizan íntegramente en forma externa (auditoría), con múltiples consultas y paneles externos de representación técnica y política, y sus procedimientos y resultados se publican en forma íntegra online (Criterio 4.7).

A la luz de estos resultados, y en consideración del contexto chileno, se elaboraron una serie de recomendaciones, que se presentan en el Capítulo de Recomendaciones de este informe.

6. DESCRIPCIÓN DE LAS METODOLOGÍAS SOLICITADAS EN LOS OBJETIVOS ESPECÍFICOS 2 Y 3

La descripción de las metodologías solicitadas en el Objetivo Específico 2 de este estudio ("identificar, describir y analizar metodologías usadas para mantener el alineamiento entre los estándares de contenido, los estándares de desempeño, y las pruebas estandarizadas establecidas") se encuentra en el Capítulo "Catastro sobre evaluación de alineamiento en diferentes sistemas educativos y descripción y análisis de tres casos en profundidad".

La descripción de las metodologías solicitadas en el Objetivo Específico 3 ("Identificar, en base a criterios, metodologías o elementos dentro de éstas, usados para evaluar y mantener el alineamiento entre los estándares de contenido y los estándares de desempeño con las pruebas estandarizadas que puedan ser transferidos o replicados en el sistema educacional chileno actual.") se encuentra en el Capítulo "Recomendaciones para el sistema educacional chileno para mantener el alineamiento entre los instrumentos en cuestión".

7. RECOMENDACIONES PARA EL SISTEMA EDUCACIONAL CHILENO PARA MANTENER EL ALINEAMIENTO

El alineamiento es fundamental para afirmar que las pruebas miden el curriculum, y para interpretar los resultados de las pruebas (puntajes promedios y niveles de desempeño) como logro de los aprendizajes curriculares. Visto de otro modo, también es central a la hora de garantizar que los objetivos del curriculum nacional se articulan de manera consistente con instrumentos como los estándares de desempeño y las pruebas estandarizadas.

La literatura académica sobre alineamiento es abundante, y ofrece variedad de modelos, criterios y procedimientos para realizar este tipo de estudios (Becker, 2005; CCSSO, 2009; Case & Zucker, 2005; Davis-Becker 2013; Martone & Sireci, 2009; Näsström & Henrrikson, 2008; Roach, Niebling & Kurz, 2008; Vockley & Lang, 2009; Webb, 1997, 2007; entre otros). La dimensión de alineamiento que recibe más atención es la de curriculum y las pruebas. Las otras dos dimensiones (curriculum y estándares de desempeño, y estándares de desempeño y pruebas) reciben considerablemente menos atención.

A diferencia de la literatura académica, la información sobre alineamiento proveniente de la práctica profesional es más escasa. En el marco de este estudio se hizo una revisión de la información sobre alineamiento existente en 52 países (seleccionados de entre los participantes en PISA 2012 y Ecuador). Los resultados muestran que en sólo unos pocos países (15%) hay información técnica que trate sobre alineamiento con suficiente profundidad. La mayoría de los países (60%) no tiene información disponible sobre el tema, y en los países que la tienen, usualmente refiere a cuestiones generales sobre alineamiento entre el currículum y las pruebas (ej., especificaciones de la prueba) (25%).

De entre los 52 países revisados, se seleccionaron tres para abordarlos como casos en profundidad: EEUU (NAEP), México (PLANEA[29]), y Ontario, Canadá (Ontario Provincial Assessment Program). Esta revisión en profundidad muestra que, al igual que Chile, estos países realizan procedimientos sistemáticos para velar por el alineamiento en sus distintas dimensiones. Sin embargo, se observan diferencias importantes en los niveles de sistematización de las revisiones hechas.

A continuación, se presentan algunas recomendaciones para el sistema educacional chileno para velar por el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas

[29] Aunque PLANEA data únicamente desde 2013, la prueba es la misma que se llamó EXCALE en años anteriores y parte de su misma serie temporal equiparada. El cambio de nombre se da como parte de la reforma educativa del 2013 en donde también se modificaron las responsabilidades de la Secretaría de Educación federal en el ámbito de la evaluación.

SIMCE, así como para velar por un alineamiento más integral de estos componentes con todo el sistema educativo. Estas recomendaciones se derivan de nuestro análisis de las prácticas en los países estudiados, y en consideración de las características del caso chileno.

Recomendación 1: Adoptar un modelo formal de alineamiento, que sirva de base para realizar revisiones y estudios periódicos y sistemáticos en este tema.

En ninguno de los casos de estudio, ni en Chile, existe un modelo formal para hacer estudios o revisiones integrales de alineamiento. En cambio, hay una serie de procedimientos que velan por el alineamiento del curriculum, estándares de desempeño, y pruebas. Usualmente estos procedimientos forman parte del desarrollo de pruebas e ítems, o del desarrollo de estándares de desempeño.

Adoptar un modelo formal sería de gran utilidad para implementar estudios periódicos y sistemáticos de alineamiento. Estos estudios se podrían realizar con dos propósitos: (1) con fines de monitoreo, durante el proceso de elaboración de cada prueba; (2) con fines evaluativos, después de elaboradas las pruebas. En el primer caso, el estudio de alineamiento debería ser una práctica habitual, y podría estar liderado desde el MINEDUC (ver Recomendación 2). En el segundo caso, estas evaluaciones se podrían hacer en modalidad de auditoría externa (ver Recomendación 3).

La Tabla 6 presenta el modelo de alineamiento sugerido por el equipo de investigación, para que el MINEDUC realice estudios de alineamiento, ya sea con propósitos de revisión o de evaluación de la alineación.

Este modelo permite orientar la recolección de evidencia sobre alineamiento en las tres dimensiones que son de mayor interés para el MINEDUC:

- Dimensión 1: Alineamiento entre el curriculum y estándares de desempeño
- Dimensión 2: Alineamiento entre el currículum y las pruebas SIMCE
- Dimensión 3: Alineamiento entre los estándares de desempeño y las pruebas SIMCE

Para ello, presenta dimensiones y criterios que pueden usarse en un estudio de alineamiento, ya sea con propósitos de revisión o de evaluación. Las dimensiones y criterios utilizados surgen de las orientaciones dadas por el equipo del MINEDUC, de la literatura especializada sobre alineamiento, y de la práctica profesional en programas de evaluación de distintos países. El modelo incorpora fuentes de evidencia de alineamiento propuestas por Davis-Becker (2013), criterios del método Webb y del modelo Achieve, así como criterios y procedimientos propios de la práctica profesional de programas de evaluación de distintos países.

Tabla 6: Modelo sugerido para realizar estudios de alineamiento entre curriculum, estándares de desempeño, y las pruebas SIMCE.

VALIDEZ DE USOS: *Se refiere a la correcta interpretación de los resultados de las evaluaciones como reflejo de los aprendizajes alcanzados en el currriculum nacional.*	
Dimensión 1: Alineamiento entre el curriculum y los estándares de desempeño	
1.1	Los estándares de desempeño se elaboran tomando como primera referencia el curriculum del ciclo evaluado.
1.2	Los estándares de desempeño describen los aprendizajes terminales o "punta de iceberg" del curriculum.
1.3	Todos los contenidos y habilidades descritos en los estándares de desempeño están especificados en el curriculum del ciclo evaluado.
1.4	El estándar de desempeño "Adecuado" describe un piso sólido de conocimientos disciplinarios que deberían saber los estudiantes en el grado evaluado.
1.5	El estándar de desempeño "Básico" describe el "mínimo sin excusas" de conocimientos disciplinarios que deberían saber los estudiantes en el grado evaluado.
Dimensión 2: Alineamiento entre el curriculum y las pruebas SIMCE	
Alineamiento entre el curriculum y el marco de evaluación de la prueba	
2.1	Hay un marco de evaluación que describe los propósitos de la evaluación, interpretación y usos esperados de los resultados, enfoque conceptual del área disciplinaria a evaluar, objetivos curriculares, contenidos, habilidades, contextos/situaciones a evaluar, formatos de la prueba y de las preguntas, tablas de especificaciones, tiempos de evaluación, entre otros.
2.2	Hay un marco de evaluación que cubre los objetivos curriculares factibles de ser evaluados en una prueba de lápiz y papel.
2.3	Hay un marco de evaluación que describe los contenidos del curriculum a evaluar.
2.4	Hay un marco de evaluación que describe las habilidades del curriculum a evaluar.
2.5	Hay un marco de evaluación que describe las situaciones/contextos en el que los estudiantes deben demonstar lo aprendido, según lo indicado en el curriculum.
2.6	Hay un marco de evaluación que describe el nivel de complejidad cognitiva o nivel de dificultad de los ítems de la prueba.
2.7	Hay un marco de evaluación que cubre los "aprendizajes terminales" del área disciplinaria y grado evaluado, según lo indicado en el curriculum.
2.8	Hay un documento técnico que operacionaliza el marco de evaluación y que sirve para guiar el desarrollo de las pruebas e ítems.

Tabla 6 (cont.)

Dimensión 2: Alineamiento entre el curriculum y las pruebas SIMCE	
	Alineamiento entre la tabla de especificaciones de la prueba y las preguntas/problemas/ítems de la prueba
2.9	Cada uno de los ítems de la prueba puede ser clasificado en una celda (combinación contenido y habilidad) de la tabla de especificaciones.
2.10	El conjunto de ítems de la prueba refleja las ponderaciones asignadas a cada uno de los ejes de contenidos de tabla de especificaciones.
2.11	Hay una distribución balanceada de los ítems de la prueba en los distintos subejes de la tabla de especificaciones.
2.12	El conjunto de ítems de la prueba refleja las ponderaciones asignadas a cada uno de los ejes de habilidades de tabla de especificaciones.
2.13	Cada uno de los ítems de la prueba puede ser clasificado en una de las situaciones/contextos descritos en el marco de evaluación.
2.14	Hay un número suficiente de ítems por nivel de complejidad cognitiva o dificultad para el grado evaluado, según lo indicado en el marco de evaluación.
Dimensión 3: Alineamiento entre estándares de desempeño y las pruebas SIMCE	
3.1	Se elaboran descripciones preliminares de los estándares de desempeño, basadas en el curriculum solamente, con el fin de guiar el desarrollo de los ítems de las pruebas.
3.2	Los estándares de desempeño que se utilizan para reportar los resultados de las evaluaciones están estrictamente referidos a los contenidos y habilidades evaluados en las pruebas.
3.3	Hay un número suficiente de ítems que anclan en torno a los puntos de corte de los niveles de desempeño.
3.4	Hay un número suficiente de ítems que evalúan los contenidos, habilidades y contextos descritos para cada nivel de desempeño.
3.5	Hay un número similar de ítems que anclan en cada nivel de desempeño.
VALIDEZ DE PROCESOS: Se *refiere a la calidad técnica del diseño estudio de alineamiento, al uso apropiado de métodos y procedimientos, y a las calificaciones de los panelistas.*	
	Panelistas
4.1	Los panelistas tienen conocimientos disciplinarios y pedagógicos adecuados para realizar la revisión de alineamiento.
4.2	Los panelistas tienen conocimiento de la población de estudiantes evaluadas (ej. nivel de aprendizajes, oportunidades de aprendizaje en la escuela).
4.3	Los paneles incluyen panelistas independientes del equipo que desarrolló el curriculum, las pruebas, y los estándares de desempeño (panelistas externos).
4.4	Los panelistas representan a distintos usuarios de la información de las evaluaciones (ej. docentes de aula, directivos, padres, funcionarios MINEDUC).
4.5	Los panelistas representan la diversidad cultural del país/estado (ej. representantes de distintas regiones y grupos étnicos, de distintos niveles socioeconómicos, hombres y mujeres).
4.6	Los panelistas representan a distintos actores sociales (ej. docentes, líderes comunitarios, empresarios).

Tabla 6 (cont.)

VALIDEZ DE PROCESOS: Se *refiere a la calidad técnica del diseño estudio de alineamiento, al uso apropiado de métodos y procedimientos, y a las calificaciones de los panelistas.*	
Métodos y procedimientos	
4.7	Se realizan estudios independentes o externos para evaluar distintos aspectos del alineamiento (tipo auditoría externa).
4.8	El estudio de alineamiento tiene un diseño técnico adecuado.
4.9	El estudio de alineamiento se implementó de acuerdo a lo esperado.
4.10	El estudio de alineamiento considera la revisión de contenidos, habilidades, niveles de complejidad o dificultad, y contextos.
4.11	Los panelistas pudieron familiarizarse con el curriculum, estándares de desempeño, y pruebas antes de emitir juicios sobre el alineamiento.
4.12	Los panelistas recibieron capacitación adecuada para entender la tarea a realizar.
4.13	Los panelistas tuvieron oportunidades de practicar las tareas a realizar antes de hacerlas "de verdad".
4.14	La metodología de alineamiento considera precauciones para evitar respuestas "socialmente deseables" o alineadas con una opinión dominante.
4.15	Los panelistas tuvieron oportunidades de evaluar su participación en el estudio de alineamiento y de dar sugerencias (ej. formulario de evaluación, discusión grupal).
Documentación	
4.16	Existe documentación técnica sobre estudios de alineamiento entre el curriculum, estándares de desempeño y pruebas SIMCE (ej., metodología, resultados).
4.17	Existe un documento con un modelo formal de alineamiento entre el curriculum, estándares de desempeño y las pruebas SIMCE.
VALIDEZ INTERNA Y EXTERNA: *Se refiere a la consistencia de los resultados, juicios y valoraciones que surgen de distintos panelistas, métodos o procedimientos utilizados en un mismo estudio de alineamiento (validez interna), o utilizados en distintos estudios de alineamiento (validez externa).*	
Validez interna	
5.1	Cada panelista suele ser consistentes en sus respuestas (consistencia intra panelista).
5.2	Los panelistas suelen tener respuestas consistentes entre sí (consistencia entre panelistas).
5.3	La varianza de puntajes entre panelistas es baja.
5.4	Los panelistas llegan a consenso fácilmente.
Validez externa	
5.5	Los resultados del estudio de alineamiento son similares a los resultados obtenidos en otros estudios de alineamiento.
5.6	El estudio de alineamiento incluye paneles internos (ej. con funcionarios del MINEDUC) y externos (ej. con académicos y docentes no relacionados al MINEDUC).
5.7	Paneles independientes llegan a las mismas conclusiones y decisiones.
5.8	Se pidió a los panelistas que clasificaran ítems con características ya conocidas (ej., itemes previamente clasificados en la tabla de especificaciones por el equipo de MINEDUC) para validar sus clasificaciones.
5.9	Los ítems con características ya conocidas fueron clasificados correctamente por los panelistas.

Antes de implementar este modelo, sería importante que el MINEDUC lo revisara y validara junto con la Agencia de Calidad. Esto dado que ambas instituciones juegan un rol clave para asegurar el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas. Adicionalmente es importante que el modelo finalmente acordado fuera actualizado a lo largo del tiempo.

Para implementar un estudio de alineamiento basado en este modelo, se sugiere trabajar con paneles de amplia representación técnica y política. Por ejemplo, con especialistas en la disciplina, en pedagogía, y en evaluación (ej. docentes, directivos de centros educativos, curriculistas, especialistas en evaluación, académicos), así como con usuarios de la información de las evaluaciones (ej., padres, funcionarios municipales), así como otros actores claves (ej. representantes de las distintas regiones del país, representantes del gremio docente, representantes de centros de estudio en educación). Los paneles deben incluir a personas externas al MINEDUC y a la Agencia de Calidad.

Cada panelista debe revisar la documentación pertinente (ej., pruebas y especificaciones de las pruebas) y hacer un juicio respecto al cumplimiento de cada criterio, marcando la opción de respuesta que mejor representa su juicio. Las opciones de respuesta a cada criterio no están definidas en esta propuesta. Posibles opciones de respuesta son SI/NO, “Muy de Acuerdo” a “Muy en Desacuerdo”, escala de puntajes del 1-5, o porcentajes, entre otros. Será tarea del MINEDUC definir las opciones de respuesta que considere más apropiadas.

En la selección de participantes en procedimientos de alineamiento (en forma similar al establecimiento de estándares) es no solo importante incluir experticia técnica o pedagógica, sino también representatividad de comunidades de interés claves. Su exclusión puede debilitar la credibilidad de las evidencias de alineamiento. Los procedimientos de alineamiento, entre otras cosas, evalúan la posible interacción de los estudiantes con tanto ítems de las pruebas, como los estándares de calidad a las que se refieren. La complejidad de los juicios, ha llevado a programas de evaluación como NAEP y OPAP a incluir comunidades de interés clave: los padres de familia de la población estudiantil a examinar y público en general. En ambos sistemas, la justificación de la inclusión de estos grupos se debe a que la validez de los juicios que afectan la política pública, como es el caso de los juicios con respecto al alineamiento de pruebas con las políticas de estándares y evaluación del sistema educativo, debe incluir las comunidades de interés afectadas en la medida de lo posible.

Se reconoce que la inclusión de padres de familia y público general a menudo está en conflicto con otros criterios técnicos para conformar equipos de trabajo: como por ejemplo su experticia en las materias evaluadas, la medición educativa, etc. Por tanto, los costos en cuanto a talleres, por ejemplo, para preparar a los padres y al público para participar en los estudios de

alineamiento (o en el establecimiento de estándares) son mayores. Sin embargo, los sistemas de evaluación que los emplean estiman que los costos se justifican por cuanto contribuyen a la credibilidad de la alineación, y responden a su naturaleza como herramienta de las políticas públicas en educación.

Las respuestas de todos los panelistas se pueden luego analizar en función de su media, moda, y dispersión. Las respuestas de dos o más paneles independientes pueden ser comparadas y también pueden compararse a lo largo del tiempo, con respecto a distintas iteraciones de las pruebas. Con esta información el MINEDUC podrá hacer una valoración integral del grado de alineamiento existente entre el curriculum, los estándares de desempeño, y las pruebas SIMCE con el fin de mejorar contínuamente el alineamiento y la evidencia que permite juzgar el nivel de alineamiento.

Finalmente es importante compartir este modelo de alineamiento con todas las autoridades vinculadas al resguardo de la coherencia del sistema. En el caso chileno es recomendable informar al Consejo Nacional de Educación sobre dicho modelo con el objetivo de que los criterios acordados puedan ser considerados por ellos a la hora de hacer sus propias revisiones de cambios o ajustes curriculares y estándares de desempeño.

Recomendación 2: Realizar revisiones internas para monitorear el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas SIMCE, cada vez que se desarrolla una nueva prueba.

Estas revisiones se pueden hacer en forma interna, con una persona del MINEDUC liderando el proceso, pero incluyendo a miembros de la Agencia de Calidad, y velando por que se recoja la evidencia sugerida para cada uno de los criterios que finalmente conformen el modelo de alineamiento. Idealmente, si se siguen las recomendaciones 1 y 3, se irán refinando y extendiendo los procedimientos internos de alineamiento de acuerdo a los resultados de estudios detallados y auditorias externas. Para ello, es central contar con un protocolo que delimite los roles, responsabilidades y tipo de decisiones que puede tomar cada participante de acuerdo a la institución a la que pertenece.

Recomendación 3: Realizar auditorías externas y periódicas para evaluar el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas SIMCE.

Las prácticas revisadas indican que los estudios de alineamiento deben realizarse regularmente y por una entidad externa, tipo auditoría. Esta revisión externa debería realizarse cada vez que hayan cambios relevantes en el curriculum, en los estándares de desempeño, o en las pruebas. Si bien no existen criterios temporales acordados al respecto, se recomienda la realización de estos estudios de acuerdo a un calendario predeterminado. El foco de la auditoría podría estar

tanto en cuestiones de procedimientos como en los resultados de alineamiento. Los resultados de estas auditorías deben estar disponibles al público.

Es importante hacer notar que los resultados de alineamiento no son un SI o NO absoluto, si no que son un juicio de valor, que indica grados de alineamiento. El fin último de la auditoría es mejorar procesos y tender a mayores niveles de alineamiento. Esto a su vez se traduce en mayores niveles de validez de las inferencias que se hacen a partir de las pruebas. Todo ello finalmente redunda en la credibilidad de las pruebas, y en su capacidad de movilizar al sistema educativo para la mejora, potenciando la coherencia del sistema, favoreciendo los aprendizajes que se consideran relevantes para los estudiantes, que en el caso chilenos están originalmente descrito en el curriculum y estándares de desempeño. En este sentido, una auditoría debe ser vista como una oportunidad para mejorar.

Recomendación 4: En las especificaciones de las pruebas SIMCE, incorporar los criterios de contextos y complejidad cognitiva para guiar la elaboración de ítems.

En Chile, las especificaciones de las pruebas SIMCE no indican los contextos (situaciones) ni la complejidad cognitiva de los ítems que se deben incluir en la prueba. Por contexto se entiende el tipo de situaciones en las que el estudiante debe demostrar lo que sabe y puede hacer. Por ejemplo, contextos de matemáticas puras o aplicadas, contextos en donde solo se entrega información relevante para resolver el problema, o en donde se incluye información no relevante también.

Por complejidad cognitiva, nos referimos a las cosas que el estudiante debe recordar, entender, analizar y hacer para resolver un ítem con éxito. Aunque en el pasado se empleaba con frecuencia la Taxonomia de Bloom[30] para especifiar la demanda cognitiva que un ítem debe tener, en la actualidad se usa con mas frecuencia el sistema de categorización de nivel de demanda cognitiva desarrollada por Norman Webb[31], denominado “Depth of Knowledge” (DOK).

Estos elementos sí están incorporados en EEUU, México y Ontario, y son importantes para asegurar el alineamiento de la prueba con el curriculum. El contexto en el que los estudiantes deben demonstrar lo que saben y pueden hacer es un elemento clave de los modelos de evaluación basados en competencias.

[30] Bloom, B.S., et al. (1956). Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: McKay

[31] Webb, Norman L., et al. “Webb Alignment Tool.” (2005, July 24). Wisconsin Center of Educational Research, University of Wisconsin-Madison. Retrieved from http://www.wcer.wisc.edu/WAT/index.asp

Recomendación 5: Usar niveles de desempeño preliminares para informar el desarrollo de ítems.

En Chile no se usan los niveles de desempeño para informar el desarrollo de ítems. En consecuencia, no se vela por desarrollar una cantidad suficiente de ítems que describan dichos niveles. En México y Ontario, se desarrollan niveles de desempeño preliminares, a partir del curriculum, para informar la elaboración de ítems. En EEUU se hace algo similar a partir del marco de evaluación NAEP (dado que no hay un curriculum nacional en EEUU). No se puede diseñar o implementar una prueba que mida confiablemente los niveles de desempeño, si no se dispone de herramientas que aseguren que los ítems representen una muestra sólida de todos los niveles de desempeño. De acuerdo a lo anterior, sería recomendable que en Chile se usen niveles de desempeño preliminares,o en su defecto, los requisitos mínimos de aprendizaje, para informar el desarrollo de ítems. Esta recomendación promueve el uso de una de las prácticas más usadas para velar por esa confiabilidad.

Recomendación 6: Asegurar la participación amplia de actores dentro y fuera del sistema educativo en los procedimientos para asegurar alineamiento.

Los equipos encargados de analizar el alineamiento entre el currículo nacional, estándares de desempeño y pruebas SIMCE, usualmente han tenido un perfil más bien técnico, correspondiendo a especialistas de curriculum y evaluación del MINEDUC, así como docentes con experiencia en el grado y área evaluada. Usualmente estos paneles no incluyen representantes de la ciudadanía y de las distintas instituciones del país.

La experiencia en otros países indica que docentes, administradores de establecimientos escolares y otros actores dentro y fuera del sistema educativo pueden ser incorporados en las tareas más técnicas, con éxito. Esto facilita el alineamiento curricular, el alineamiento con respecto a prácticas pedagógicas y oportunidades educativas. También contribuye a la legitimación social de los estándares, propicia su conocimiento amplio, e incrementa su “face validity” o validez aparente.

En la recomendación 1, al aboradar la inclusión de padres de familia y público general hemos hecho referencia a los sustentos teóricos de la inclusión de participantes de comunidades no expertas en medición o las asignaturas en los procedimientos de alineación. Corresponde a esfuerzos para aumentar su validez como herramientas de políticas públicas y aprovecha su conocimiento acerca de los fines sociales que se persiguen en las mismas. Como hemos indicado arriba, los costos de la inclusión de estos grupos son mayores que en el caso de poblaciones expertas en términos de su preparación (en talleres, por ejemplo) para participar

en las distintas acciones. Sin embargo, en sistemas como NAEP Y OPAP los costos se consideran congruentes con el valor agregado de incluirlos.

En Chile, los padres de familia y público en general entienden que las pruebas SIMCE tienen un impacto importante en las escuelas y la sociedad en general, y son comunidades con interés manifiesto en la alineación de pruebas precisamente por su naturaleza de herramienta de políticas públicas. La participación de padres de familia y público en general debe ser vista como una estrategia para lograr apoyo de estos actores claves.

Recomendación 7: Revisar inclusión de elementos en los estándares que no cuentan con evidencia proveniente de las pruebas SIMCE.

Chile se diferencia de otros países por incluir en las descripciones de los estándares de desempeño, elementos que no son necesariamente evaluados en las pruebas SIMCE. Por ejemplo, los estándares de desempeño de matemáticas de 4º grado refieren a la capacidad de medir longitudes en centímetros, siendo que las pruebas SIMCE no evalúan esta habilidad. Esto se hace así dada la importancia de que los estándares de desempeño estén alineados con las expectativas curriculares, y no se restrinjan solamente a lo que evalúan las pruebas.

Sin embargo, incluir en los estándares de desempeño elementos no evaluados en las pruebas es, desde un punto de vista psicométrico, una práctica inadecuada. Esto representa un desafío que debe resolverse cuanto antes. Se sugiere considerar alguna de las siguientes opciones, las que han sido ordenadas según su grado de alineamiento con las pruebas SIMCE (y por lo tanto, su grado de validez para interpretar los resultados de las pruebas en función de los estándares de desempeño):

Opción 7.1: Elaborar descripciones de los estándares de desempeño que estén estrictamente alineadas con lo que miden las pruebas. Esto se puede abordar de dos formas: (a) Evaluando todos los requisitos mínimos de aprendizaje descritos en los estándares de desempeño vigentes. Esto implicaría, por ejemplo, evaluar el uso de la regla en el SIMCE; (b) Elaborando o revisando los estándares de desempeño de modo tal que sólo se refieran los aprendizajes efectivamente evaluados por las pruebas. Esto implicaría, por ejemplo, eliminar las referencias al uso de la regla.

Opción 7.2. Elaborar descripciones de los estándares de desempeño que incluyan elementos que no son necesariamente evaluados en las pruebas (práctica actual), agregando una indicación clara y explícita de que este es el caso. Todos los contenidos y habilidades descritos en los estándares (estén o no en las pruebas) deben estar apoyados en evidencias de validez rigurosamente recolectadas y analizadas. Deben especificarse los métodos y fuentes utilizados para su verificación empírica, ya sean pruebas SIMCE,

evaluaciones internacionales, o evidencia de aula. Esto requeriría, por ejemplo, verificar que efectivamente medir longitudes corresponde al nivel de desempeño de matemáticas especificado. Para ello, habría que revisar si hay evidencia sobre esto en las evaluaciones internacionales. De no ser así, habría que consultar con docentes, revisar trabajos y evaluaciones de aula que demuestren que los estudiantes que alcanzan dicho estándar son efectivamente capaces de medir longitudes.

Opción 7.3. Elaborar descripciones de los estándares de desempeño que estén estrictamente alineados a las expectativas curriculares. Esto permitiría incluir en los estándares de desempeño áreas que tradicionalmente no son evaluadas por el SIMCE -- tales como cálculo mental y expresión oral-- pero que constituyen áreas claves del curriculum. Estándares de desempeño así desarrollados podrían describir trayectorias de aprendizaje a lo largo de todo el ciclo escolar.
De seguirse esta opción, sería necesario recoger evidencia de aprendizaje de múltiples fuentes, que respalden los estándares de desempeño. Por ejemplo, resultados de pruebas SIMCE y de evaluaciones internacionales, evaluaciones de aula, trabajos y actividades de los estudiantes, y entrevistas con docentes, entre otros[32]. Estos estándares de desempeño podrían utilizarse como documentos de apoyo profesional. Por ejemplo, en instrumentos orientadores de la evaluación aula.

Recomendación 8: Reforzar el alineamiento entre el curriculum, estándares de desempeño, y pruebas SIMCE, por un lado, con prácticas pedagógicas y experiencias de aprendizaje, por otro.

El uso de los estándares de desempeño con fines de apoyo pedagógico es una tarea aún pendiente en Chile. El rigor técnico con el que se desarrollan estos estándares no tiene un paralelo en mecanismos concretos que aseguren su uso para apoyar a los docentes en sus prácticas pedagógicas, ni para apoyar a los estudiantes en sus aprendizajes. El alineamiento de los estándares con lo que pasa en el aula es fundamental para tender a la mejora educativa.

Para avanzar en esta línea, se recomienda que el MINEDUC:

- Opción 8.1: Realice estudios de alineamiento de los estándares de desempeño con prácticas docentes. Estos estudios deberán, por un lado, indagar en qué medida los estudiantes tienen oportunidades de desarrollar los conocimientos y habilidades descritos en los estándares. También deberán indagar acerca de las consecuencias en la

[32] Adoptar esta recomendación vendría a ser equivalente a introducir instrumentos similares a los "mapas de progreso" que Chile desarrolló en los años 2002 a 2010.

práctica docente que pueden tener estos estándares de desempeño, tanto positivas (enseñanza diferenciada de acuerdo a nivel de aprendizaje) como negativas (estrechamiento curricular). Estos estudios deberían luego informar directamente a la política educativa en cuanto al uso de evaluaciones estandarizadas, a la formación inicial y continúa de los docentes, así como sobre los procesos de supervisión educativa, entre otros. Estudios de este tipo se han realizado para evaluar los estándares de desempeño en el NAEP de EEUU.

- Opción 8.2: Ponga a disposición de los docentes y directivos material de apoyo pedagógico que fomente el alineamiento de los estándares de desempeño con las prácticas pedagógicas, incentivando la implementación de estrategias pedagógicas que reconozcan los diferentes niveles de aprendizaje que se pueden observar en el aula. Por ejemplo, (a) un centro de recursos de aprendizaje online con ejemplos de preguntas SIMCE, tareas, actividades, y evaluaciones de aula, en donde se indique para qué nivel de desempeño son más apropiadas; (b) videoteca con testimonios y "muestras" de cómo otros docentes usan los niveles de desempeño para planificar y hacer sus clases; (c) guías para la planificación de clases, en donde se refuerce la necesidad de ofrecer actividades de aprendizaje apropiadas para estudiantes que tienen distintas necesidades de aprendizaje.

Estas medidas deben resguardar que las evaluaciones de aula no se vean forzadas a transformarse en evaluaciones orientadas a evaluar los estándares de desempeño, sino a visualizar la información provista por los estándares de desempeño como un complemento a la obtenida por los docentes a través de sus propias evaluaciones.

La adopción de estas recomendaciones servirá para asegurar en mayor medida el alineamiento entre el curriculum, los estándares de desempeño, y las pruebas SIMCE. Ello permitirá interpretar, con mayor certeza, los resultados del SIMCE como el grado de cumplimiento de distintos niveles de aprendizaje, según lo indicado en el curriculum nacional. También servirá para asegurar un alineamiento más integral de todo el sistema educativo. Esto es esencial para hacer sinergias que fomenten mayores niveles de aprendizaje, tendiendo así al mejoramiento de la calidad y equidad de la educación.

8. BIBLIOGRAFÍA

AERA, APA, NCME. (2014). Standards for Educational and Psychological Testing. Washington DC: AERA.

Agencia de Calidad de la Educación (2015). Informe Técnico Simce 2013. Descargado en Agosto 2016 desde: http://archivos.agenciaeducacion.cl/documentos-web/InformeTecnicoSimce_2013.pdf

Baker, E. L. (2005). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. *Measurement and research in the accountability era*, 315-335.

Bloom, B.S., et al. (1956). Taxonomy of Educational Objectives, Handbook I: Cognitive Domain. New York: McKay

Case, B., & Zucker, S. (2005). Horizontal and vertical alignment. In *the China–US Conference on Alignment of Assessments and Instruction, Beijing, China: Pearson Education, Inc.*

CCSSO, 2009. Alignment and the States: Three approaches to aligning the National Assessment of Educational Progress with state assessments, other assessments, and standards.

Clarke, M. (2012). What matters most for student assessment systems: A framework paper. Washington DC: World Bank.

Comisión Europea/EACEA/Eurydice, 2015. *La garantía de la calidad en la educación: Políticas y enfoques para la evaluación de los centros educativos en Europa.* Informe de Eurydice. Luxemburgo: Oficina de Publicaciones de la Unión Europea.

Consejo Nacional de Educación (2012). Acuerdo 075/2012 del Consejo Nacional de Educación, Descargado en Agosto de 2016 desde: http://www.cned.cl/public/Secciones/SeccionEducacionEscolar/acuerdos/Acuerdo_075_2012.pdf

Cox, C., Meckes, L. & De Padua, E. (2013). Learning Standards. En OECD, *Learning Standards, Teaching Standards and Standards for School Principals: A Comparative Study* (pp. 18-31), OECD Education Working Papers, No. 99, OECD Publishing.

Davis-Becker, S. L. and Buckendahl, C. "A Proposed Framework for Evaluating Alignment Studies," *Educational Measurement: Issues & Practice* 32, no. 1 (Spring 2013): 23–33,

Fulmer, G. “Estimating Critical Values for Strength of Alignment among Curriculum, Assessments, and Instruction,” *Journal of Educational and Behavioral Statistics* 36, no. 3 (June 1, 2011): 381–402

La Marca, Paul M. “Alignment of Standards and Assessments as an Accountability Criterion,” *Practical Assessment, Research and Evaluation* 7, no. 21 (2001)

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research*, *79*(4), 1332-1361.

Ministerio de Educación [MINEDUC] (2014). Fundamentos Estándares de Aprendizaje Matemática Lenguaje y Comunicación: Lectura II Medio. Documento de trabajo de la Unidad de Curriculum y Evaluación del Ministerio de Educación.

Näsström, G., & Henriksson, W. (2008). Alignment of standards and assessment: A theoretical and empirical study of methods for alignment. *Electronic Journal of Research in Educational Psychology*, *6*(3), 667-690.

Parveva, T., De Coster, I., & Noorani, S. (2009). *National Testing of Pupils in Europe: Objectives, Organisation and Use of Results*. Education, Audiovisual and Culture Executive Agency, European Commission. Available from EU Bookshop.

Phelps, R. P. (2014). Synergies for better learning: an international perspective on evaluation and assessment. *Assessment in Education: Principles, Policy & Practice*, *21*(4), 481-493.

Roach, A. T., Niebling, B. C., & Kurz, A. (2008). Evaluating the alignment among curriculum, instruction, and assessments: Implications and applications for research and practice. *Psychology in the Schools*, *45*(2), 158-176.

Rothman, R. (2003). Imperfect matches: the alignment of standards and tests. Paper commissioned by the Committee on Test Design for K-12 Science Achievement. *Center for Education, National Research Council: National Academy of Sciences*.

Rothman, R., Slattery, J. B., Vranek, J. L., & Resnick, L. B. (2002). Benchmarking and alignment of standards and testing. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Ángeles. Recuperado a partir de http://cse.ucla.edu/products/reports/TR566.pdf

Vockley, M. & Lang, V. (2009). Three Approaches to Aligning the National Assessment of Educational Progress with State Assessments, Other Assessments, and Standards. Documento de trabajo preparado para el Council of Chief State School Officers and the U.S. Department of Education, National Center for Education Statistics.

Webb, N. (2007). Issues related to judging thet alignment of curriculum standards and assessments. Applied Measurement in Education, 20(1), 7-25.

Webb, N. L. (1997). Criteria for Alignment of Expectations and Assessments in Mathematics and Science Education. Research Monograph No. 6.

SITIOS WEB

Eurydice: https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Main_Page

Catalogue of Learning Assessments del Instituto de Estadísticas de UNESCO: http://www.uis.unesco.org/nada/en/index.php/catalogue/learning_assessments

Center on International Education Benchmarking del NCEE: http://www.ncee.org/programs-affiliates/center-on-international-education-benchmarking/

OECD Reviews of Evaluation and Assessment in Education: http://www.oecd-ilibrary.org/education/oecd-reviews-of-evaluation-and-assessment-in-education_22230955

SABER-Student Assessment. Banco Mundial.
http://saber.worldbank.org/index.cfm?indx=8&pd=5&sub=1

9. ANEXOS

ANEXO 1: PLANTILLA EMAIL DE CONTACTO PARA CATASTRO

ESPAÑOL
Asunto: Solicita información sobre Sistema de Evaluacion en <pais>. Referencia: <nombre referencia>

Estimado <Nombre>
<titulo>

Le escribo por sugerencia de <nombre referencia>, para solicitar su ayuda con un estudio sobre sistemas de evaluación en distintos países del mundo, incluido <nombre pais>. Estamos haciendo este estudio con un equipo del State University of New York at Albany, a pedido del Programa de las Naciones Unidas para el Desarrollo (PNUD) y del Ministerio de Educación de Chile.

En concreto, estamos buscando información sobre:

1. Desarrollo, uso y comunicación de estándares de desempeño (o niveles de aprendizaje) asociados a los resultados de las evaluaciones nacionales a gran escala: ¿nos podría confirmar que existe este tipo de estándares en su país? Por ejemplo, son los estudiantes clasificados en niveles Avanzado, Básico, o Menos que básico? De ser así ¿cómo se desarrollan, utilizan, comunican, actualizan y evalúan estos estándares?
2. ¿Qué procedimientos se emplean para asegurar el alineamiento o la coherencia entre estos (a) estándares de desempeño, (b) los instrumentos de evaluación y (c) el currículum nacional (u otro referente relevante)?

Después de revisar el sitio web de su institución, necesitamos complementar la información ya recabada. Por lo mismo, le agradeceré enormemente cualquier información al respecto. Hay algún documento técnico que aborde estos temas? Sería posible acordar una reunión telefónica or por Skype para conversar sobre estos temas?

De antemano muchísimas gracias por su colaboración,
<Nombre>
<Titulo>

ENGLISH
Dear <name>
<Title>

I am contacting you as suggested by <name>, to ask for your help for a study on assessment systems around the world, including <country>. We are conducting this study with a team from the State University of New York at Albany, by request of the United Nations Development Program (UNDP) and the Ministry of Education of Chile.

We are looking for information on:

1. Performance standards or levels. Are results of standardized tests reported using performance levels standards (for example, are students categorized as being at Advanced, Basic, Below Basic, levels – or according to proficiency levels?). If yes, how are these performance standards developed, communicated, used, updated, and evaluated?
2. Alignment methods and procedures. What procedures are used to ensure the alignment among: (a) curriculum/learning standards, (b) achievement tests, and (c) performance standards

After reviewing the website of your institution, we would like to collect more detailed information about these topics. I would greatly appreciate if you could provide me with any additional information. Is there any technical documentation available that you might be able to direct me to or send me? Would it be possible to have a Skype meeting or conference call to discuss these topics?

In advance many thanks for your help.

<Nombre>
<Titulo>

ANEXO 2: CATASTRO DE METODOLOGÍAS DE ALINEAMIENTO

Ver archivo Excel adjunto

ANEXO 3: CONTACTO DE INSTITUCIONES O PROFESIONALES QUE PUEDAN REALIZAR ASESORÍAS EN LA MATERIA.

NAEP:
Teresa Neidorf Smith
AIR -- American Institute of Research
Ha sido la investigadora principal en varios estudios de validez y alineamiento del NAEP, realizados por AIR en contrato con el NCES.
tneidorf@air.org

Ontario:
Richard G. Wolfe, Professor Emeritus, Ontario Institute for Studies in Education – University of Toronto: wolferg@gmail.com

México:
Felipe Martínez Rizo, ex – Director General del INEE (Investigador Honorifico del INEE): felipemartinez.rizo@gmail.com

Margarita Zorrilla Fiero, Consejera, Junta de Gobierno del INEE: margarita.zorrilla@gmail.com

ANEXO 4: LISTADO DE DOCUMENTOS EMPLEADOS EN ESTE ESTUDIO

A continuación se entrega un listado de los documentos electrónicos que fueron recopilados durante el desarrollo de este proyecto.
Cada uno de estos documentos se encuentra disponible en una carpeta electrónica entregada al MINEDUC junto con el envío de este informe.

Pais	Titulo	Autor	Descripción	Nombre de Archivo
Archivo general	Issues related to judging the Alignement of Curriclulum Standards ans Assessments.	Norman Webb	Documento analiza cuatro criterios para analizar el alineamiento.	G_Al_Webb_Four Criteria
Archivo general	Alignment of the national standards for learning lenguages with the common core state standards.		Análisis del alineamiento de estándares	G_Al_Alignment national standard
Canada - Ontario	EQAO's Technical report for the 2013-2014 assessments	Education Quality and accountability Office. EQAO	Este documento entrega los antedecentes técnicos para las evaluaciones de 2013 - 14	Can_Ont_EQAO_Tecnical report_2013
Canada - Ontario	Understanding levels of achivement 2012.	EQAO	Este documento entrega información para vincular la evaluación de aula con las evaluaciones que desarrolla EQAO. Educación Primaria	Can_Ont_EQAO_Und levels PD_2012
Canada - Ontario	EQAO: Ontario's Provincial Assessment Program. Its history and influence.	EQAO	Este documento revisa los antecedentes e influencias de la oficina EQAO	Can_Ont_EQAO_EQAO history_2013
Canada - Ontario	Framework. Assessment of Reading, Writing and Mathematics, junior Division. (Grades 4 - 6)	EQAO	Este documento entrega los antecedentes para las evaluaciones de escritura, lectura y matemáticas para las pruebas para los grados 4 - 6.	Can_Ont_EQAO_Frame Assess_2007
Canada - Ontario	The Ontario Curriculum Grades 1 - 8. Mathematics.	Ministry of Education	Este documento actualiza el curriculum de matemáticas desde los grados 1 al 8.	Can_Ont_Med_Math Curr_2005
Canada - Ontario	The power of Ontario's provincial testing program	EQAO	Este documento describe las evaluaciones llevadas adelante por el EQAO	Can_Ont_EQAO_Prov testing_2012
Canada - Ontario	EQAO's Technical report for 2013-2014 Assessments.	EQAO	Este documento entrega los antecedentes técnicos sobre las evaluaciones desarrolladas por EQAO durante los años 2013 - 2014	Can_Ont_EQAO_Tech Report_2013
Canada –	Understanding levels of achivement 2012. Junior Division	EQAO	Este documento entrega información para vincular	Can_Ont_EQAO_Und Levels JD_2012

Ontario			la evaluación de aula con las evaluaciones que desarrolla EQAO. Junior Division.	
Canada - Ontario	What is the quality of EQAO assessment?	W. Todd Rogers. Education Quality and accountability Office.	Este documento describe el procedimiento y parametros para la elaboración de las evaluaciones impartidas por EQAO	Can_Ont_Todd_EQAO Assess_2013
EE..UU	Constructed alignmed assessments using automated test construction	Porter, Polikoff y otros.	Describe la prueba de un sistema automático de alinación de test evaluació.	EEUU_Al_Porter_Constructed alignmed
EE.UU	A proposed framework for evaluating alignment studies	Susan Davis-Beker y otros	Levanta información para identificar validez y fortaleza de estudios de alineamiento	EEUU_Al_Davis_Evaluating aligment
EE.UU	Three approaches to alignment the national assessment of educational progress with state assessment, other assessment, and standards	Vokley y Lang	Describe tres acercamientos de alineamiento.	EEUU_Al_Vokley_Three approaches
EE.UU	Curriculum mapping in higher education: a vehicle for collaboration	Uchiyama y Radin	Describe una metodología de mapeo de curriculum	EE.UU_AL_Uchiyama_Curriculum mapping
Europa	Issues in aligning assessments with the Common European Framework of Reference.	Spiros Papageorgious	Alineamiento entre sistemas de evaluación europeos	EU_Al_Papageorgious_Europe Aligning
Suecia	Alignment standards and assessment: A theoretical and empirical studies of methods of alignment	Näström y Henriksson	Evalúa modelos de análisis de alineamiento	SW_Al_Nastrom_Alignment strandards

ANEXO 5: PRESENTACIÓN (PPT) CON RESUMEN DEL ESTUDIO Y SUS RESULTADOS.

Ver archivo Power Point adjunto.