INFORME 3: FINAL (REVISADO)

SDP N° 006/2016 “MODELOS INTERNACIONALES DE ELABORACIÓN DE ESTÁNDARES DE DESEMPEÑO Y EL SEGUIMIENTO Y EVALUACIÓN QUE SE LES REALIZA”

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 final es dar cuenta en del estudio realizado por el equipo de investigación en el proyecto "*Modelos Internacionales de Elaboración de Estándares de Desempeño y el Seguimiento y Evaluación que se les Realiza*".

Evaluaciones referidas a estándares de desempeño son una herramienta que se utiliza en muchos países como parte de una arquitectura de instrumentos de política educativa que busca promover la calidad en los resultados de la escolarización.

La elaboración y continuo refinamiento y evaluación de los estándares de desempeño, constituyen elementos de creciente importancia al consolidarse este tipo de instrumentos como parte importante de la política curricular. Las experiencias de distintos países de la OECD y América Latina proporcionan una importante fuente de conocimiento y oportunidad para identificar prácticas ejemplares, relevantes para ser consideradas en Chile.

El objetivo general que guía este estudio es describir y analizar el "proceso de elaboración, seguimiento y evaluación de estándares de desempeño en distintos sistemas educativos a nivel internacional". Este objetivo general se traduce en los siguientes objetivos específicos (Términos de Referencia, ps. 36 y 37):

- "Caracterizar y describir sistemas educativos que elaboran, implementan y evalúan estándares de desempeño en su jurisdicción.

- Caracterizar y describir el proceso de elaboración, implementación y evaluación de estándares de desempeño, incluyendo la evaluación que se realiza respecto al uso de los estándares y las evaluaciones del impacto del uso de estándares en el aprendizaje, en los sistemas educativos seleccionados.

- Analizar comparativamente las características de cada uno de los sistemas educacionales analizados y sus procesos de elaboración de estándares, identificando tanto elementos comunes como particularidades relevantes.

- Generar recomendaciones para el sistema educacional chileno respecto de metodologías de elaboración, seguimiento y evaluación de estándares de desempeño, considerando el contexto y las particularidades del país."

Para responder a estos objetivos, se usó una metodología de trabajo consistente en: (a) Una revisión documental sobre los estándares de desempeño y pruebas estandarizadas, en los países participantes en PISA 2012, más Ecuador; (b) Un análisis en profundidad para doce sistemas educativos (países, provincias o estados); y (c) Un diagnóstico o línea de base sobre la metodología actualmente utilizada en el MINEDUC para el establecimiento de estándares de desempeño con propósitos de comparación. Los doce sistemas educativos seleccionados por el MINEDUC para los casos en profundidad fueron: Australia; la provincia de Ontario, Canadá;

España; Inglaterra; Escocia; México; Holanda; Nueva Zelanda; Perú; EEUU y los estados de New York y Virginia, de EEUU.

Los resultados muestran que en los doce programas de evaluación revisados en profundidad se siguen procedimientos técnicos similares en el establecimiento de estándares, generalmente versiones de las metodologías Bookmark o Angoff, pero con distintas formas de emplear el juicio de actores clave del sistema educativo y del público. Aunque el curriculum oficial es siempre el referente principal en el establecimiento de estándares de desempeño (cuando existe) hay diferencia en los métodos específicos que se siguen para usar ese curriculum. Con respecto a la evaluación o seguimiento de los estándares, hay más variabilidad entre países. Con frecuencia, hay reconocimiento de que los mismos estándares deben ser evaluados con independencia mediante estudios que confirmen su confiabilidad y validez, pero de momento lo más común es simplemente documentar cuán fielmente se siguen los procesos formales de establecimiento de estándares.

Todos los países enfrentan los desafíos de contar con una política curricular dinámica, con cambios y reformas del curriculum oficial a los que debe responder el sistema de evaluación de estándares de desempeño con los dos objetivos, difíciles de compatibilizar, de resguardar la comparabilidad psicométrica en el tiempo para monitorear la trayectoria de mejoría, deterioro o estabilidad en los resultados de la escolarización a lo largo de los años y la necesidad de resguardar una alineación óptima con un curriculum oficial que cambia a lo largo del tiempo. También los sistemas educativos consideran la importancia de evaluar los estándares, es decir, de producir y analizar evidencias de su confiabilidad y validez, para contribuir a su continuo refinamiento y mejora. Otro propósito de estos esfuerzos de evaluación es contribuir a justificar las decisiones que se toman a partir de la evaluación de estándares ante los actores del sistema educativo y la sociedad.

Recomendamos que en Chile se adopten prácticas para asegurar tanto la comparabilidad interanual de los estándares y el alineamiento óptimo con un curriculum cambiante mediante una estrategia dual semejante al que se sigue en el NAEP de los EEUU. También recomendamos prácticas para estrechar la relación de los estándares de desempeño con el trabajo de docentes y otros actores dentro y fuera del sistema educativo, incorporándolos más tanto en el proceso de establecimiento de estándares, como en los procesos de evaluación de los estándares y en la traducción de los estándares en términos que permitan su uso en la generación de oportunidades de aprendizaje en las aulas. En este sentido, dos cosas son importantes: generar estándares que no se limiten a un lenguaje técnico y por lo tanto se hagan más cercanos a los términos usados por sus usuarios potenciales, y producir materiales de apoyo a los estándares con información y recursos que faciliten su uso por parte de los actores clave del sistema educativo.

2. ANTECEDENTES SOBRE EL TEMA

2.1. MARCO CONCEPTUAL SOBRE ESTÁNDARES DE DESEMPEÑO

El objetivo de este capítulo es proveer de un marco conceptual sobre estándares de desempeño, y describir distintos modelos de desarrollo de los mismos. El capítulo se basa en la revisión de distintos programas de evaluación, así como en la literatura especializada sobre el tema.

De acuerdo con Mineduc (2014), los estándares de desempeño definen el criterio o expectativa con que se debe cumplir para ser clasificado en un determinado nivel o categoría de desempeño. Comprenden un conjunto de niveles, en los cuales se clasifica el grado de logro de los aprendizajes de los estudiantes. Estos niveles pueden ser dos, de tipo *pass-fail* (por ejemplo, posee o no los conocimientos necesarios) o más de dos por ejemplo en las evaluaciones TIMSS (*Trends in International Mathematics and Science Study*) y PIRLS (*Progress in International Reading Literacy Study)* que distinguen los niveles Avanzado, Alto, Intermedio y Bajo. Además, estos niveles pueden o no estar asociados a juicios de valor. Por ejemplo, pueden estar nombrados con rótulos que no explicitan un juicio cualitativo respecto de qué tan bueno es ubicarse en cada categoría, rótulos asociados a números, como es el caso de la prueba PISA (*Programme for International Student Assessment*) o como en el caso de NAEP (*National Assessment of Educational Progress*) donde los niveles están nombrados con rótulos que llevan asociado un juicio cualitativo (Avanzado, Competente y Básico). Por su parte, Cizek (2012b) define a los estándares de desempeño como el rendimiento que es necesario tener en una prueba para que un estudiante sea clasificado en una determinada categoría de desempeño (ej. Adecuado, Elemental, Suficiente). El desarrollo de estándares de desempeño se refiere a los métodos y procedimientos a través de los cuales se definen dichas categorías de desempeño.

Los estándares de desempeño suelen considerar los siguientes elementos comunes (CEPPE, 2013):

- Una descripción precisa del desempeño exhibido por el estudiante que alcanza el estándar de desempeño.
- Indicadores del tipo de trabajo característico que muestra el estudiante que logra demostrar el nivel de desempeño que se busca describir en los estándares.
- Ilustraciones del tipo de evidencia que muestra el estudiante que logra el estándar. Esto se hace mostrando las preguntas de la prueba que típicamente es capaz de responder un estudiante que alcanza el estándar. En el caso de evaluaciones de desempeño que no se basan en pruebas de lápiz y papel, la evidencia puede incluir trabajos concretos (proyectos, experimentos) o videos. Estas ilustraciones suelen ir acompañadas de comentarios que explican por qué se considera que representan el logro del estándar.
- Un puntaje de corte que determina si el estándar se ha alcanzado o no.

Así, por ejemplo, los estándares de desempeño de Nueva Zelanda contienen todos estos elementos. En otros casos, como Brasil o los estándares de desempeño desarrollados por el Consejo de Ministros de Canadá, simplemente se listan indicadores que representan los logros implicados en las preguntas que son capaces de resolver los estudiantes en los tramos de puntaje de la prueba asociados a los distintos niveles de logro (CEPPE, 2013).

2.1.1 Métodos para desarrollar estándares de desempeño

Hay una variedad de métodos para desarrollar estándares de desempeño. Todos aspiran a hacer clasificaciones que sean claras, defendibles, justas, reproducibles, y en base a información de calidad. A pesar de las diferencias en los métodos, hay muchas comunalidades también. Los métodos se distinguen principalmente en el tipo de información que utilizan para tomar decisiones, y en los procedimientos seguidos para tomar dichas decisiones.

Más allá del método particular, lo importante es que el método sea apropiado a los propósitos del programa de evaluación y a las características de las pruebas (Cizek, 2012b). El método también debe ser apropiado al contexto en general en que se desarrollan los estándares. Así, por ejemplo, en una evaluación con fines diagnósticos en donde solo hay recursos para dar atención remedial al 10% de los estudiantes con más bajos niveles de desempeño, un método de tipo normativo sería más adecuado. Por otra parte, en una evaluación que reporta resultados en función del logro de expectativas curriculares, un método referido a criterio sería más adecuado.

Por otro lado, la necesidad de involucrar a panelistas que representan a distintas zonas geográficas ha promovido el desarrollo de métodos online. La necesidad de asegurar que los estándares de desempeño de programas nacionales de evaluación estén relativamente alineados con los estándares utilizados en las evaluaciones internacionales, ha fomentado el desarrollo de métodos tipo benchmark.

Los métodos para desarrollar estándares de desempeño pueden clasificarse de distintas maneras: en función de su referente (referidos a normas o a criterios), de los procedimientos específicos que utiliza para fijar los puntos de corte (ej., Angoff, Bookmark, Grupos Contrastantes), de las características de las pruebas (métodos para pruebas adaptativas, para evaluaciones con preguntas abiertas, para evaluaciones con proyectos), del tipo de escala de puntajes (ej. escalas verticales), de la cantidad y tipo de niveles de desempeño, de la modalidad de trabajo utilizada (panel presencial u online), entre otros.

A continuación, se describen brevemente algunos de los métodos más utilizados en diferentes sistemas de evaluación y más mencionados en la literatura especializada.

Métodos de tipo normativo. Estos métodos fijan los estándares de desempeño tomando como principal consideración la distribución de estudiantes en la escala de puntajes. Los estándares se fijan *a posteriori*, identificando el punto de corte que distingue a un determinado porcentaje

de estudiantes del resto de la población evaluada. Estos métodos son utilizados en los programas internacionales de evaluación de LLECE, PIRLS, PISA, y TIMSS, entre otros. Así, por ejemplo, TIMSS fijó originalmente sus puntos de corte de modo tal de clasificar a los estudiantes del conjunto de países participantes en cinco grupos: Avanzado (incluye al 10% de estudiantes de mejor desempeño), Alto (incluye al 30% de mejor desempeño), Intermedio (incluye al 50% de mejor desempeño), Bajo (incluye al 60% de mejor desempeño), y quienes no alcanzan el nivel Bajo.

Métodos referidos a criterio. Estos métodos fijan los puntos de corte tomando como principal insumo lo que se espera que los estudiantes sepan y puedan hacer, según lo establecido en el curriculum o marco de evaluación. Los métodos referidos criterios son los usualmente utilizados en los países que tienen un curriculum nacional como Chile (ej. NMSSA en Nueva Zelanda, ECE en Perú), o algún otro tipo de referente curricular único (ej. NAEP en Estados Unidos).

Métodos de tipo aprobado/reprobado. Son comunes en exámenes con fines de certificación de la educación primaria o secundaria, o exámenes de selección para la educación terciaria. Por ejemplo, en Ontario, Canadá, el OSSLT (Ontario Secondary School Literacy Test) es un examen de certificación de la educación secundaria que clasifica a los estudiantes en dos categorías (Completo o Incompleto).

Métodos con múltiples niveles de desempeño. Son usualmente utilizados en sistemas de evaluación cuyo principal propósito es monitorear el rendimiento de los estudiantes. Para ello, deben describir el desempeño en un espectro amplio de la escala de puntajes. Estos métodos son utilizados en los programas de evaluación de Australia, Chile, Francia, y México, por mencionar algunos países.

A continuación, se describen en mayor profundidad algunos métodos que pudieran ser de especial interés para Chile.

<u>Método Bookmark</u>

El método Bookmark es uno de los más utilizados en los sistemas de evaluación de estudiantes. Los estándares de desempeño se elaboran: (1) definiendo qué significa estar en cada nivel de desempeño (ej., qué es lo que debería ser capaz de hacer un estudiante mínimamente competente para ser clasificado en el nivel Adecuado), siendo este el componente referido a criterio; (2) identificando los ítemes –previamente ordenados en un cuadernillo según su dificultad-- que dicho estudiante debería ser capaz de responder correctamente; (3) identificando el punto de corte en la escala de puntajes asociado al ítem de mayor dificultad que dicho estudiante debería ser capaz de responder correctamente. Este método es ampliamente utilizado en EEUU (tanto a nivel del sistema de evaluación federal como estatal), Hong Kong (China), Inglaterra, México, Nueva Zelanda, y Chile, por mencionar algunos. El método Bookmark está ampliamente descrito en la literatura (por ejemplo, Lewis et al., 2012). También está descrito en documentos del MINEDUC (UCE, 2014).

Aunque existen variantes de este método, un componente fundamental del método es el Cuadernillo de Ítemes Ordenados (OIB – Ordered Item Booklet). Este cuadernillo contiene un conjunto de ítemes ordenados según su dificultad. La dificultad se determina empíricamente, usualmente después de calibrar los ítemes en TRI (Teoría de Respuesta al Ítem). Esta presentación de ítemes ordenados a panelistas en el proceso de fijación de estándares los ayuda a comprender la dificultad relativa de ítemes y enfoca su atención sobre los tipos de ítem que tienen mayor probabilidad de ser contestados con éxito por parte de estudiantes en distintos niveles de desempeño. Las variaciones más importantes de Bookmark buscan, mediante una variedad de estrategias principalmente de análisis gráfico, comunicar con mayor claridad información referida a la dificultad de ítemes para grupos de estudiantes con distintos niveles de desempeño, a los grupos de panelistas.

Una variación del método que podría ser de interés para Chile es el "Mapmark con Dominios". El Mapmark viene a ser el mapa de la escala de puntajes IRT, en donde se muestra por un lado la distribución de las preguntas en la escala de puntajes y, por el otro, la distribución de estudiantes en la misma escala. Esta información es entregada a los panelistas en forma adicional a la información tradicionalmente entregada (ej. cuadernillos con ítemes ordenados de menor a mayor dificultad, ítemes clasificados según contenidos y habilidades evaluados). El Mapmark aporta una dimensión espacial que ha sido valorada en la definición de estándares de desempeño. (Para más información, ver la Ficha NAEP de este informe).

Método Angoff

Este era probablemente el método de desarrollo de estándares más utilizado antes de la introducción del método Bookmark. Es un método referido a criterio, que pide a los participantes identificar los ítemes de la prueba que deberían ser correctamente respondidos por un estudiante mínimamente competente que alcanza un determinado nivel. Estos ítemes son marcados con un "1", y la suma de los "1" constituye el punto de corte. En vez de "1", el método ha variado a la utilización de la probabilidad de respuesta correcta asociada a cada pregunta. El método ha sido modificado para ser utilizado en diversos programas de evaluación, por ejemplo, para trabajar con preguntas abiertas no dicotómicas. Este método es ampliamente utilizado en EEUU, Chile, Corea del Sur, Hong Kong (China), y Nueva Zelanda, por mencionar algunos. Este método es ampliamente descrito en la literatura especializada (por ejemplo, en Plake y Cizek, 2012), así como en documentos del MINEDUC (UCE, 2014).

Aunque, como hemos dicho anteriormente, el Método Angoff ha sido menos utilizado desde la introducción del Método Bookmark, muchos autores consideran que tiene importantes ventajas. La ventaja que con mayor frecuencia se menciona es el hecho que los juicios de panelistas en el caso de Bookmark se hacen sobre la base de toda la prueba. Esto significa que los estándares fijados en ese procedimiento requieren de un esfuerzo completamente nuevo de fijación de estándares cada vez que la evaluación sufre cambios importantes. En cambio, en Angoff los juicios se basan en ítemes, y se pueden usar los juicios de panelistas ítem-por-ítem para volver a calcular puntajes de corte cuando cambian las evaluaciones.

Método Benchmark

Este método surge como respuesta al problema de que el significado asociado a los distintos estándares de desempeño puede variar enormemente de un país a otro, o dentro de un mismo país. Así, por ejemplo, estar en el nivel "Adecuado" en un estado puede tener un significado muy distinto a estar en ese mismo nivel en otro estado. Esto lleva asociado consigo problemas de comunicación, credibilidad, y transparencia, por mencionar algunos.

Philips (2012) muestra las grandes variaciones existentes en EEUU, en el porcentaje de estudiantes que alcanzan el nivel "*Proficient*" en los distintos estados. Paradójicamente, los estados con mejores sistemas educativos son los que tienen menor porcentaje de estudiantes en este nivel, y los con peores sistemas educativos son los que tienen más. Estas diferencias se deben a las distintas exigencias y significado asociado a estar en el nivel en cuestión, y no al desempeño real de los estudiantes.

Para resolver este problema, Philips propone desarrollar estándares de desempeño alineados con estándares internacionales. Para ello, propone hacer un equating entre las escalas de puntajes nacional e internacional, para luego utilizar los puntos de corte de la escala internacional en la escala nacional.

Método para trabajo online

La definición de estándares de desempeño tiene una dimensión tanto técnica como política. La dimensión política es clave para que los estándares sean validados, reconocidos y adoptados por todos los actores del sistema educativo. Velar por la dimensión política requiere convocar a panelistas que representen a distintos grupos de interés, regiones geográficas, y etnias, entre otros. Sin embargo, reunir a esta diversidad de panelistas en un mismo lugar suele ser costoso y difícil de llevar a cabo, dadas las dificultades asociadas a viajes, agenda y logística, entre otros. Como consecuencia, muchas veces los estándares no son revisados por paneles con representación adecuada.

Para atender a este problema, se han desarrollado plataformas de trabajo virtual para el desarrollo de estándares. Entre los procedimientos utilizados hay videos para entrenamiento de los participantes, software para que los participantes ingresen sus datos, además de recursos como el *chat* y videoconferencia. Zieky (2012) anticipa que este tipo de métodos y tecnologías serán de uso cada vez más común, a medida que las tecnologías mejoran y las personas están cada vez más acostumbradas al trabajar y a tener reuniones online.

Más allá del método y procedimientos utilizados, es clave que éstos se implementen en forma sistemática, y cumpliendo con estándares de calidad. Hambleton (1999) propone 20 criterios para evaluar el proceso de desarrollo de estándares de desempeño. Estos criterios incluyen la adecuada composición de los paneles, la adecuada capacitación de los panelistas, la validación/triangulación de resultados, adecuada documentación del proceso, realización de una marcha blanca, y evaluación final del proceso por los panelistas, entre otros.

2.1.2 Impacto de los estándares de desempeño en el sistema educativo

De acuerdo a CEPPE (2013), es difícil establecer las consecuencias de los estándares en sí mismos o de políticas basadas en estándares desvinculándolos del modo en que son evaluados, de las medidas que se adopten a partir de su logro o no logro, y de su relación con sistemas de *accountability*. Dada la estrecha relación entre estándares y evaluación, la mayor parte de la literatura acerca del impacto de los estándares se refiere más bien a las consecuencias de los sistemas externos de evaluación –especialmente cuando se asocian altas consecuencias y medidas de *accountability*-- que a los estándares mismos. Las consecuencias que tiene implementar estándares para el mejoramiento de la calidad de la educación están mediadas por el tipo de medidas o políticas que se asocian a ellos y, en especial, por el modo en que se evalúan y se toman decisiones a partir de los resultados de dichas evaluaciones.

Dos polos en el uso de estándares son, en un extremo, vincularlos con evaluaciones externas con altas consecuencias y medidas de *accountability*, y en el otro, su uso como orientaciones o guía para la enseñanza y el aprendizaje en las escuelas. En este último caso, el otro el tipo de monitoreo externo del logro de los estándares tiene bajas consecuencias para los evaluados ya sea porque no se publican los resultados por escuela, o porque la evaluación es muestral (CEPPE, 2013).

Otro ángulo para analizar el impacto de los estándares en el sistema educacional, es examinando el efecto recíproco que se puede dar entre estándares y medición. En Estados Unidos, se ha argumentado y encontrado evidencia de que cuando los resultados de las pruebas están asociados a consecuencias, los estándares como elementos orientadores de la enseñanza pasan a un segundo plano (Hamilton, 2008, Ravitch, 2010).

Sin embargo, también puede darse el efecto contrario en casos en que las mediciones con altas consecuencias han precedido a los estándares de desempeño. En estos casos, la incorporación de estándares de desempeño puede aportar sentido educativo al reporte de los resultados, al mostrar la proporción de estudiantes que alcanza los aprendizajes descritos en los niveles de logro. Tal es el caso de Chile, donde desde 1995 existe un sistema de medición (SIMCE) que publica resultados por escuela desde 1995. Si bien la prueba estaba alineada al curriculum nacional y sus respectivos objetivos, hasta 2005 solo reportaba puntajes promedio, lo que hacía poco claro para los profesores y para el público en general, la relación entre los resultados de la medición y el aprendizaje que debían alcanzar los estudiantes. Desarrollar estándares de desempeño que describen e ilustran el aprendizaje alcanzado por los estudiantes en distintos niveles de logro y publicar resultados por escuela indicando la distribución de los estudiantes en dichos niveles fue, en este caso, una respuesta a la necesidad de que la medición nacional adquiriera significado pedagógico (Meckes & Carrasco, 2010; Mineduc, 2003; OECD, 2004).

2.2. DIAGNÓSTICO SOBRE DESARROLLO DE ESTÁNDARES DE DESEMPEÑO EN CHILE

2.2.1 Contexto

El sistema de evaluación de la educación chilena se rige por los lineamientos del Sistema de Aseguramiento de la Calidad de la Educación Parvularia, Básica y Media (SAC), el cual fue creado mediante la promulgación de la Ley 20529 de agosto de 2011. El SAC establece como un deber del Estado propender a asegurar una educación de calidad y equidad, entendiendo por esta última que todos los alumnos tengan las mismas oportunidades de recibir una educación de calidad. Además, a través de esta ley se crean dos nuevos organismos: la Superintendencia de Educación y la Agencia de Calidad.

Así, forman parte de este sistema de aseguramiento de calidad cuatro instituciones: Ministerio de Educación, Superintendencia de Educación, Consejo Nacional de Educación y Agencia de Calidad de la Educación. Esta última está a cargo de desarrollar y aplicar el Plan de Evaluaciones Nacionales e Internacionales, el cual es diseñado por el Ministerio de Educación y aprobado por el Consejo Nacional de Educación.

Antes de la promulgación de la ley SAC, la Unidad de Curriculum y Evaluación (UCE) 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 como son denominados en la Ley) asociados a las pruebas SIMCE eran desarrollados por el equipo de evaluación, mientras que los estándares de contenido (el curriculum nacional), estaban a cargo del equipo de curriculum. Tal como se señaló anteriormente, luego de la promulgación de la ley SAC la evaluación nacional pasó a estar a cargo de la Agencia de Calidad de la Educación.

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 Ministerio de Educación al Consejo Nacional de Educación y, luego de que este los aprueba, la Agencia de Calidad de la Educación los emplea para analizar y comunicar los resultados de las pruebas nacionales.

La UCE es también responsable de asegurar el alineamiento entre el curriculum nacional, los estándares de desempeño, y las pruebas SIMCE. Este alineamiento es condición necesaria para que los resultados de las pruebas SIMCE puedan ser interpretados como reflejo de 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 currículo vigente. Sus cuestionarios indagan sobre información del entorno del estudiante, relevante para comprender los niveles de aprendizaje observados. El al artículo 37 de la Ley General de Educación (N.º 20370), establece

que 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 como 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 currículo 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 20529 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 otros diez indicadores de calidad (como, por ejemplo, clima de convivencia escolar, hábitos de vida saludable o tendencia en el tiempo de 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 Desarrollo de estándares de desempeño

En 2004, especialistas en educación de la OCDE realizaron una evaluación de las políticas educacionales en Chile que derivó en recomendaciones para el SIMCE. Coincidentes con las sugerencias de la Comisión SIMCE (Ministerio de Educación, 2003), las recomendaciones principales de la revisión de la OCDE (2004) fueron que el SIMCE debía establecer estándares que permitieran conocer la proporción de la población escolar que alcanzaba los aprendizajes esperados, y que se debían destinar importantes esfuerzos y recursos para capacitar a docentes y directivos en la interpretación y uso de la información.

La promulgación de la ley que da origen al Sistema Nacional de Aseguramiento de la Calidad (Ley SNAC 20.529) en el año 2011 marcó un nuevo hito para el SIMCE, poniendo un mayor énfasis en los estándares de desempeño o estándares de aprendizaje (como se les denomina en la Ley) como mecanismo para asegurar la calidad de la educación y responsabilizar a las escuelas por sus resultados. Esto implicó revisar los estándares desarrollados anteriormente y desarrollar una nueva versión de estos (Ministerio de Educación, 2014).

El desarrollo de los estándares de desempeño en Chile contempla la definición de dos componentes: uno cualitativo y uno cuantitativo. El componente cualitativo corresponde a los tres Niveles establecidos con sus respectivos rótulos, una definición que da cuenta de lo que significa quedar clasificado en cada uno de ellos, y un listado con los requisitos mínimos establecidos para alcanzar los niveles Adecuado y Elemental. Este componente se desarrolla a través de un proceso que considera las expectativas del marco curricular, así como los logros efectivamente demostrados por los estudiantes chilenos. A continuación, se describen los pasos específicos que se siguieron para desarrollar el componente cualitativo de los estándares de 4º y 8º básico y 2º medio para lectura, matemáticas, ciencias naturales y ciencias sociales (la metodología para el desarrollo de estándares de escritura es diferente, pero no se cuenta –hasta el momento- con los documentos oficiales para dar cuenta de ella).

El primer paso para desarrollar el componente cualitativo de los estándares fue en establecer las definiciones generales: la cantidad de niveles que se van a emplear, la exigencia genérica asociada a cada uno de ellos, y los rótulos que se utilizarán para nombrarlos. El segundo consistió en elaborar los requisitos mínimos teóricos para alcanzar cada uno de los niveles a partir del currículo vigente, según la asignatura y grado evaluados. Luego, en un tercer paso se contrastaron los requisitos mínimos teóricos establecidos para cada nivel con la evidencia empírica nacional e internacional disponible, en cada asignatura y grado. Este paso buscaba ajustar las exigencias definidas para que los Estándares resulten desafiantes y alcanzables para los estudiantes y escuelas del país. En cuarto lugar, los requisitos mínimos ya ajustados con evidencia son validados por especialistas y docentes de aula, en sesiones de trabajo para cada asignatura y grado. Finalmente, los requisitos mínimos propuestos se ajustan en base a las observaciones de los especialistas y se obtienen las descripciones de los niveles con los listados de los requisitos mínimos para alcanzar los niveles Adecuado y Elemental, en base a lo cual posteriormente se definirá el componente cuantitativo de los Estándares (Ministerio de Educación, 2014).

Esta descripción es luego empleada para establecer el puntaje en las pruebas SIMCE que permite distinguir entre los estudiantes que alcanzan o no cada nivel de desempeño. El establecimiento de puntajes de corte para los estándares actualmente publicados se hace empleando el método Bookmark[1] (Ministerio de Educación, 2014).

2.2.3 Comunicación y uso de estándares de desempeño

Los estándares de aprendizaje se comunican tanto a través de los resultados de las pruebas SIMCE que corresponda, así como a través de folletos donde se presentan tanto el componente

[1] Ver antecedentes de este documento para un mayor detalle acerca de este método

cualitativo como el cuantitativo, además de ejemplos de preguntas que ilustran el desempeño en cada nivel.

Es importante señalar que nuevos estándares deben ser revisados cada seis años o si se presentan cambios en el referente curricular nacional, ambas condiciones estipuladas por la Ley que regula el SAC. Estos cambios se rigen por el siguiente calendario:

Curso	Área de aprendizaje	Población evaluada	Vigencia estándares de desempeño
2º Básico	Lectura	Aplicación voluntaria	2014 - 2020
4º Básico	Lectura	Censal	2013 - 2019
	Matemática	Censal	2013 - 2019
6º Básico	Lectura	Censal	2017 - 2023
	Escritura	Censal	2017 - 2023
	Matemática	Censal	2017 - 2023
	Ciencias Naturales	Censal	2018 - 2024
	Ciencias Sociales	Censal	
8º Básico	Lectura	Censal	2013 - 2019
	Matemática	Censal	2013 - 2019
	Ciencias Naturales	Censal	2013 - 2019
	Ciencias Sociales	Censal	2013 - 2019
	Educación Física y Salud	Muestral	
	Formación ciudadana	Muestral	
2º Medio	Lectura	Censal	2015 - 2021
	Matemática	Censal	2015 - 2021
	Ciencias Naturales	Censal	2018 - 2024
	Ciencias Sociales	Censal	2018 - 2024
3º Medio	Inglés	Muestral	
4º Medio	Competencias genéricas EMTP	Muestral	

Como ha sido ampliamente documentado en diversas investigaciones internacionales, los sistemas de evaluación basados en estándares con altas consecuencias pueden generar efectos negativos no deseados en el sistema, tales como el estrechamiento curricular o de las prácticas de enseñanza y evaluación de los docentes. Frente a esto, Mineduc (2014) plantea que se tendrán en cuenta los siguientes elementos para evitar este tipo de consecuencias:

- Coordinación con otros instrumentos curriculares y políticas educativas: alineamiento con el marco curricular vigente y sus respectivos programas de estudio; consideración de requisitos de política de ordenamiento o clasificación de escuelas del Sistema de Aseguramiento de la Calidad de la Educación.

- Resguardos para evitar estrechamiento curricular: se elaborarán Estándares para diversas áreas, como Matemática, Lectura, Ciencias Naturales, Historia, Geografía y Ciencias Sociales, Inglés y Escritura, lo cual –según Mineduc (2014)- es en sí mismo un resguardo para hacer frente al posible estrechamiento curricular. Además, para evaluar de manera más integral la calidad de la educación, se integrará la información entregada por los establecimientos respecto a Otros Indicadores de Calidad, como el desarrollo de la autoestima académica y la convivencia escolar, para no restringir la calidad de la educación solo al ámbito del aprendizaje de contenidos y habilidades de las asignaturas.
- Resguardos para que los requisitos mínimos no se conviertan en la meta: los Estándares de desempeño se basan en evidencia cuya exigencia sería lo suficientemente desafiante para movilizar al sistema y en directa asociación al currículo vigente. Además, en documento de difusión que se entregaría a los docentes se presentan los objetivos de aprendizaje del currículo como las metas de aprendizaje y se describirán aquellos conocimientos y habilidades que se espera alcancen aquellos estudiantes que obtienen puntajes significativamente más altos que el puntaje mínimo exigido para lograr el Nivel Adecuado, es decir, aquellos que alcanzan aprendizajes que van más allá de los requisitos descritos, de modo que los aprendizajes no se restrinjan a lo mínimo exigido para alcanzar ese Nivel.
- Resguardos para evitar que se enseñe para las pruebas: Para evitar que los profesores se focalicen en preparar a los estudiantes de manera espuria para obtener mejores resultados en las pruebas SIMCE, sin el consiguiente aprendizaje de los objetivos planteados en el currículo, Mineduc (2014) señala que se debe cuidar la construcción de las pruebas. “El mejor resguardo es contar con un banco de preguntas amplio y variado que permita cubrir correctamente el currículo evaluado, y construir pruebas que no sean predecibles, de modo que obtener buenos puntajes sea sinónimo de haber logrado el aprendizaje del currículo.” (p.186).

Respecto de una adecuada incorporación en el sistema educacional, Mineduc (2014) señala que es necesario diseñar e implementar dispositivos que permitan monitorear el uso e impacto de los estándares de desempeño. Entre estos dispositivos para monitorear el uso de los estándares estarían mecanismos para recibir retroalimentación de parte de los docentes sobre el o los medios de difusión más efectivos para acceder a la información y las posibles dificultades de acceso; sobre cómo utilizan la información que entregan los estándares en su gestión pedagógica y para evaluar si la información disponible es suficiente e identificar qué información se sugiere incorporar. Respecto de la evaluación del impacto de los estándares en el sistema, a largo plazo, se propone que la UCE realice una evaluación que permita dar cuenta si los Estándares han servido para movilizar el sistema hacia el logro de mejores aprendizajes de los estudiantes y si resulta conveniente aumentar su exigencia después de seis años de vigencia.

Sin embargo, de acuerdo a la información entregada en entrevista con miembros del equipo a cargo de la elaboración de los estándares, se plantea que habría deficiencias en la

comunicación de los estándares, puesto que estos no serían suficientemente conocidos en el sistema. Esto se debería a que los estándares de desempeño aun no se presentan de modo articulado con el currículo nacional, sino como un elemento más asociado a la evaluación nacional. Por otro lado, la agencia habría dejado de presentar, en sus comunicaciones en prensa, la distribución de estudiantes en los niveles de los estándares para evitar juicios imprecisos sobre cambios en los resultados a nivel nacional.

Al profundizar en el impacto de los estándares de desempeño en el sistema educacional chileno, no fue posible encontrar estudios que permitieran distinguir el efecto específico de estos estándares en el aprendizaje de los estudiantes. No obstante, fue posible revisar las recomendaciones del equipo de tarea sobre SIMCE (Ministerio de Educación, 2015); las investigaciones de Meckes, Taut & Espinoza (2016), Manzi, Bogolasky, Gutiérrez, Grau & Volante (2014), Elacqua, Martínez, Santos, Urbina, Treviño, & Place (2013) y Flórez (2013); y un artículo escrito por García-Huidobro (2014), donde se señalan algunos elementos a considerar a la hora de revisar el impacto de los estándares en la educación chilena:

- De acuerdo con el equipo de tarea a cargo de la revisión del SIMCE (Ministerio de Educación, 2015), los estándares de aprendizaje conllevan una mejorar significativa en el tipo de información reportada a partir de las pruebas SIMCE, dotándola de mayor significado pedagógico. Sin embargo, aun es necesario que estos estándares puedan ser empleados efectivamente por los docentes para retroalimentar la enseñanza y el aprendizaje.

- De acuerdo con el equipo de tarea a cargo de la revisión del SIMCE (Ministerio de Educación, 2015), se deben revisar las implicancias de considerar la distribución de los estudiantes de las escuelas en los niveles de aprendizaje de los estándares, especialmente cuando se otorga a este indicador un peso mayoritario en la ordenación de escuelas. Al respecto, sugieren verificar si esto genera una atención exclusiva de las escuelas en el grupo de estudiantes de más bajos resultados en desmedro de los demás, o si por el contrario, estas dirigen sus esfuerzos a los estudiantes en todos los niveles de logro.

- De acuerdo a los estudios realizados por Meckes, Taut & Espinoza (2016), Manzi et al. (2014) y Elacqua et al. (2013) la estrategia predominante de las escuelas para mejorar el aprendizaje sería simplemente evaluar más, en lugar de buscar alternativas para fortalecer la enseñanza de sus profesores. Estas evaluaciones tenderían a adoptar un formato similar al de las pruebas SIMCE. De acuerdo a Meckes, Taut & Espinoza (2016), este tipo de estrategias predomina por sobre estrategias organizacionales relacionadas con el desarrollo de capacidades docentes.

- Según estudio realizado por Flórez (2013) acerca de la validez del SIMCE en relación con los propósitos de evaluar, medir o diagnosticar resultados de aprendizaje en función del logro de los objetivos del Marco Curricular, la evidencia obtenida indica que la interpretación debiera estar limitada a ciertos contenidos y ciertas habilidades, especialmente las más básicas y rutinarias, de determinadas áreas del currículum y no extenderse a habilidades de carácter más complejo. Según esta investigadora, además, habría evidencia que pone en entredicho la calidad de las preguntas y las pautas de corrección empleadas en las pruebas SIMCE, lo que podría también poner en cuestión esta interpretación limitada. Si bien esta autora no indaga acerca de los estándares de desempeño específicamente, se podría inferir que habría que tomar ciertas precauciones si se desea incorporar habilidades de orden superior en los estándares

- De acuerdo con García-Huidobro (2014) los estándares promueven una distorsión de lo que se entiende por calidad de la educación. Según este autor, se reduciría la educación a las destrezas medibles por un test, dejando de lado aspectos sociales, políticos y culturales de la educación. Por otro lado, las escuelas destinarían gran parte de sus esfuerzos a preparar los test con el consiguiente empobrecimiento de la vida escolar y de la educación que ellas proveen.

2.2.4 Preguntas para orientar el presente estudio

Al analizar la información disponible sobre el desarrollo, comunicación y uso de los estándares de desempeño en Chile, resulta destacable que todo el proceso de desarrollo de estos se encuentra meticulosamente documentado y se ha llevado a cabo considerando las recomendaciones internacionalmente más aceptadas. Sobre el desarrollo de estándares de desempeño en Chile, el principal desafío pareciera ser el diseño e implementación de procedimientos para una revisión y actualización de los estándares, dado que no se ha podido encontrar información igualmente detallada que anticipe cómo se llevará a cabo este proceso.

Si bien el Ministerio de Educación fue claro en señalar que se tomarían determinados resguardos para evitar consecuencias negativas no deseadas y apoyar un uso adecuado y provechoso de los estándares al interior de los establecimientos educacionales, dichas medidas no han podido ser implementadas del todo o, al menos, no se ha podido obtener evidencia sobre la realización de estas. Es más, la comunicación de los estándares ha sido, incluso según los propios miembros de la UCE, un elemento que se ha abordado de manera insuficiente. No resulta extraño, entonces, que investigaciones y documentos como los presentadas en este documento (Ministerio de Educación (2015); Meckes, Taut & Espinoza (2016), Manzi et al. (2014), Elacqua et al. (2013) y Flórez (2013); García-Huidobro (2014)), alerten sobre impactos

no deseados de una política de evaluación basada en estándares sobre los cuales, hasta ahora, no se han observado medidas paliativas o algunas de ellas dependen de otras instituciones como la Agencia de Calidad (calidad y variedad de las preguntas incluidas en las pruebas, por ejemplo). Es más, entre las medidas señaladas para evitar el estrechamiento curricular se indica que se desarrollarán estándares para diversas asignaturas, pero en el nuevo plan de evaluaciones aprobado en el año 2015 muchas de estas asignaturas pasan a ser evaluadas a través de pruebas de carácter muestral o voluntaria (Inglés en 3º medio y Lectura en 2º básico) o han sido eliminadas (Ciencias Naturales e Historia, Geografía y Ciencias Sociales en 4º básico), por lo que sus respectivos estándares de desempeño dejan de ser requeridos por ley y pierden su efecto mandatorio en los establecimientos educacionales.

En síntesis, se aprecia un destacable desarrollo en lo que refiere al diseño de estrategias para la elaboración de los estándares de desempeño, pero se requiere repensar y profundizar las estrategias tendientes a apoyar una adecuada implementación de estos estándares en el sistema que garantice un uso acorde con el mejoramiento de la calidad de los aprendizajes de todos los estudiantes.

3. DESCRIPCIÓN DE LA METODOLOGÍA DE TRABAJO LLEVADA A CABO

Este estudio describe y analiza el proceso de elaboración, seguimiento e implementación de estándares de desempeño en distintos sistemas educativos. Para ello, se realizaron:

(a) Doce casos de estudio (análisis documental y entrevistas) de sistemas educativos en los que se emplean estándares de desempeño asociados a pruebas aplicadas a gran escala o a nivel nacional.

(b) Un diagnóstico o línea de base sobre el actual proceso de elaboración, seguimiento e implementación de estándares de desempeño en Chile.

A continuación, se especifican las tareas que permitieron caracterizar el proceso de elaboración, seguimiento e implementación de los estándares de desempeño en distintos sistemas educativos, y el monitoreo y evaluación de estos estándares respecto de su implementación, uso e impacto.

3.1. CASOS DE ESTUDIO

Para la realización de estos casos de estudio se analizó y sistematizó la información de doce sistemas educativos (países o estados) seleccionados en conjunto con la contraparte técnica del MINEDUC. Tal como se indica en los TDR que orientan este proyecto, estos sistemas educativos contemplaron, al menos, tres sistemas de Europa, dos de América del Norte, dos de América Latina y uno de Oceanía.

3.1.1 Selección de casos

En conjunto con el MINEDUC, se seleccionaron 12 casos (países o estados) para estudio en profundidad (ver informe 2 de este proyecto para mayor detalle sobre selección de estos casos):

Europa (4): Escocia, España, Inglaterra, Holanda
América del Norte (4): Ontario (Canadá), Estados Unidos (nivel federal), el estado de Nueva York (EEUU), y el estado de Virginia (EEUU)
América Latina (2): México y Perú
Asia y Oceanía (2): Australia y Nueva Zelandia

3.1.2 Procedimientos para recolección de información

Se han utilizaron los siguientes procedimientos de recolección de información:

a. Revisión documental de publicaciones académicas (libros, revistas científicas); informes técnicos de programas internacionales, nacionales y estatales de evaluación de aprendizaje; y sitios web institucionales.
b. Recolección de información vía email, teléfono, o videoconferencias con actores claves (ej. funcionarios de sistemas de evaluación de otros países).
c. En el caso de EEUU a nivel federal, también se sostuvieron varias entrevistas en persona con una profesional que trabaja para el programa federal de evaluación de este país.

3.2. CRITERIOS O CATEGORÍAS DE ANÁLISIS

Se analizó la información levantada aplicando criterios o categorías de análisis que reflejaran las principales características de programas de evaluación y pruebas estandarizadas, así como las metodologías y procedimientos típicamente utilizados para desarrollar estándares de desempeño, utilizarlos, comunicarlos y actualizarlos, y evaluar su impacto. También se recogió información sobre la estrategia de comunicación de los estándares. Ej., mecanismos de comunicación y capacitación, audiencias claves para la comunicación (ej. directivos, docentes, instituciones formadoras de docentes).

Estas categorías ya fueron presentadas y aprobadas por el equipo del MINEDUC a través del informe 1 correspondiente a este proyecto de investigación y corresponden a las siguientes:

- Descripción del sistema de evaluación
    - Nombre de la evaluación
    - Referente orientador de las evaluaciones
    - Organismo responsable del programa de evaluación y del curriculum
    - Áreas disciplinarias evaluadas
    - Grados evaluados
    - Características de las pruebas
- Descripción de los estándares de desempeño
    - Organismo a cargo
    - Características
    - Historia
- Desarrollo de estándares de desempeño
    - Instituciones y profesionales involucrados
    - Metodología
- Comunicación de estándares de desempeño
- Uso de estándares de desempeño

3.3 DIAGNÓSTICO DEL DESARROLLO DE ESTÁNDARES DE DESEMPEÑO EN CHILE

Finalmente, el equipo hizo una revisión de la metodología de desarrollo de estándares de desempeño en el MINEDUC. Para ello, se entrevistó con la contraparte del MINEDUC, se revisó la documentación facilitada por este sobre estándares de desempeño ("Fundamentos Estándares de Aprendizaje Matemática, Lenguaje y Comunicación: Lectura, II Medio" y ("Fundamentos Estándares de Aprendizaje 4º y 8º Básico") y se revisó la documentación disponible en el *website* de la Agencia de Calidad. Esta información ha sido de gran valor para comprender los procesos que actualmente se llevan a cabo en Chile para desarrollar estándares de desempeño, recabar información sobre los desafíos e interrogantes de la contraparte, y orientar las preguntas de investigación del estudio.

4. DESCRIPCIÓN DE LA INFORMACIÓN REQUERIDA DE LOS SISTEMAS SOLICITADOS

A continuación, se presentan doce fichas con información respecto de los estándares de desempeño desarrollados en diferentes países o sistemas educacionales. Estas fichas son:

Europa:
Ficha #1: Escocia
Ficha #2: España
Ficha #3: Inglaterra
Ficha #4: Holanda

América del Norte:
Ficha #5: Ontario (Canadá)
Ficha #6: Estados Unidos (nivel federal - NAEP)
Ficha #7: Estado de Nueva York (EEUU)
Ficha #8: Estado de Virginia (EEUU)

América Latina:
Ficha #9: México
Ficha #10: Perú

Asia y Oceanía:
Ficha #11: Australia
Ficha #12: Nueva Zelandia

FICHA #1
SCOTTISH SURVEY OF LITERACY AND NUMERACY (SSLN) - ESCOCIA
ENCUESTA ESCOCESA DE LENGUAJE Y MATEMÁTICAS

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la Evaluación

Con el objetivo de informar acerca de la calidad del sistema escocés de educación, el gobierno de ese país, las autoridades locales, Education Scotland, la Autoridad Escocesa de Calificación (Scottish Qualifications Authority - SQA), y la Asociación de Directores de Educación de Escocia (Association of Directors of Education in Scotland - ADES) realizan la denominada Encuesta Escocesa de Lenguaje y Matemáticas (Scottish Survey of Literacy and Numeracy - SSLN) para levantar información actualizada sobre una muestra de establecimientos escolares acerca del desempeño de los estudiantes en lenguaje, matemática y actitud hacia el aprendizaje. Su propósito general es apoyar el aprendizaje e informar de manera segura y detallada sobre avances en el aprendizaje.
El SSLN tiene tres objetivos específicos: a) planificar e impulsar mejores políticas para el beneficio de los estudiantes; b) mejorar el entendimiento de los factores que influyen en el desempeño escolar en lenguaje y matemática; c) compartir buenas prácticas orientadas al mejor uso de los recursos en la sala de clase.

1.2 Referente orientador de las evaluaciones.

El SSLN está alineado con el curriculum de Escocia, llamado Curriculum for Excelence (CfE) , que define expectativas para cinco niveles a lo largo de la trayectoria escolar: "early stage level" (últimos dos años de la educación preescolar); "First level" (final del grado 4) ; "Second level" (final grado 7) ; "Third and Fourth level" (grado 8 a 10); and "Senior Phase level" (grado 11 a 13).
Los principios de CfE apuntan a lograr un alineamiento entre experiencias, resultados, aprendizaje, enseñanza y la práctica de la evaluación. En este sentido, todas las evaluaciones, entre las que se cuenta el SSLN, tienen que vincular las necesidades de los estudiantes respecto del curriculum. Las evaluaciones deben seguir, reforzar y promover el aprendizaje y enseñanza de alta calidad, y deben adecuarse, mediante diferentes tipos de evaluación, a las características y necesidades de estudiantes de diferentes etapas.
Para lograr que las evaluaciones apoyen la implementación del CfE, el gobierno escocés elaboró el documento de Marco para las Evaluaciones (Curriculum for Excelence building the Curriculum 5 a Framework for assessments (5)). Este documento es parte de la serie Building the Curriculum y busca apoyar la planificación, el diseño y la puesta en práctica del currículo y las evaluaciones en las escuelas.
El SSLN, como parte del sistema de evaluación, refleja los valores y principios de Curriculum: i) apoyando el desarrollo de conocimientos, aprendizajes, herramientas, atributos y capacidades ; ii) garantizando a los actores del sistema educativo (padres, estudiantes) el progreso en sus aprendizajes; iii) entregando información precisa acerca

de lo logrado; iv) contribuyendo a la planificación de los próximos pasos de los estudiantes; v) e informando sobre progresos en el aprendizaje y en la enseñanza.

1.3 Organismo responsable del programa de evaluación y del curriculum.

La autoridad escocesa de evaluación (Scottich Qualifications Authority - SQA) trabaja junto a otras instituciones para asegurar que los estándares y las expectativas de las evaluaciones nacionales sean consistentes con los valores, propósitos y principios del Curriculum for Excellence (CfE).

SQA junto con "Education Scotland" está encargada del desarrollo de los criterios de evaluación para Lenguaje y Matemática en todos los ciclos evaluados.

El resguardo del alineamiento curricular de las evaluaciones SSLN está bajo la responsabilidad de "Assessment and Qualifications Unit".

1.4 Áreas disciplinares

El SSLN evalúa lenguaje y matemáticas (Literacy and Numeracy) alternadamente. El año 2015 se evaluó matemática.

1.5 Grados evaluados

En educación primaria los grados 4 (P4 / 7-8 años de edad) y 7 (P7 / 10-11 años de edad), y en educación secundaria el grado 10 (S2 / 12-13 años de edad).

1.6 Características de las pruebas

La prueba se realiza anualmente a una muestra representativa a nivel nacional de estudiantes dentro de cada una de las escuelas del país. Busca evaluar el desempeño nacional en lenguaje y matemáticas, de manera alternada en cada grado evaluado.

Este test mide un amplio rango de habilidades, conocimientos y actitudes de aprendizaje, todos contenidos en el CfE. En Lenguaje, los estudiantes deben completar una prueba de lápiz y papel y una prueba online. Además, deben participar en un grupo de discusión que busca evaluar la comprensión y expresión oral (listening and talking) o presentar trabajos escritos hechos en clases. Por último, deben completar un cuestionario. En Matemática, los estudiantes deben completar una prueba de lápiz y papel, participar en una evaluación interactiva entre el estudiante y el profesor que considera alrededor de 12 preguntas sobre cálculo mental, una tarea de estimación y redondeo (estimation and rounding) y una tarea sobre moneda, medición o azar e incertidumbre. Este componente interactivo equivale a 12 puntos de la prueba en cada grado evaluado. Por último, deben completar un cuestionario que recoge información sobre factores asociados a los aprendizajes, tales como actitudes y experiencias de los estudiantes en la clases.

El número de preguntas varía dependiendo del nivel evaluado y cada pregunta equivale a un punto de la prueba. En promedio, 60% de los puntos de la prueba derivan de preguntas de respuesta corta. En el grado 4 (P4) cada cuadernillo tiene 16 preguntas de respuesta corta y una tarea múltiple consistente en 6 preguntas. Para grado 7 y 10, cada cuadernillo contiene 20 preguntas de respuesta corta y una tarea múltiple consistente en 8 preguntas. Estas pruebas tienen una duración aproximada de 40

minutos para P4, y 60 minutos para P7 y S2. En el caso de la prueba de 2015 - prueba de matemáticas – consistió en que los estudiantes debían trabajar sobre dos cuadernillos de una hora de trabajo cada uno (uno con preguntas de desarrollo y otro con preguntas de selección múltiple) y una prueba interactiva online entre el estudiante y el profesor.
Adicionalmente, profesores de los niveles evaluados deben llenar un cuestionario online que busca recoger información sobre las experiencias de los profesores en el desarrollo curricular de las áreas evaluadas. Esto va dirigido a todos los profesores de las escuelas primarias, y hasta 10 profesores que enseñen en segundo año de educación secundaria, los cuales son seleccionados al azar entre diferentes áreas curriculares.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

Son muchos los organismos que trabajan en el desarrollo del SSLN, pero considerando el foco de este estudio, consideraremos preferentemente dos: SQA y "Education Analytical Services Division". SQA está encargada de los procesos operacionales, tales como la impresión, distribución de los materiales de evaluación, administración de la prueba y la captura de los datos. "Education Analytical Services Division" es la institución encargada del diseño, análisis y comunicación de los resultados de SSLN.

2.2 Características

Se han desarrollado cuatro categorías para catalogar los resultados de los estudiantes en las pruebas: "Aun no trabaja al nivel [esperado]", "Trabaja dentro del nivel [esperado]", "Trabaja bien dentro del nivel [esperado]" o "Trabaja muy bien dentro del nivel [esperado]". El nivel se refiere a la expectativa curricular para el grado evaluado. Así, estar en la categoría "Sobre el nivel" significa que el estudiante demuestra competencias más avanzadas de lo establecido en el curriculum, para el grado evaluado. Más allá de estos rótulos, no hay descripciones asociadas a los estándares de desempeño que indiquen lo que los estudiantes deben saber y poder hacer en cada categoría.
Los estudiantes son clasificados en una categoría u otra en función del porcentaje de respuestas correctas. El porcentaje necesario para estar en un nivel varía para cada grado evaluado (4º, 7º, y 10º) y para cada administración de la prueba. Los porcentajes de respuestas correctas necesarios asociados a cada estándar en SSLN 2015 se muestran en la siguiente tabla.

Tabla: Porcentaje de respuestas correctas asociado a niveles de desempeño de Matemáticas en cada grado evaluado el 2015

Nivel de desempeño	Significado del nivel de desempeño	Porcentaje de Respuestas Correctas necesario para ser clasificado en cada nivel		
		4o grado	7o grado	10o grado
Aún no trabaja en el nivel [esperado] (*Not yet working within the level*)	No logra aún ninguno de los resultados esperados para el nivel evaluado	<9%	<19%	<34%
Trabaja dentro del nivel [esperado] (*Working within the level*)	Logra algunos de los resultados esperados para el nivele valuado, pero aun no logra otros	9% - 50%	19% - 50%	34% - 50%
Trabaja bien dentro del nivel [esperado] (*Performing well at the level*)	Logra la mayor parte de los resultados esperados para el nivel evaluado	50% - 75%	50% - 75%	50% - 75%
Trabaja muy bien dentro del nivel [esperado] (*Performing very well at the level*)	Logra casi todos los resultados esperados para el nivel evaluado	>75%	>75%	>75%

2.3 Historia

SSLN reemplazó la Encuesta Escocesa de Desempeño (Scottish Survey of Achievement - SSA), la que había sido implementada desde el 2004 al 2009. SSLN fue desarrollado el 2009 y entrega información desde el año 2011 con el objetivo de apoyar la evaluación del curriculum. Sus resultados por lo tanto no son comparables con los resultados de evaluación de años anteriores al 2011.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones

La institución encargada del análisis y comunicación de los resultados es "Education Analytical Services Division".

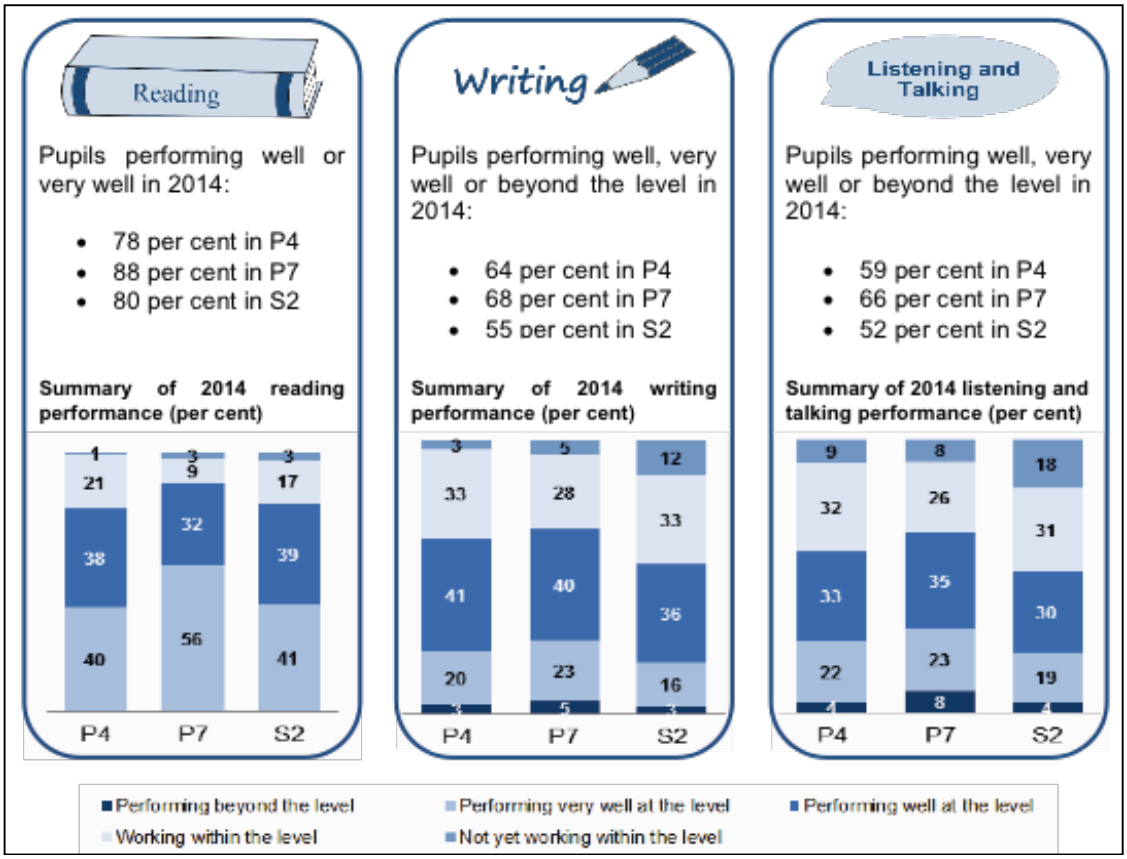
3.2 Metodología

En primer lugar, las preguntas de la prueba son desarrolladas específicamente para el SSLN por profesores en ejercicio y expertos en evaluación. En el caso de ser tareas anteriores de SSA (la prueba anterior a SSLN), estas son reevaluadas en función de los niveles curriculares y los resultados asociados a esos ítemes. En términos de la relación entre la prueba y el currículum, las preguntas se buscan incluir tareas con diferentes grados de complejidad y a lo largo de todos los ejes curriculares establecidos por el curriculum.

En segundo lugar, y en relación los estándares de desempeño mismos, los estudiantes son asignados de acuerdo al porcentaje de preguntas respondidas correctamente a uno de los cuatro niveles. Los porcentajes asociados a cada nivel de desempeño son anualmente definidos en una consulta donde participa Education Scotland, SQA y profesores; esta consulta está basada en el juicio profesional y el análisis de las tareas que son consideradas en la evaluación.

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

El logro de los estándares de desempeño establecidos para 4t^{o} grado (P4), 7^{o} grado (P7) y 10^{o} grado (S2) son reportados a través de un promedio nacional y promedios nacionales desagregados por género y por vulnerabilidad socio económica.
Los resultados son publicados cada año hacia el final del año escolar (abril-junio) siguiente a la administración de la prueba.
Debido al propósito asociado a SSLN de entregar retroalimentación a las escuelas y apoyo al mejoramiento educativo, el diseño de esta evaluación no permite entregar datos a nivel de estudiantes ni agregados a nivel de escuelas o de autoridades locales.



Source: Statistic Publication notice disponible en: http://www.gov.scot/Resource/0047/00475898.pdf [Consultado en diciembre 2016)

A diferencia de la prueba anterior SSA, la SSLN no cuenta con escalamiento vertical que permita evaluar a los estudiantes de un ciclo en relación al logro de los estándares de otros ciclos.

5. USO ESTÁNDARES DE DESEMPEÑO

Basada en los resultados de la evaluación SSLN, Education Scotland produce recursos de aprendizajes disponible para las escuelas y profesores (Professional Learning Resources http://www.educationscotland.gov.uk/learningandteaching/assessment/ssln/resources) con el propósito de facilitar las mejoras en aprendizaje, enseñanza y evaluación al nivel de la sala de clases, en especial en relación a las áreas y conceptos claves, así como en las habilidades de los estudiantes.

6. OTRA INFORMACIÓN RELEVANTE

El año 2016 se ha implementado un cambio en el sistema de evaluación que implica el cese de la encuesta SSLN.
Como parte del Marco Nacional de Mejoras para la Educación de Escocia (National Improvement Framework for Scottish Education), nuevas evaluaciones nacionales estandarizadas se introducirán en las escuelas el año 2017 con el fin de apoyar el desarrollo del juicio profesional de profesores al interior de las escuelas. De este modo, el SSLN será reemplazado por una recopilación de datos de los Niveles de Logro del Currículo para la Excelencia (CfE), basados en juicios profesionales de profesores, que cubrirán a todos los alumnos en los grados 1 (P1), 4 (P4), 7(P7) y 10 (S3).

7. REFERENCIAS

1 http://www.educationscotland.gov.uk/Images/AssessmentforCfE_tcm4-565505.pdf
2 http://www.gov.scot/Topics/Statistics/Browse/School-Education/SSLN
3 http://www.sqa.org.uk/sqa/70972.html
4 http://www.gov.scot/Resource/0050/00500750.pdf
5 http://www.educationscotland.gov.uk/learningandteaching/thecurriculum/buildingyourcurriculum/curriculumplanning/whatisbuildingyourcurriculum/btc/btc5.asp
6 SSLN2015_InformeResultados

8. CONTACTO

Marion MacRury
Scottish Survey of Literacy and Numeracy Team
Scottish Government
2-D South (mail 28)
Victoria Quay
Edinburgh EH6 6QQ
marion.macrury@gov.scot

FICHA #2
Evaluación General de Diagnóstico, España

1. Descripción del Sistema de Evaluación

1.1 Nombre de la evaluación

En el contexto de un país altamente descentralizado, España experimentó con la creación de un sistema de evaluación en dos niveles: La Evaluación General de Diagnóstico a nivel de país, y las Evaluaciones de Diagnóstico a nivel de cada comunidad autónoma. Este sistema de evaluación alcanzó a implementarse en los años 2009 y 2010 respaldado por la Ley Orgánica de Educación de 2006, para ser luego descontinuado en 2013. Este cambio de política se debió a las presiones desde las comunidades autónomas de controlar el sistema de evaluación, sin que haya un ente a nivel de país que las oriente.

La Evaluación General de Diagnóstico (EGD) a nivel de país tenía "...como finalidad contribuir a la mejora de la calidad y la equidad de la educación, orientar las políticas educativas, aumentar la transparencia y eficacia del sistema educativo y ofrecer información sobre el grado de adquisición de las competencias básicas [...] El objetivo inmediato de la Evaluación General de Diagnóstico es obtener datos representativos del grado de adquisición de las competencias básicas del currículo en enseñanza Primaria y Secundaria" (INEE, 2009, p.11). Esta evaluación pretendía también medir tendencias en el tiempo, comparando los resultados de los distintos años evaluados. La evaluación fue administrada a muestras representativas de estudiantes que asistían a centros públicos en los grados evaluados. Reportó resultados a nivel país y de las comunidades autónomas.

La Evaluación Diagnóstica a nivel de las comunidades autónomas, por su parte, "... tendrá carácter formativo y orientador para los centros e informativo para las familias y para el conjunto de la comunidad educativa." (INEE, 2009, p. 9). La ley establecía que "en ningún caso, los resultados de estas evaluaciones podrán ser utilizados para el establecimiento de clasificaciones de los centros". La evaluación se administró a un censo de escuelas y de estudiantes matriculados en centros públicos. Estas evaluaciones reportaban resultados a nivel de centros educativos, los que son analizados por los consejos escolares y el profesorado.

Ambas evaluaciones medían las competencias básicas del curriculum en primaria y secundaria. Los resultados daban cuenta del nivel de logro de las competencias a través del reporte por estándares de desempeño. Para cada competencia también informaban sobre el desempeño por contenidos y procesos cognitivos evaluados.

El plan era que los resultados dieran lugar a compromisos de revisión y mejora educativa. Las dos evaluaciones se administraron en los mismos grados: 4º grado (fin del segundo ciclo de educación primaria) y 8º grado (fin del segundo ciclo de educación

secundaria obligatoria), aunque algunas comunidades autónomas también evaluaron otros grados adicionalmente. Ambas evaluaciones iban a ser administradas anualmente, alternando la evaluación de cada grado. (ej. en 2009 se evaluó 4º grado, en 2010 el 8º grado).

1.2 Referente orientador de las evaluaciones

la Ley Orgánica de Educación de 2006 establecía la realización de evaluaciones de diagnóstico a nivel estatal y a nivel de cada comunidad autónoma. A nivel estatal, el INEE tenía el mandato de realizar la EGD, con la colaboración de los organismos de las administraciones educativas de las comunidades autónomas. A nivel de las comunidades autónomas, las administraciones educativas tenían a cargo las Evaluaciones Diagnósticas, las que usaban como marco de referencia a la EGD.

La ley establecía que la EGD debía evaluar el conjunto de las competencias básicas del currículo. "Estas competencias básicas se relacionan con contenidos curriculares que suponen conocimientos, habilidades y actitudes transferibles y útiles para hacer frente a situaciones y problemas que se presentan en la vida real. En definitiva, se trata de valorar en qué medida la escuela prepara para la vida y forma a los estudiantes para asumir su papel como ciudadanos en una sociedad moderna. Las competencias básicas del currículo se refieren a las capacidades de los sujetos para utilizar sus conocimientos, habilidades y actitudes en la comprensión de la realidad y en la resolución de problemas prácticos planteados en situaciones de la vida cotidiana; en resumen, la aplicación de los conocimientos en un contexto determinado para la resolución de un problema" (INEE, p. 11).

El curriculum de España tiene un componente mínimo común para todo el país (equivalente al 70% del curriculum aproximadamente) y un componente variable definido por cada comunidad autónoma (equivalente al 30% aproximadamente). La EGD se basaba en el componente mínimo común del país.

El curriculum de España establece ocho competencias (enseñanza mínimas), tales como: Competencia en comunicación lingüística, competencia matemática, competencia para aprender a aprender, y Autonomía e iniciativa personal. Estas competencias son transversales a las materias o áreas disciplinarias, y no se pueden identificar con una sola materia.

Para evaluar las competencias, primero se desarrolló un marco de evaluación que da los lineamientos generales para todas las pruebas. Este marco fue encomendado por el Consejo Rector del INEE a un equipo de trabajo externo que trabajó en colaboración con el Grupo Técnico del INEE. En el equipo de trabajo participaron profesionales y académicos, tanto nacionales como internacionales, así como representantes de todas las administraciones educativas de las comunidades autónomas.

El marco especificaba poblaciones y muestras evaluadas, contextos en los que aprenden los estudiantes (ej. contexto socioeconómico), especificaciones técnicas de las pruebas (ej. formato de pruebas y preguntas, pautas para la elaboración de preguntas, tiempo de duración), objeto de evaluación (competencias básicas del curriculum), criterios para el análisis de resultados, e informes y difusión (ej., tipos de informes, audiencias claves).

Respecto de las especificaciones técnicas de las pruebas, el marco de evaluación establecía que "al evaluar competencias, los métodos de evaluación más adecuados han sido los que basan la valoración en la información obtenida a partir de las respuestas del alumnado ante situaciones que exigen la aplicación de conocimientos." (INEE 2009, p. 27).

El marco de evaluación mostraba la relación entre las competencias curriculares y las áreas disciplinarias. Una matriz indicaba la intensidad de la relación entre ambos elementos. El marco también especificaba tres dimensiones claves de las competencias a evaluar: (a) situaciones y contextos en los que se aplica la competencia, (b) procesos cognitivos involucrados; (c) contenidos. Dado el enfoque de competencias, no habían límites estrictos entre las áreas disciplinarias evaluadas.

La matriz de especificaciones era el elemento clave dentro de las especificaciones técnicas de las pruebas. La matriz permitía asegurar el alineamiento de las pruebas con el curriculum estatal. Para ello, operacionalizaba el curriculum, cubriendo los componentes de las competencias básicas establecidas en el curriculum nacional. Para cada prueba y competencia evaluada, la matriz especificaba: 1. Contextos y situaciones, 2. Procesos cognitivos, 3. Contenidos, 4.Actitudes. La matriz también especificaba la cantidad y tipos de ítemes a incluir por proceso cognitivo y contenido evaluados. Cada ítem debía corresponder a una celda (cruce de procesos cognitivos y contenidos) dentro de la matriz, y el conjunto de ítemes debía reflejar las proporciones especificadas en la matriz.

La matriz de especificaciones era la base para la elaboración de preguntas o ítemes de las pruebas. Los estándares de desempeño no fueron utilizados como insumo para esta tarea. Ello dado que la EGD sólo se aplicó una vez en cada grado, y por lo tanto los estándares de desempeño fueron desarrollados en dicha ocasión. No se elaboraron más ítemes después dado que la EGD fue descontinuada.

1.3 Organismo responsable del programa de evaluación y del curriculum.

Organismo responsable de la EGD a nivel estatal: INEE (Instituto Nacional de Evaluación Educativa) del Ministerio de Educación. El INEE opera bajo la dirección de un Consejo Rector a cargo de definir y velar por la correcta implementación de los lineamientos estratégicos de la evaluación (descritos en el marco de evaluación). El Consejo Rector cuenta entre sus miembros a representantes de todas las comunidades autónomas, y opera por consenso.

Organismos responsables de las ED a nivel de las comunidades autónomas: Las administraciones educativas de cada comunidad autónoma.

Organismo responsable del curriculum a nivel estatal: Ministerio de Educación.

El Ministerio de Educación participó en la elaboración de estándares de desempeño aportando profesionales para los paneles de elaboración de estándares.

1.4 Áreas disciplinarias evaluadas

La EGD no mide áreas disciplinarias específicas, si no que competencias que son transversales a distintas áreas disciplinarias. Las competencias evaluadas son:

- Competencia matemática
- Competencia en Comunicación lingüística
- Competencia en el conocimiento e interacción con el mundo físico
- Competencia social y ciudadana

Así, por ejemplo, la prueba de competencia matemática se relaciona fuertemente con el área disciplinaria de matemáticas, pero también se relaciona con las áreas de conocimiento del medio.

1.5 Grados Evaluados

4º grado (fin del segundo ciclo de educación primaria)
8º grado (fin del segundo ciclo de educación secundaria obligatoria)

1.6 Características de las pruebas

La EGD evaluaba las competencias del curriculum estatal a través de administración grupal e individual. La administración grupal utilizaba pruebas de lápiz y papel, con preguntas de selección múltiple, abiertas breves, y abiertas de desarrollo. La administración individual fue utilizada para medir comprensión oral con registro de audio.

El uso de distintos cuadernillos de pruebas que utilizaban un diseño matricial permitía una mayor cobertura curricular. Esto favorecía el alineamiento de las pruebas con el curriculum.

Los resultados de las pruebas fueron reportados por competencia evaluada. Para ello se utilizó una escala de puntajes IRT (Teoría de Respuesta al Ítem) con media igual a 500 puntos y desviación estándar igual a 100 puntos. Esta escala se basó en el modelo de Rash (de un parámetro: dificultad/habilidad).

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

INEE (Instituto Nacional de Evaluación Educativa) del Ministerio de Educación.

2.2 Características

La EGD reportaba niveles de rendimiento (estándares de desempeño) en dos competencias: Competencia de comunicación lingüística, y en la Competencia Matemática.

Los estándares de desempeño o "Niveles de Rendimiento" eran: Nivel 1, Nivel 2, Nivel 3, Nivel 4, y Nivel 5. Además, había un "Nivel < 1", que correspondía a los estudiantes que no alcanzaban el Nivel 1. Estos niveles describían qué caracteriza al dominio de la competencia de los estudiantes cuya puntuación se encuentra en cada uno de dichos niveles (INEE 2009, p. 42). Ver Figura 1.

Figura 1. Estándares de Desempeño de la Evaluación General Diagnóstica de España.

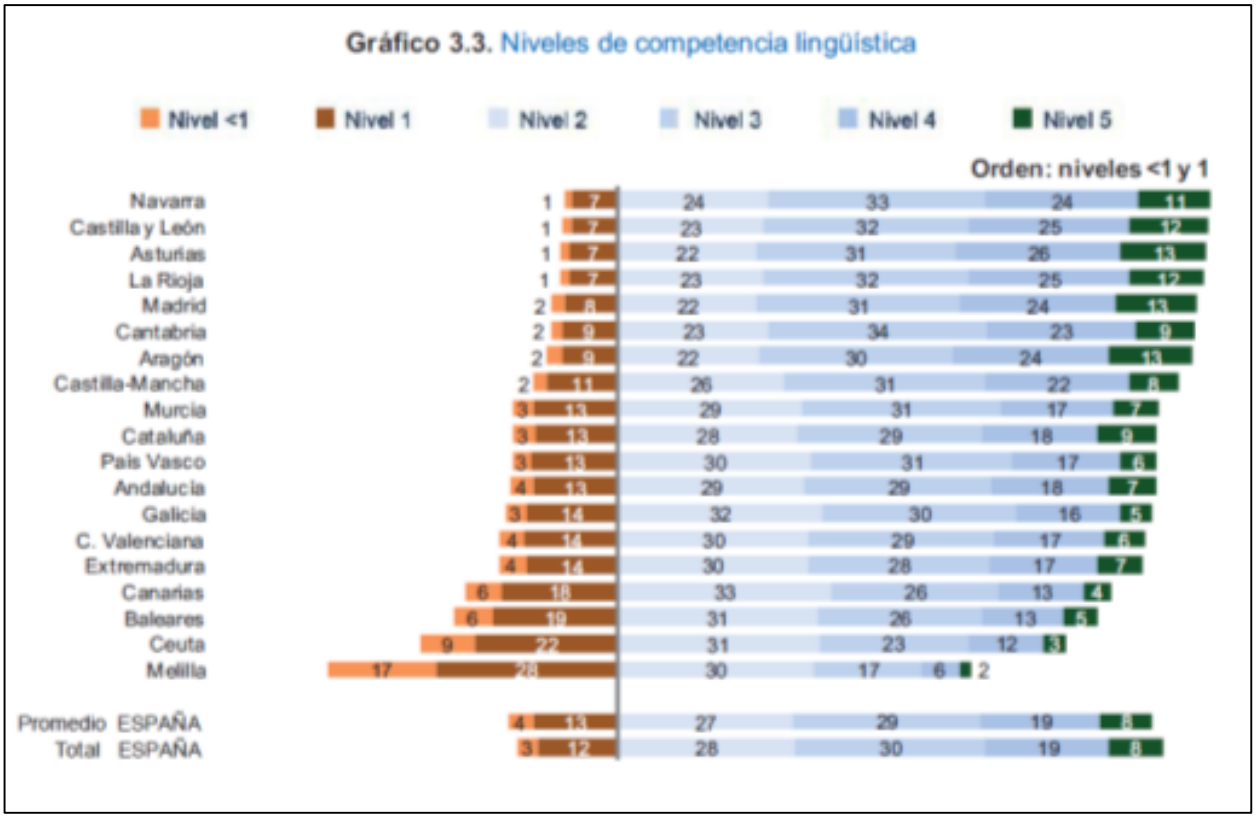


Figura 2. Descripciones asociadas a los Estándares de Desempeño de la Evaluación General Diagnóstica de España.

Tabla 3.1. Descripción de los niveles de Competencia en comunicación lingüística

Nivel	Lo que saben y lo que saben hacer los alumnos en cada uno de los niveles de rendimiento
5 650	En el *nivel 5* los alumnos, además de los conocimientos y destrezas de los niveles anteriores, son capaces de: • sintetizar un texto divulgativo, • elaborar un texto organizando sus partes de forma que expresen una progresión temática que permita seguir la información que se pretende dar, utilizando al mismo tiempo los mecanismos más relevantes para conseguir una cohesión léxica y gramatical, con legibilidad y presentación correctas, • resumir por escrito la información extraída de un texto, eliminando la información no relevante.
4 575	En el *nivel 4* los alumnos además de los conocimientos y destrezas de los niveles anteriores pueden: • organizar la información de un texto reconociendo palabras poco usuales e identificando las partes del mismo y sus relaciones, • valorar las acciones de los personajes a partir del conjunto de informaciones que aparecen aisladas a lo largo de un texto y escribe sus reflexiones, • adaptar, recrear y aplicar a otros contextos lo leído en un texto, • elaborar un texto -a partir de otro leído- con coherencia, cohesión léxica y gramatical y progresión temática de las ideas que quieren expresar.
3 485	En el *nivel 3* los alumnos, además de los conocimientos y destrezas de los niveles anteriores, pueden: • sintetizar información práctica que les permite actuar adecuadamente en la vida cotidiana, • buscar información en textos con mayor extensión y dificultad que los de anteriores niveles, • reflexionar sobre los valores o formas de ser de los personajes a partir de la forma de expresarse de los mismos, • organizar la información y reconociendo e identificando las relaciones entre partes concreta del texto, • integrar el significado de frases literarias por el contexto de texto, • realizar descripciones con una cierta coherencia y cohesión léxica y gramatical basadas en un texto previamente leído usando un vocabulario adecuado a la situación a la que se destina el texto, controlando aspectos como la legibilidad de caligrafía.
2 408	En el *nivel 2* los alumnos, además de los conocimientos y destrezas del nivel anterior, son capaces de: • obtener e identificar la información de un texto corto, sintetizarla eligiendo la frase que mejor lo consigue, • reflexionar sobre la forma de actuar de un personaje ante un hecho concreto, valorar lo que dice y las afirmaciones que se vierten en el texto, • organizar y localizar en un texto hechos y tiempos concretos que les permitan orientarse en situaciones concretas de la vida cotidiana, • sustituir palabras que aparecen en un texto por otras de su vocabulario sin que pierda cohesión, o elegir las palabras que mejor se adecuen a otra que no son de su vocabulario usual, • escribir un texto coherente, bien a partir de frases independientes bien libremente, de forma que coincida con las ideas de un texto leído previamente, • reconocer el significado de símbolos gráficos en la escritura.
1 333	En el *nivel 1* los alumnos tienen capacidad para: • conocer el destinatario de un escrito, • conocer el significado de expresiones comunes insertas en un texto, • Integrar el significado de expresiones comunes insertas en un texto y el de algunas palabras de uso relativamente frecuente, • sintetizar información relativa a hechos concretos o a rasgos más destacados de un personaje, • identificar situaciones y acciones concretas vinculadas a dicho personaje, • reflexionar sobre la definición de un objeto a partir de la información obtenida de un texto, • localizar el espacio en que se desarrolla una historia, organizando la información del texto, • completar un texto de forma coherente con palabras dadas.

Los niveles de desempeño son principalmente referidos al curriculum. El Nivel 1 corresponde a los estudiantes que saben lo más básico del grado evaluado. El nivel 2 describe a un estudiante que cumple con la expectativa curricular, es decir, a un estudiante competente. Los demás niveles fueron fijados a la misma distancia (rango de puntajes) que puntos de corte de los niveles 1 y 2 (70 puntos de la escala de puntajes, o 0.7 desviaciones estándares, aproximadamente).

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

La elaboración de las descripciones asociadas a los puntos de corte es hecha por los expertos en la competencia que han colaborado en la preparación de la prueba. El INEE conformó paneles de especialistas internos, y externos (personal del Ministerio de Educación, académicos, especialistas internacionales).

3.2 Metodología

A. Metodología para definir los puntos de corte

Para definir los puntos de corte asociados a los estándares de desempeño, se usó un procedimiento similar al bookmark. Primero, se ordenaron las preguntas de cada prueba según su nivel de dificultad. Segundo, se marcaron los ítemes que correspondían al Nivel 1 y 2. Tercero, se identificaron los puntos de corte en la escala de dificultad que correspondían a los ítemes que separaban los niveles 1 y 2 (INEE 2010, p. 38).

B. Metodología para elaborar las descripciones

Los niveles se describen de acuerdo con los ítemes que en función de su nivel de dificultad se sitúan en cada uno de ellos. En la "elaboración de las descripciones se tiene en cuenta, primero, las relaciones entre los criterios de evaluación del currículo y las dimensiones de cada competencia que se presentan en el marco de la evaluación; segundo, las unidades de evaluación y las preguntas correspondientes elaboradas a partir de esos criterios de evaluación; tercero, la distribución resultante de los ítemes en cada nivel de dificultad y, finalmente, qué caracteriza al grado de adquisición de la competencia por parte de los alumnos." (INEE 2009, p. 69). Estas descripciones están estrictamente basadas en las preguntas utilizadas en las pruebas.

Los estándares de desempeño fueron revisados y aprobados por las comunidades autónomas.

Dado que la EGD fue descontinuada, los estándares de desempeño elaborados nunca fueron actualizados, revisados ni auditados.

No se encontró información más detallada sobre la metodología utilizada.

4. Comunicación estándares de desempeño

Los estándares de desempeño fueron comunicados a través de los informes de resultados de las dos evaluaciones que se alcanzaron a hacer. No se publicaron otros documentos para apoyar la comprensión o uso pedagógico de los estándares.

5. Uso estándares de desempeño

El plan era que los resultados se usaran para elaborar planes de mejora. Sin embargo, este componente quedó sin implementar, luego de que la EGD se descontinuara.

6. Referencias:

Castillo, J.; Martín-Montalvo, J. González, J. Spain. In Mullis, I; Martin, M.; Minnich, C.; Stanco, G.; Arora, A.; Centurino, V.; Castle, C. (2012). TIMSS 2011 Encyclopedia: Education Policy and Curriculum in Mathematics and Science. Vol. 2. TIMSS and PIRLS International Study Center: Lynch School of Education, Boston College.

Eurydice. 2009. National Testing of Pupils in Europe: Objectives, Organization, and Uses of Results. Brussels: Educational, Audiovisual and Culture Executive Agency.

INEE 2009. Evaluación General de Diagnóstico 2009. Marco de Evaluación. Gobierno de España, Ministerio de Educación. http://www.mecd.gob.es/dctm/ievaluacion/evaluaciongeneraldiagnostico/egd-2009-marco-evaluacion.pdf?documentId=0901e72b8044a2e5

INEE 2010. Evaluación General de Diagnóstico 2009. Educación Primaria. Cuarto Curso. Informe de Resultados. Gobierno de España, Ministerio de Educación. http://www.mecd.gob.es/dctm/ievaluacion/evaluaciongeneraldiagnostico/pdf-completo-informe-egd-2009.pdf?documentId=0901e72b8015e34e

INEE 2011. Evaluación General de Diagnóstico 2010. Educación Secundaria Obligatoria. Segundo Curso. Informe de Resultados. Gobierno de España, Ministerio de Educación. http://www.mecd.gob.es/inee/publicaciones/evaluacion-diagnostico.html#EGD_2010_2

Sitios web:

INEE -- Instituto Nacional de Evaluación Educativa:
http://www.mecd.gob.es/inee/portada.html

Evaluación General de Diagnóstico:
http://www.mecd.gob.es/inee/publicaciones/evaluacion-diagnostico.html

7. CONTACTO:

Ruth Martín Escanilla
Jefe de área de evaluación
Instituto Nacional de Evaluación Educativa
Ministerio de Educación, Cultura y Deporte
http://www.mecd.gob.es/inee
Paseo del Prado, 28. 4ª planta. Despacho 417
28014 MADRID- ESPAÑA
Tfno. +34 91 7459210 (Ext. 73724)
Fax . +34 91 7459249
ruth.martin@me

Ruth Martín fue entrevistada en video conferencia via Skype.

FICHA #3
SISTEMA NACIONAL DE EVALUACION ESCOLAR
Inglaterra

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

Inglaterra tiene un sistema de evaluación escolar asociado a la rendición de cuentas por parte de las escuelas y de altas consecuencias a partir de sus resultados. Es uno de los sistemas escolares más antiguos en el mundo donde las escuelas enfrentan la intervención o el cierre en base a la evaluación que, entre otros, factores considera el desempeño de sus estudiantes evaluado a través de la aplicación de pruebas estandarizadas. El sistema de evaluación inglés cuenta con varias décadas más de desarrollo que la mayoría de sistemas de evaluación en el mundo.

1.1 Nombre de la evaluación

Las pruebas administradas durante la educación primaria son conocidas por el nombre National Curriculum Assessments, aunque coloquialmente se les llama SAT. La prueba administrada en educación secundaria es conocida por el nombre GCSE (General Certificate of Secondary Education).

1.2 Referente orientador de las evaluaciones.

El referente para la actual evaluación de los estudiantes es el currículum nacional diseñado por el *Department for Education* (DfE) e introducido en 1988 por la ERA (Education Reform Act).[2]

1.3 Organismo responsable del programa de evaluación y del currículo.

El *Department for Education* (DfE) es el organismo encargado del diseño curricular y de encargar el desarrollo de pruebas en educación primaria y secundaria.

Durante la *educación primaria* las evaluaciones externas a la escuela están bajo la responsabilidad de *Standards and Testing Agency* (STA) que es una agencia ejecutiva dependiente del DfE y encargada de proveer un sistema nacional de evaluación, pruebas y moderación para medir el progreso de los estudiantes desde su incorporación al sistema educacional hasta finalizar la educación primaria. Esta agencia es responsable de establecer y mantener los estándares de desempeño asociados a las pruebas que se aplican en dichos momentos.[3]

[2] Para más información sobre esto ver:
https://form.education.gov.uk/fillform.php?self=1&form_id=cCCNJ1xSfBE&type=form&ShowMsg=1&form_name=Contact+the+Department+for+Education&noRegister=false&ret=%2Fmodule%2Fservices&noLoginPrompt=1

[3] Para más detalles ver:
https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538564/STA_Annual_report_and_accounts.PDF

Durante la *educación secundaria* la institución a cargo es la Office of Qualification and Examinations Regulations (Ofqual), resguardando la comparabilidad entre las evaluaciones y los estándares de desempeño que administran los Consejos de Examinación (Exam boards)[4].

1.4 Áreas disciplinares

Matematica, lectura y escritura (grmática, ortografía y puntuación).

1.5 Grados evaluados

El sistema Inglés de evaluación nacional cuenta con una serie de pruebas estandarizadas censales para medir el desempeño de los estudiantes tanto en la educación primaria y secundaria. Estas pruebas son una combinación entre pruebas nacionales y pruebas que desarrollan los profesores para cada estudiante. Este conjunto de pruebas son administradas censalmente a los alumnos de la educación primaria y secundaria al finalizar los ciclos educacionales: Early Stage (Reception / 4-5 años), Key Stage 1 (Year2 / 6-7 años), Key Stage 2 (Year6 / 10-11 años) y Key Stage 4 (Year 11 / 14-16 años). No existe una evaluación al final de Key Stage 3.

1.6 Características de las pruebas

La primera evaluación en la trayectoria escolar diseñada por la STA es el test **Phonics Screening** (test de fonemas) que evalúa a todos los estudiantes de Year 1 (5-6 años). El propósito principal de esta evaluación es monitorear el cumplimiento del estándar nacional y hacer que las escuelas rindan cuenta del logro alcanzado por sus estudiantes. El Phonics Screening lo administra el docente del curso evaluado en base a la lectura de 40 palabras en idioma inglés definidas nacionalmente. Una peculiaridad de este test implementada en 2015 es que considera la lectura de palabras extrañas (aliens words). Dichas palabras, a pesar de que parecen similares a otras palabras en uso, son palabras que no existen en el idioma de la prueba y de este modo desafían al estudiante a leer los Phonics (sonidos) que les solicitan leer y así evitar la lectura por similitud o asociación a palabras conocidas previamente. El Phonics screening se repite en Year 2 (6-7 a años) a todos los estudiantes que no cumplieron el estándar nacional el año anterior. Considerando que esta prueba se repite dos años seguidos, la primera aplicación puede además implicar un propósito diagnóstico que permite a las escuelas y docentes reforzar la enseñanza de fonemas con aquellos estudiantes que deben lograr el desempeño esperado en la segunda oportunidad.

---

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/477453/STA_Business_Plan_2015-16.pdf

[4] Para más detalles generales sobre Ofqual ver: https://www.gov.uk/government/organisations/ofqual/about

Adicionalmente, en Year 2 y Year 6 se desarrollan las pruebas SAT, que son evaluaciones también diseñadas por la STA y administradas por los docentes, cuyo propósito es monitorear el cumplimiento del estándar de desempeño nacional esperado para ese momento de la trayectoria escolar en las áreas evaluadas (Inglés, Matemática y Ciencias). Adicionalmente, los resultados de las pruebas de Year 6 pueden ser utilizados como información diagnóstica por la escuelas secundarias donde emigran los estudiantes de primaria. Algunas escuelas secundarias usan esta información para distribuir a los estudiantes de acuerdo su nivel de desempeño en una o varias áreas, principalmente matemática.

La **prueba SAT administrada en Year 2**, final del primer ciclo de primaria (KS1), implica una prueba nacional de aplicación censal que incluye un test estandarizado y el juicio de los profesores en base a su trabajo con el estudiante en las áreas de matemáticas, inglés (lectura, gramática, escritura, dictado y puntuación) y ciencias. Cada prueba contempla dos cuadernillos.

La **prueba SAT administrada en Year 6** (final del segundo ciclo de primaria), implica una prueba estandarizada nacional de aplicación censal que se aplica en las áreas de matemática, inglés (lectura, gramática, escritura, dictado y puntuación) y ciencias. Esta prueba, si bien sigue siendo administrada internamente por la escuela, ya no considera el juicio de los profesores. Al igual que en la prueba de KS1, cada uno de los tests estandarizados contempla dos cuadernillos distintos y el puntaje final bruto es la suma de respuestas correctas entre esos dos cuadernillos.

A diferencia de lo que sucede en la educación primaria, en la educación secundaria no existe una agencia dependiente del DfE que diseñe las pruebas, como lo es la STA. En secundaria las pruebas son diseñadas y aplicadas por diferentes 'Exam boards' que se encargan de desarrollar una oferta amplia de pruebas de evaluación que luego las escuelas tienen el derecho de escoger. Cada escuela escoge el 'Exam board' que quiere que examine a sus estudiantes. Como ya se mencionó, es Ofqual el que resguarda la comparabilidad entre los distintos dispositivos de evaluación.

Los test aplicados en educación secundaria se conocen como el General Certificate of Secondary Education (GCSE), aplicados en Year 11 (14-16 años). Adicionalmente, se administran los A levels en Year 12 (17-18years), AS Levels y Vocacional qualification. El propósito principal de estos exámenes es certificar el rendimiento alcanzado por los estudiantes al terminar su educación secundaria y aportar información para el ingreso de los estudiantes a la educación superior técnica y universitaria.

El referente para la elaboración de los GCSE es el currículum nacional diseñado por el DfE. A través de los GSE se evalúan tres asignaturas obligatorias: matemáticas, inglés y ciencias. También pueden incorporar las siguientes áreas opcionales: Historia, Geografía, Lenguaje, ICT, Educación Religiosa, Educación Física, Música, Teatro, Arte, Diseño y Tecnología, y Salud Personal, Social y Educacional.

2. DESCRIPCIÓN ESTÁNDARES DE DESEMPEÑO

2.1 Organismos a cargo

En *educación primaria* la STA elabora los estándares de desempeño. En *educación secundaria*, el DfE elabora los estándares de contenido y Ofqual se encarga de mantener los estándares y la validez de evaluaciones GCSE, A levels, AS Levels y Vocacional Qualification.

2.2 Características de los estándares

En *educación primaria* cada evaluación tiene un único estándar que se cumple al alcanzar o superar los 100 puntos en las pruebas SAT. Durante el 2016, los puntajes obtenidos por los estudiantes en las pruebas (la suma bruta de las respuestas correctas de los estudiantes en los dos cuadernillos) se convirtieron en puntajes estandarizados (scales scores) en base a la dificultad de la prueba durante ese año (que varía ligeramente cada año). Una vez estandarizados los puntajes, en el caso de Year 2 el resultado puede variar entre 85 y 115 puntos, donde 100 marca el logro del estándar de desempeño. En el caso de Year 6 el puntaje de la prueba varía entre 80 y 110, y nuevamente es 100 el puntaje de corte que refleja el logro del estándar.

En relación a los exámenes GCSE en *educación secundaria*, Ofqual es la agencia encargada de regular que los distintos 'exams board' (Oficinas evaluadores, que actualmente son 8 en todo el país) se ajusten a los estándares de contenido (currículum), estándares de evaluación (reglas sobre cómo se evalúa que establece Ofqual para resguardar la comparabilidad) y estándares de desempeño (grade standard). Para cada prueba se consideran 9 estándares de desempeño.

2.3 Historia

El año 2015 se incorporó el nivel de desempeño 9, nivel más alto, por primera vez en los exámenes GCSE[5]. Adicionalmente de denominar los niveles de desempeño a través de letras, pasan a ser nombrados del 1 al 9. (Para más información ver 'Reforms to GCSEs in England from 2015 Summary')

Las razones entregadas para estos cambios son las siguientes:

(1) Razón métrica: Provee mayor diferenciación entre los estudiantes que se ubican por sobre el nivel medio de logro, y además permite seguir midiendo las diferencias y el progreso entre los estudiantes que obtienen más bajos resultados.

[5] Para ver más información sobre la reforma de GCSE en 2015 ver: https://www.gov.uk/government/collections/reform-of-gcse-qualifications-by-ofqual#documents y documento online 'Reforms to GCSEs in England from 2015 Summary'.

(2) Cambio curricular mayor: Generar un nuevo sistema de calificación de los GCSE entrega una señal clara de que la prueba ha cambiado, y permite disminuir la errónea comparación entre los resultados obtenidos entre el anterior y el actual GCSE.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales a cargo

En educación primaria la STA elabora los estándares de desempeño. En educación secundaria Ofqual define los estándares y vela por su mantención entre los distintos 'exam boards'.

3.2 Metodología

En el caso de Year 2 y la prueba SAT que se administra en ese año, el estándar de desempeño está definido por un puntaje en la escala igual a 100. En 2016 se usó la metodología de paneles de profesores que definieron ese puntaje de la escala (100 puntos) para establecer cuándo el estudiante cumplía con lo exigido por el estándar en la prueba de KS1 y 2[6]. No existe una definición cualitativa del estándar asociado a cada prueba, sino más bien se refiere al cumplimiento de las metas definidas para cada ciclo de la educación primaria.

En el caso de educación secundaria y las pruebas administradas por cada 'exam board' que escoja la escuela, Office of Qualification and Examinations Regulations (Ofqual), tal como se mencionó anteriormente, está encargada de definir los estándares y velar por su mantención entre los distintos 'exam boards'. El puntaje para ubicar a estudiante en un nivel de desempeño dentro de los nueve posibles (grade standards) es definido a través de un grupo de examinadores expertos de cada 'exam board' que determinan un puntaje mínimo, medio y máximo necesarios para alcanzar los niveles de desempeños más altos, medios y más bajos. Luego, con eso y considerando criterios aritméticos, se construyen los intervalos, teniendo en cuenta que cada intervalo contenga la misma cantidad de sub-intervalos[7]. Ofqual es el encargado de velar porque cada uno de las 'exam boards' que toma las pruebas en las escuelas no tenga diferencia significativa entre sus propias pruebas de años pasados, ni entre las pruebas implementadas por otros 'exam boards'.

[6] Para más detalles sobre KS1 ver: https://www.gov.uk/guidance/key-stage-1-tests-standard-setting and https://www.gov.uk/guidance/scaled-scores-at-key-stage-1
Para ver más detalles sobre KS2, especialmente características de la escala, ver: https://www.gov.uk/guidance/scaled-scores-at-key-stage-2

[7] Para más información sobre la definición de niveles de desempeño ver: http://webarchive.nationalarchives.gov.uk/20141110161323/http://comment.ofqual.gov.uk/setting-the-grade-standards-of-new-gcses-april-2014/ y https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/529862/Setting_grade_standards_part_2.pdf

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

En primer ciclo de educación primaria (KS1), el resultado según estándares de desempeño en pruebas SAT se combina con una evaluación hecha por el docente en base a su trabajo en clases y es entregada al final del año por el docente a los padres. El docente transmite su evaluación a los apoderados a través de un informe elaborado por él o ella mismo y se añade el informe con los resultados de la prueba SAT, los que no consideran ninguna información adicional a los puntajes, además de la definición si ha cumplido el estándar esperado. La evaluación docente es más amplia y puede no coincidir con los resultados SAT puesto que considera más y diversa evidencia sobre el desempeño de los estudiantes.

En segundo ciclo de educación primaria (KS2) se informa si el estudiante cumplió el estándar y no se vincula con la evaluación hecha por el docente. El docente entrega a los apoderados el informe con los resultados de la prueba SAT, los que no consideran ninguna información adicional a los puntajes además de la definición si ha cumplido el estándar esperado.

Adicionalmente, Office for Standards in Education, Children's Services and Skills (OFSTED) desarrolla un informe con altas consecuencias para las escuelas[8]. En base a una amplia fuente de información, dentro de la que se considera el logro de los estudiantes de las escuelas en las evaluaciones SAT y GCSE, esta oficina cataloga a los establecimientos en una de cuatro categorías: 'Outstanding', 'Good', 'Requires Improvement' y 'Inadequate', lo que puede llevar a emprender cambios en el trabajo de las escuelas hasta su cierre.

Si bien los resultados de las evaluaciones a estudiantes, que informan el dominio de los estudiantes en las distintas áreas curriculares evaluadas, permanecen como información privada para la escuela y los padres de los alumnos, esta información complementa información agregada, en base a la cual las escuelas son evaluadas y clasificadas en categorías de efectividad escolar que son diseminadas públicamente y que, entre otros fines, busca proveer antecedentes útiles para que los padres escojan la escuela a la que asistirán sus hijos, dentro de la zona geográfica en la que viven.

5. USO ESTÁNDARES DE DESEMPEÑO

Esta evaluación puede ser asociada a distintos usos tales como:

1. Rendición de cuentas de las escuelas acerca de sus resultados.
2. Selección de estudiantes para entrar a la educación superior.
3. Diagnóstico del logro del estándar nacional en las áreas evaluadas.

[8] Para más detalles generales sobre OFSTED ver: https://www.gov.uk/government/organisations/ofsted/about y https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/547229/School_inspection_handbook-section_5.pdf

4. Provisión de información al público en general sobre el rendimiento alcanzado por las escuelas

6. REFERENCIAS

Ofqual (2016). Setting the grade standards of new GCSEs in England – part 2. Recuperado en Julio de 2016 de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/529862/Setting_grade_standards_part_2.pdf

Ofqual. Reforms to GCSEs in England from 2015 Summary. Recuperado en Julio de 2016 de: https://www.gov.uk/government/collections/reform-of-gcse-qualifications-by-ofqual#documents

Ofsted. Sitio oficial https://www.gov.uk/government/organisations/ofsted/about

Ofsted (2005). School inspection handbook Handbook for inspecting schools in England under section 5 of the Education Act 2005. Recuperado en Agosto de 2016 de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/547229/School_inspection_handbook-section_5.pdf

Standards and Testing Agency (2016). Annual Report and Accounts For the year ended 31 March 2016. Recuperado en Agosto de 2016 de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/538564/STA_Annual_report_and_accounts.PDF

Standards and Testing Agency (2016). Business plan 1 April 2015 - 31 March 2016. Recuperado en Agosto de 2016 de: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/477453/STA_Business_Plan_2015-16.pdf

7. CONTACTO

Liz Twist
Interim Deputy Director for Test Development
Standards & TestingAgency (STA)
liz.twist@education.gsi.gov.uk

Liz Twist fue contactada por email.

FICHA #4
Holanda: "School Leavers' Attaintment test" o "Cito Test"

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la Evaluación

Holanda cuenta con un sistema de evaluación altamente desarrollado a lo largo de la trayectoria escolar[9] que abarca diversos propósitos: evaluar y monitorear el sistema educacional, las escuelas, el desempeño docente y el rendimiento de los estudiantes; apoyar la enseñanza; y certificar las competencias de los estudiantes para el mundo del trabajo.

Con el propósito de identificar y reportar tendencias en los logros de los estudiantes y proporcionar información al proceso de toma de decisiones de política educativa, docentes y público en general, se desarrollan en Holanda la **Encuesta Periódica de Educación (Periodical Survey of Education - PPON)**[10] y **Encuesta Anual de Niveles Educacionales (Annual Survey of Educational Levels – JPON)**[11]. Los resultados de dichas pruebas no están asociadas directamente a consecuencias para las escuelas ni para los estudiantes.

Junto con estas pruebas, también existen las pruebas **"Cito LVS"** que conforman un sistema de evaluación cuyo propósito es apoyar la enseñanza[12]. Adicionalmente, existen

[9] La educación en Holanda es obligatoria para todos los niños desde los 5 años hasta los 16 años y está dividida en tres niveles: educación básica, secundaria y terciaria. La educación básica, dividida en ocho grados, considera estudiantes que tienen entre 4 y 12 años de edad. La educación secundaria considera estudiantes que tienen entre 12 y 18 años y está divida en distintos planes de estudio: (1) 'pre-vocational education' (VMBO), 12 a 16 años; (2) 'individualised pre-vocational education' (IVBO), 12 a 16 años; (2) 'senior general secondary education' (HAVO), 12 a 17 años; (3) 'pre-university education' (VWO), 12 a 18 años.

[10] PPON es una encuesta muestral aplicada a estudiantes de grado 8 (11-12 años), y en algunos casos, estudiantes de grado 4 (7-8 años). Ha sido desarrollada periódicamente desde 1987 y evalúa habilidades en matemáticas y holandés, ciencias sociales, historia, geografía, biología, inglés, artes visuales, música y educación física. Está diseñada para proveer información robusta sobre los cambios que se dan en el tiempo y en las diferentes áreas del curriculum. Esta prueba monitorea matemáticas y holandés durante un ciclo de 5 años.

[11] JPON es una encuesta muestral introducida en 2008 para monitorear el progreso del programa Escuelas de Mañana (School for tomorrow) del Ministerio de Educación, Cultura y Ciencia. Evalúa el manejo de los estudiantes en holandés y matemática en dos puntos del ciclo educacional (grado 4 y grado 8). Esta prueba está diseñada para proveer retroalimentación oportuna en áreas específicas que buscan ser trabajadas por la reforma llevada adelante por el Ministerio de Educación, Cultura y Ciencias.

[12] También existen varias evaluaciones desarrolladas por Cito con propósitos formativos y de apoyo a la enseñanza, llamadas pruebas "Cito LVS", que consisten en un set pruebas nacionales estandarizadas para desarrollar evaluaciones longitudinales del desempeño de los estudiantes a lo largo de la educación primaria, de los estudiantes de entre el grado 5to y 7mo ("Cito entrance") y de los primeros dos años de la educación secundaria

certificaciones de competencias para estudiantes al finalizar su enseñanza secundaria, de cara a su acceso al mercado laboral[13].

Finalmente, existe el "**school leavers' attaintment test**" o "Cito test", que es un examen aplicado a todos los estudiantes al concluir la enseñanza básica. Tiene el doble propósito de: (1) evaluar los conocimientos adquiridos por ellos durante los primeros ocho años de la educación primaria; y (2) de asesorar a los estudiantes en la selección del plan de estudios secundarios más idóneo para ellos. Desde 2012 esta prueba comienza a ser obligatoria para todas las escuelas del país y desde el 2013 todos los estudiantes del grado 8 deben rendir esta evaluación en Lenguaje y Matemáticas.

La ficha correspondiente a Holanda estará centrada en el "School Leavers' Attaintment Test", por ser la prueba que se utiliza masivamente por las escuelas y por estar asociada a consecuencias a nivel de estudiantes y de escuelas. Este examen ha sufrido modificaciones en los últimos tres años. Con certeza podemos afirmar que hasta el 2013 el Instituto Central para el Desarrollo de Pruebas (Central Institute for Test Development – CITO), responsable de los proyectos de evaluación mencionados con anterioridad, era el principal proveedor del "School Leavers' Attaintment Test" (Cito test).

Debido a la falta de información actualizada no es posible describir con certeza uno de los cambios que parece haber sido implementado desde 2015, momento en que CITO comenzaría a ser la única proveedora de este test al final de la enseñanza básica. De todos modos, cabe destacar que el "Cito test", versión CITO del "school leavers attaintment test", ha sido la prueba más utilizada voluntariamente por las escuelas[14]. Hasta 2013 las escuelas asumían el costo monetario del test. No hay información actualizada disponible acerca de quién asume el costo actual del "Cito test", aunque diversos autores destacan lo naturalizado dentro del sistema holandés de este mecanismo de costo a las escuelas, único en el mundo.

1.2 Referente orientador de las evaluaciones.

El referente orientador de las pruebas fue estipulado por el Decreto de Educación Primaria de 2010 (Act on Primary Education 2010), el Decreto de Educación secundaria (Act on Secondary Education) y especialmente el Decreto de Lenguaje y Matemática de agosto del 2010 (Language and Numeracy Act). En este último decreto se establecen niveles de referencia, "reference levels", o puntos de corte "Attaintment benchmarks" para Lenguaje y Matemáticas, los que son estándares de aprendizaje que describen lo que los estudiantes deberían haber aprendido y practicado en los distintos ciclos

[13] Existe una certificación de competencias para los estudiantes para tres tipos de educación secundaria, una para VMBO, una para HAVO y una para VWO. Estas evaluaciones reconocen competencias de entrada al mercado laboral, así como permiten el acceso a educación superior.

14 En Holanda las escuelas tienen la facultad de elegir que examenes administran, de entre distintos proveedores. Cito es el proveedor principal (con 85% escuelas), correspondiente a 144.708 estudiantes.

educativos[15]. "Cito test" tiene el propósito de evaluar en qué medida los estudiantes demuestran la expectativa esperada y descrita por los niveles de referencia (Scheerens et al. 2012).

En el caso de Lenguaje, existen cuatro niveles básicos o fundamentales (F) y cuatro niveles avanzados o 'Targets' (S) para educación primaria, secundaria y vocacional. En el caso de Matemáticas, existen tres niveles básicos y tres niveles avanzados. Los niveles básicos (F) se espera que lo cumplan la mayor cantidad de estudiantes, mientras que los niveles avanzados (S) están enfocados para estudiantes con mayores niveles de aprendizaje. Los niveles más avanzados incluyen a los más básicos. En la actualidad, para la educación básica, el estándar esperado de Lenguaje es 1F y el avanzado es 1S[16].

Cada año el 'examination syllabus' especifica los elementos que finalmente serán evaluados por la prueba nacional al final de educación básica. Este documento es aprobado por el Ministerio de Educación, Cultura y Ciencia.

1.3 Organismo responsable del programa de evaluación y del currículo.

La responsabilidad de llevar adelante estas pruebas en Holanda recae en el Ministerio de Educación, Cultura y Ciencia. También le cabe un rol a la Inspectoría de Educación (Inspectorate of Education), institución a cargo del monitoreo sobre la calidad del sistema de evaluación, donde por encargo constitucional, debe elaborar anualmente un estado de situación de la educación holandesa. Ambas instituciones cuentan con una serie de instituciones y organismos colaboradores para realizar sus labores, entre ellas destaca CITO.

1.4 Áreas disciplinares

El "Cito test" cubre las materias de lenguaje, aritmética/matemáticas y habilidades de estudio. Opcionalmente también evalúa temas del área ambiental.

[15] El documento con los niveles de referencia de ambas áreas puede ser encontrado en el siguiente link. Lamentablemente, solo existe una versión en holandés. Ver:
http://www.taalenrekenen.nl/downloads/referentiekader-taal-en-rekenen-referentieniveaus.pdf/

[16] El siguiente es un ejemplo de la complejidad de los estándares básico y avanzado en relación a escritura al finalizar la enseñanza básica. Se trata de una traducción muy preliminar del texto en holandés basada en el traductor Google. Mayor precisión se alcanzaría con una traducción oficial desde el documento antes citado. (1F) Students can write short, simple texts on everyday topics in the form of a letter, card or email. Can use the most common punctuation marks; (1S) Students can write coherent texts with a simple, linear structure, across diverse and familiar topics. The text contains a sequence of introduction, core and lock.

1.5 Grados evaluados

"Cito Test" se aplica a todos 'school leavers', es decir, los estudiantes del último año del grado 8.

1.6 Características de las pruebas

El "Cito test" es una prueba diseñada por la oficina Cito y administrada por las escuelas. Es un instrumento individual de selección múltiple, aplicado hasta el 2012 durante el mes de febrero y diseñado para medir habilidades más que contenidos en las áreas evaluadas. Cada año se informa a través del "examination syllabus" el número y largo de las pruebas. En general está conformada por alrededor de 200 preguntas de selección múltiple en los tres dominios mínimos, Matemáticas, Lenguaje y habilidades de estudio y por otra sección optativa de 90 preguntas evalúa temas ambientales.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

En 2010, el Ministerio de Educación Cultura y Ciencias, a través del decreto de lenguaje y matemáticas (Language and Numeracy Act), establece niveles de referencia con puntos de corte (benchmarks) acerca del conocimiento y habilidades que los estudiantes deben tener, tanto para lenguaje como para matemáticas tanto al final de la educación primaria y para cada uno de las trayectorias (tracks) de la educación secundaria (indicando niveles fundamentales o básicos y avanzados o 'targets').

2.2 Características de los estándares

CITO test entrega resultados asociados a bandas de puntajes que refieren a nueve distintos tipos de educación secundaria, donde cada banda se asocia a puntajes de corte en la escala de la prueba. CITO informa el resultado del estudiante en el test a través de un puntaje absoluto que, asociado a las bandas de puntaje definidas por la prueba, permite indicar el tipo de recomendación entregada a escuelas, padres y estudiantes. La siguiente tabla fue utilizada por CITO en la evaluación de 2012 y asocia un tipo específico de recomendación a los rangos de puntajes en ese test.

TABLE FOR THE USE Of Cito BAND-WIDTHS			
Advice	Additional research required	Consultation with primary school required	automatically accepted
Practical education	No Cito bandwidths applicable		
Learning routes supporting education	No Cito bandwidths applicable		
vmbo-basic	514 and lower	515 t/m 520	521 and higher
vmbo-basic/cadre	517 and lower	518 t/m 522	523 and higher
vmbo-cadre	520 and lower	521 t/m 528	529 and higher
vmbo-mixed	526 and lower	527 t/m 533	534 and higher
vmbo-theoretical	526 and lower	527 t/m 533	534 and higher
vmbo-theoretical / havo	528 and lower	529 t/m 535	536 and higher
havo	531 and lower	532 t/m 537	538 and higher
havo / vwo	535 and lower	536 t/m 540	541 and higher
vwo	539 and lower	540 t/m 544	545 and higher
kopklas	No Cito bandwidths applicable		

Nota: La primera columna indica distintos tipos de educación secundaria.

2.3 Historia

En términos generales, durante los últimos cinco años, el sistema holandés de educación ha profundizado la implementación de un sistema educativo que define estándares nacionales monitoreados a través de evaluaciones externas asociadas a cada vez más fuertes consecuencias para estudiantes y escuelas.

Esta profundización de la rendición de cuentas a través de pruebas estandarizadas de desempeño académico se ancla a una estructura ya existente en el sistema, fundada en las diversas evaluaciones que voluntariamente las escuelas utilizaban para monitorear el desempeño de sus estudiantes, pero avanza a la definición de estándares y la obligatoriedad de la evaluación a final de los ciclos educativos. Este camino comienza el año 2010 con una serie de decretos que reforman la educación que gradualmente hacen los estándares mínimos de desempeño académico de los estudiantes indicadores cada vez más centrales de la calidad de la educación sistema educativo holandés.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

El desarrollo de las evaluaciones de estudiantes y el reporte de los estándares son responsabilidad de CITO (Central Institute for Test Development).

3.2 Metodología

No ha sido posible acceder a documentos que detallen con más especificidad la metodología utilizada para establecer las bandas de puntajes ni tampoco para conocer con precisión la vinculación de esos puntajes y el logro de los estándares básico (1F) y avanzado (1S).

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

Los resultados de los estudiantes en el "Cito Test" son informados al alumno y su familia a través de un informe de la escuela, emanado desde CITO. Este informe indica el resultado en el test a través de un puntaje absoluto, el cual indica su desempeño en la prueba y, en base a una escala elaborada por Cito, qué tipo de colegio es el más adecuado para que el estudiante continúe con sus estudios secundarios. Esta decisión puede tomarse, tal como lo sugiere CITO, considerando únicamente este reporte o alguna otra información adicional con que cuente el establecimiento educacional.

5. USO ESTÁNDARES DE DESEMPEÑO

Según todas las fuentes consultadas, el principal uso de los resultados del "Cito test" es sugerir el tipo de establecimiento educacional para que el estudiante continúe su educación secundaria.

La información que entrega el "Cito test" (corregido por nivel socioeconómico de la escuela) así como las evaluaciones durante la educación secundaria son los primeros indicadores para realizar un análisis de riesgos tempranos dentro del sistema educacional holandés, ayudando a identificar las escuelas de bajo rendimiento. Una vez detectado el riesgo, se realiza una inspección adicional para evaluar la calidad de la escuela.

En el caso que las escuelas no mejoren - no muestren una mejora mínima en el desempeño de los estudiantes o cuando no cumplan con lo especificado en los decretos emanados desde el Ministerio de Educación y Ciencia para la educación primaria y secundaria - el Departamento de Educación pueden ejecutar sanciones administrativas o financieras.

Finalmente, el "Cito test" puede ser utilizado a nivel de escuela para orientar la discusión sobre los objetivos educativos y curriculares, y a su vez, a nivel de sistema de educación, utilizado dentro de los reportes que entrega la Inspectoría de Educación, y los informes de "Key Figures and Trends" que elabora el Ministerio de Educación, Cultura y Ciencia.

6. REFERENCIAS

OECD Reviews of Evaluation and Assessment in Education. Netherland. Deborah Nusche, Henry Braun, Gábor Halász and Paulo Santiago.

EDUCATIONAL EVALUATION AND ASSESSMENT IN THE NETHERLANDS. Country background report for the OECD study on Evaluation and Assessment Frameworks for Improving School Outcomes. Jaap Scheerens, Melanie Ehren, Peter Sleegers and Renske de Leeuw. University of Twente, the Netherlands.

https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Netherlands:Quality_Assurance

7. CONTACTO

No fue posible contactar a un especialista de CITO para la elaboración de esta ficha.

FICHA #5
ONTARIO PROVINCIAL ASSESSMENT PROGRAM
Programa Provincial de Evaluación de Ontario, Canadá

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la 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:

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

El propósito básico 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 currículo provincial. Las evaluaciones se desarrollan en forma paralela en los dos idiomas oficiales de Ontario: inglés y francés. Se procura tener los mismos criterios de calidad en estas pruebas paralelas.

1.2 Referente orientador de las evaluaciones.

Todas las evaluaciones deben ser referidas a las expectativas del currículum de Ontario. El currículum establece expectativas específicas para cada área disciplinar y grado desde preescolar (kindergarten) hasta final de la secundaria.

1.3 Organismo responsable del programa de evaluación y del currículo.

Un organismo ministerial está a cargo del currículo, y otro organismo semiautónomo está a cargo de la evaluación. La División de Aprendizaje y Currículum (*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 Currículum de Ontario).

El programa de evaluación está a cargo de una agencia *'arm's length'* (semiautónomo) 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 Accountabilty 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 currículo de Ontario en el sistema educativo de la provincia.

1.4 Áreas disciplinares

El programa de evaluación de Ontario evalúa las áreas disciplinares de lectura, escritura, matemáticas y *literacy*[17], en los ámbitos que sus técnicos estiman son evaluables mediante pruebas estandarizadas. Todos los años, los diseños de las pruebas (*frameworks*) por área disciplinar son publicados para el uso de docentes y estudiantes. La EQAO estima que la única preparación necesaria para sus pruebas es la enseñanza y el aprendizaje del currículo de Ontario.

1.5 Grados evaluados

Como se indica en la Tabla 1, las primeras evaluaciones del EQAO se administran en 3º grado, son las pruebas de *Primary Division* y cubren el currículo de lectura, escritura y matemáticas de 1º, 2º y 3º grado. Pruebas en estas tres áreas toman lugar de nuevo en 6º grado y cubren las expectativas curriculares de *Junior Division* (4º a 6º grado). En 9º se administra una prueba de matemáticas, con distintas versiones según los estudiantes estén matriculados en cursos aplicados o académicos de matemáticas. Finalmente, en 10º grado se administra el Ontario Secondary School Literacy Test (OSSLT) (Evaluación de "literacy" a nivel secundario de Ontario, que es un requisito para obtener el diploma de estudios secundarios de Ontario[18].

Tabla 1: Grados y áreas evaluadas en el programa de evaluación de la provincia de Ontario

3º Grado	Lectura, escritura y matemáticas evaluadas en el último de los tres grados de *primary division*
6º Grado	Lectura, escritura y matemáticas evaluadas en el último de los tres grados *junior division*
9º Grado	Matemáticas, evaluadas en el primer año de educación secundaria
10º Grado	*Literacy*, evaluado como requisito de graduación.

[17] "Literacy" en este caso se refiere a proficiencias de lectura y escritura en el contexto de cada una de las asignaturas hasta final de 9º grado.

[18] Esta última evaluación, en una prueba general voluntaria de la versión en-línea, fue objeto de un ciberataque en octubre de 2016, presentando un importante desafío para el programa del EQAO de convertir todo el programa de evaluación a versiones en-línea.

1.6 Características de las pruebas

Educadores de la provincia participan en todos los aspectos de las pruebas: construcción, administración y logística, corrección, diseño de informes, entre otros, con el propósito expreso de ayudar a garantizar la relevancia de las evaluaciones y su alineamiento con el currículum y las prácticas docentes.

En las evaluaciones de 3º y 6º hay tres cuadernillos de pruebas: dos para lenguaje (lectura y escritura) y uno para matemática. Cada cuadernillo toma dos horas, para un total de seis horas[19]. Los cuadernillos cubren tres grados (correspondiente al nivel o "división") de expectativas curriculares. En la evaluación de matemática de 9º grado, hay dos cuadernillos que toman una hora cada uno, para un total de 2 horas. Hay dos cuadernillos de 75 minutos en el OSSLT (*Ontario Secondary School Literacy Test*), para un total de dos horas y media.

Todos los cuadernillos contienen ítemes comunes que cuentan para sus resultados y un pequeño número de ítemes piloto e ítemes "matriz" que se usan para la equiparación interanual de resultados, que no se toman en cuenta para los resultados del estudiante. En todas las pruebas hay tanto ítemes de selección múltiple e ítemes de respuesta abierta.

Los ítemes de selección múltiple se corrigen por "scanner" con la excepción de las evaluaciones de 3º grado. Los ítemes de respuesta abierta son corregidos por parte de examinadores, que en su mayoría son educadores de la provincia. Una pauta de corrección general para cada tipo de ítem describe el tipo de trabajo que corresponde a cada nivel de desempeño y sirve como herramienta para asegurar consistencia entre los ítemes de una misma prueba, y consistencia interanual. En base a la pauta de corrección general, que es una guía para describir tipos de ítemes, se escriben pautas concretas para cada ítem específico. Cada nivel de desempeño en cada ítem cuenta con un *anchor paper* (ejemplo de anclaje) que es una muestra de una respuesta autentica, tomada de los pilotos del ítem, que ilustra ese nivel de desempeño para ese ítem.

Estos materiales para la corrección de ítemes de respuesta abierta, se identifican y desarrollan durante el proceso de determinación del rango de desempeños (*range finding*). Este proceso es supervisado por expertos en psicometría y currículum del EQAO con participación de 8 a 25 educadores de la provincia que se reúnen tres veces al año para estudiar muestras de respuestas de estudiantes y recomendar los ejemplos de anclaje. Estos comités también revisan y hacen recomendaciones acerca de todos los materiales para el entrenamiento de evaluadores, incluyendo la prueba de certificación

[19] Se ha reducido a la mitad el tiempo de evaluación. En el pasado tomaba un total de 12 horas. Esta reducción a la mitad del tiempo con respecto a evaluaciones anteriores, se cumple en todas las pruebas actuales.

de evaluadores, además de recomendar materiales para actividades de calibración de pruebas.

Los evaluadores reciben entrenamiento exhaustivo para asegurar una interpretación común de los materiales para la corrección de ítemes, y cada evaluador, supervisor de evaluación y líder de equipo de evaluación debe aprobar una evaluación escrita que lo califica como corrector de ítemes. Los evaluadores son responsables de corregir entre uno a cuatro ítemes máximo, para asegurar mayor consistencia y menos errores. Los líderes de equipos de evaluadores y los supervisores, monitorean la validez y confiabilidad de cada evaluador todos los días para fines de monitoreo y control de calidad. Cuando estiman necesario, organizan entrenamientos individuales y grupales adicionales. En el caso de la prueba OSSLT, que es un requisito de graduación, cada ítem es evaluado por dos evaluadores independientes[20].

En la determinación del puntaje de los estudiantes se utiliza un proceso de equiparación de pruebas basados en la Teoría de Respuesta al Ítem (TRI) para garantizar la equivalencia de puntajes de año en año. También a veces se hacen revisiones al currículo de Ontario que resultan en modificaciones del contenido – y por tanto de las características psicométricas – de las pruebas que también requieren del uso de TRI en un proceso de escalamiento, para garantizar equivalencias en los puntajes. Los procesos de equiparación y escalamiento son similares pero sus propósitos son distintos. La equiparación se usa para ajustar diferencias en dificultad entre pruebas similares en contenido y especificaciones estadísticas. El escalamiento se usa para ajustar diferencias entre pruebas que no son similares en contenido y especificaciones estadísticas. Cuando no ha habido ajustes en el currículum, no se emplea escalamiento. Los modelos generales que se usan en el programa de evaluación son el Modelo Logístico de 3 Parámetros (3PL) y el modelo Generalizado de Crédito Parcial (GPC).

[20] El enorme volumen de trabajo escrito que se recoge de todos los estudiantes de la provincia, en ocasiones contiene textos que sugieren situaciones en que los estudiantes pueden estar sufriendo abuso o experimentando situaciones peligrosas. También ha habido contenido inapropiado u ofensivo e inclusive que indican posibles situaciones de criminalidad. La ley exige que EQAO cuente con un protocolo detallado para tratar estos temas que incluyen situaciones que requieren referir casos al *Children's Aid Society* (entidad protectora de los derechos de menores), la dirección de la escuela, la familia y otras agencias responsables.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

La EQAO es una agencia formada en 1996 en respuesta a las recomendaciones de 1994 de la Comisión Real sobre el Aprendizaje (*Royal Commission on Learning*) que concluyó que la provincia necesitaba una agencia para el escrutinio público, pero independiente, del sistema educativo y sus logros. La EQAO es administrada por un Director Ejecutivo (*Chief Executive Officer*) y una Junta de Directores (*Board of Directors*). Las responsabilidades de la EQAO se especifican en un Memorándum de Entendimiento entre su Junta de Directores y el Ministerio de Educación, que debe ser ratificado por el Gabinete del Primer Ministro Provincial y la Junta Administrativa del Gabinete (*Management Board of the Cabinet*) del gobierno de Ontario. El actual Memorándum está vigente desde el año 2003. La Junta de Directores está compuesta por líderes educativos tanto del sector público como del privado, líderes de negocios, abogados, líderes comunitarios y de universidades. En la actualidad, el Director (*Chair*) de la Junta de Directores es un exministro de educación de Ontario y el Director Ejecutivo es un educador con experiencia en administración educativa, currículum y evaluación.

2.2 Características

El programa de evaluación de Ontario fija en forma separada, los puntos de corte correspondientes a los niveles de desempeño en las evaluaciones en inglés y francés. En las pruebas de 3º y 6º y 9º se fijan cuatro niveles (ver Tabla 2). En las pruebas de 10º grado, requisito de graduación del secundario, son dos niveles de desempeño: completo e incompleto. En el caso de estas últimas pruebas, el nivel de desempeño "completo" es el estándar provincial para aprobar ese requisito para la certificación de término de la educación secundaria.

Tabla 2: Descripción de Estándares de Desempeño en Pruebas de Ontario de 3º, 6º y 9º.

Nivel	Descripción del Estándar
4	El estudiante demuestra dominio de todos los conocimientos y desempeños del currículum. Supera el estándar provincial.
3	El estudiante demuestra dominio de la mayoría de las expectativas del currículum. Logra el estándar provincial.
2	El estudiante demuestra dominio de algunas de las expectativas del currículum. Se aproxima al estándar provincial.
1	El estudiante demuestra dominio limitado de algunas de las expectativas del currículum. Muy por debajo del estándar provincial.
(sub-1)	El estudiante no muestra suficiente evidencia del dominio de las expectativas del currículum para ser categorizado al Nivel 1.

2.3 Historia

El EQAO introdujo sus pruebas, y los estándares de desempeño asociados a los mismos, en los siguientes años:

- 1996-97: Estándares y pruebas de lectura, escritura y matemática en 3º Grado
- 1998-99: Estándares y pruebas de lectura, escritura y matemática en 6º Grado
- 2000-01: Estándares y pruebas de matemáticas de 9º Grado
- 2002: Estándares y pruebas de *literacy* en 10º grado

Hasta 2002, los únicos cambios en los estándares de desempeño fueron ajustes recomendados por los contratistas externos, para corregir problemas menores de calibración de año a año – principalmente problemas relativos al comportamiento de ítemes a lo largo del tiempo. La metodología para las correcciones fue la misma que el establecimiento original de los primeros estándares de desempeño (explicado en el apartado 3.2).

En 2002 el EQAO contrató una revisión externa de todos los procesos logísticos, psicométricos, comunicacionales y en otros ámbitos con el fin de asegurar continuo control de calidad. Esta revisión no se hizo según un calendario establecido de antemano, sino por decisión de la junta directiva que estimó necesaria una auditoria externa comprensiva. La revisión tuvo el propósito de determinar cuan bien el sistema de evaluación ejemplifica estándares internacionales de calidad en evaluaciones estandarizadas a gran escala y que los productos del EQAO satisfagan las necesidades de los usuarios en los ámbitos de *accountability*, planificación de mejoras, y formación del cuerpo docente. La revisión estuvo a cargo del *Ontario Institute for Studies in Education* (OISE – Instituto de Ontario para Estudios en Educación) de la Universidad de Toronto, la universidad pública de mayor prestigio en la Provincia. Personal del OISE realizó la revisión con la participación de un conjunto de académicos y especialistas de instituciones especializadas en todo el mundo. También hubo extensas consultas con usuarios del sistema educativo y el público general.

En el ámbito de estándares de desempeño, las recomendaciones fueron principalmente en lo que se refiere a la comunicación de las mismas a distintas audiencias. En términos de procedimientos, en la revisión del 2002 se recomendó el uso de ítemes de respuesta abierta, y se recomendó el procedimiento de calibración pare ese tipo de ítemes para su uso en la definición de estándares de desempeño que se usa en la actualidad, y que se describe en el apartado 3.2.

Finalmente, los revisores recomendaron establecer un calendario de revisión sistemática de todos los procesos y productos del EQAO, incluyendo los estándares de desempeño y los procedimientos para establecerlos cada siete años. Los procedimientos descritos en esta ficha son aquellos que se siguen después de la revisión de 2009.

Las revisiones periódicas se llaman Revisiones de Aseguramiento de Calidad de Pruebas (Ensuring Quality Assessment reviews) incluyen cuatro etapas:

1. Investigación y Estudio: El EQAO reliza una revisión de "mejores prácticas" mediante el estudio de los procesos y procedimientos de organizaciones y agencias de evaluación de prestigio en el ámbito internacional. Al mismo tiempo, un equipo de evaluación externo de reconocidos expertos internacionales en evaluación a gran-escala es organizado y liderado por la universidad pública provincial (Ontario Institute for Studies in Education / University of Toronto) conduce una auditoría externa de todos los procedimientos del EQAO. El informe de la evaluación externa incluye una auditoría exhaustiva de todos los procesos, procedimientos y protocolos del EQAO y proporciona recomendaciones para mejorar los mismos.

2. Consulta: El EQAO realiza foros conjuntos entre su consejo asesor (Assessment Advisory Committee) y más de 20 grupos organizados de comunidades de interes que representan a directores de centros educativos públicos y privados, agentes de supervisión, docentes, consejos educativos ("boards of education and trustees"), padres de familia y estudiantes para recoger información acerca de las pruebas y su administración en las escuelas, la pertinencia de los datos para propósitos de rendición de cuentas y mejoría educativa que proporciona el EQAO en sus distintos informes y el impacto de las pruebas en la formación en servicio de docentes. Al mismo tiempo, el EQAO invita a comentario público abierto al informe de auditoría externa que es producto de la primera etapa descrita arriba.

3. Analisis y síntesis: El EQAO considera toda la información recogida en las auditorías, revisiones y consulta pública. Evalúa en particular las recomendaciones de la evaluación externa a la luz de los resultados a consultas de comunidades de interés, mejores practicas internacionales, impacto potencial en escuelas y entidades educativa, impacto potencial en la calidad psicométrica de las pruebas, e impacto en el EQAO mismo en términos de recursos, costos y tiempos.

4. Implementación: Como resultado de la etapa 3, el EQAO asume compromisos de acción para el siguientes periodos. En evaluaciones externas y consultas públicas futuras, se evalúa el buen curso de la implemención de estos compromisos.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

El EQAO lleva a cabo todo el trabajo psicométrico de calibración en relación con los estándares de desempeño. Como un control de calidad, contrata compañías independientes para replicar todos sus cálculos, a fin de identificar, y resolver, posibles problemas. EL EQAO también coordina el trabajo de comités de educadores y otros especialistas de la provincia, en el proceso de establecer los estándares.
Coordina este trabajo en forma separada para las pruebas en inglés y francés.

3.2 Metodología

En las evaluaciones de 3º. 6º y 9º, las pruebas incluyen ítemes de selección múltiple y de respuesta abierta. Los niveles se asignan usando el θ calculado en la calibración (se usan 3PL -*Three Parameter Logistic* - para ítemes de selección múltiple y GPC - *Generalized Partial Credit*- para ítemes de respuesta abierta). 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.

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

El procedimiento para las pruebas OSSLT, la prueba de *literacy* que es requisito de graduación, es más cualitativo. De nuevo, como es el caso de todas las pruebas, los estándares de desempeño se fijan en forma separada para las pruebas en inglés y francés.

El proceso, organizado por EQAO, utiliza paneles de jueces que representan miembros del público y personas empleadas en sistemas escolares de primaria y secundaria. Estas personas se identifican en colaboración con líderes educativos y comunitarios de la provincia, y expresamente no se busca representar ninguna organización o grupo de intereses (por ejemplo, la política es no tener representación designada de sindicatos, partidos políticos, asociaciones profesionales, etc.). Cada panel de jueces cuenta con 18 miembros: 9 educadores y 9 representantes del público. Los paneles cuentan con asesores, expertos en psicometría y currículum, a menudo de universidades de Ontario. Inclusive han asesorado expertos del OCDE.

Los paneles estudian muestras de trabajo estudiantil en los cuadernillos de la prueba correspondientes a pilotajes y pruebas de campo. Las muestras de cuadernillos se escogen para representar la diversidad geográfica, socio-económica y étnica de la provincia. Basados en estas muestras, se discute procurando decidir que ejemplos corresponden a estudiantes que están a nivel de completar los requisitos de graduación (nivel completo) o por debajo de ese nivel (incompleto). Se discuten ampliamente las descripciones del trabajo correspondiente a cada nivel de desempeño, e inclusive en algunos casos se resuelven desacuerdos mediante votos. El referente es el currículum provincial de Ontario y la interpretación del mismo que hacen los miembros de los paneles. La Tabla 3 muestra un ejemplo de las descripciones de cada nivel de desempeño para lectura en inglés. Existen además estándares de desempeño en lectura en francés; los estándares se fijan independientemente en inglés y francés.

Tabla 3: Ejemplo de estándares de desempeño para 10° Grado

Descripciones de niveles de desempeño en lectura, OSSLT en inglés	
Incompleto	Completo
El estudiante, con competencia y exactitud limitada: • Demuestra comprensión limitada de información directamente expresada. • Raramente conecta ideas e informaciones relevantes entre sí para comprender el significado del texto. • Tiene dificultad integrando su experiencia y conocimiento personal con el texto para extender su comprensión del texto	El estudiante con suficiente competencia y exactitud • Demuestra comprensión de información directamente expresada. • Usualmente conecta ideas e informaciones relevantes entre sí para comprender el significado del texto. • Generalmente usa vocabulario y estructura de oraciones apropiados. • Tiene éxito moderado integrando su experiencia y conocimiento personal con el texto para extender su comprensión del texto

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

La estrategia de comunicación de los estándares cumple distintos propósitos para distintos niveles del sistema educativo.

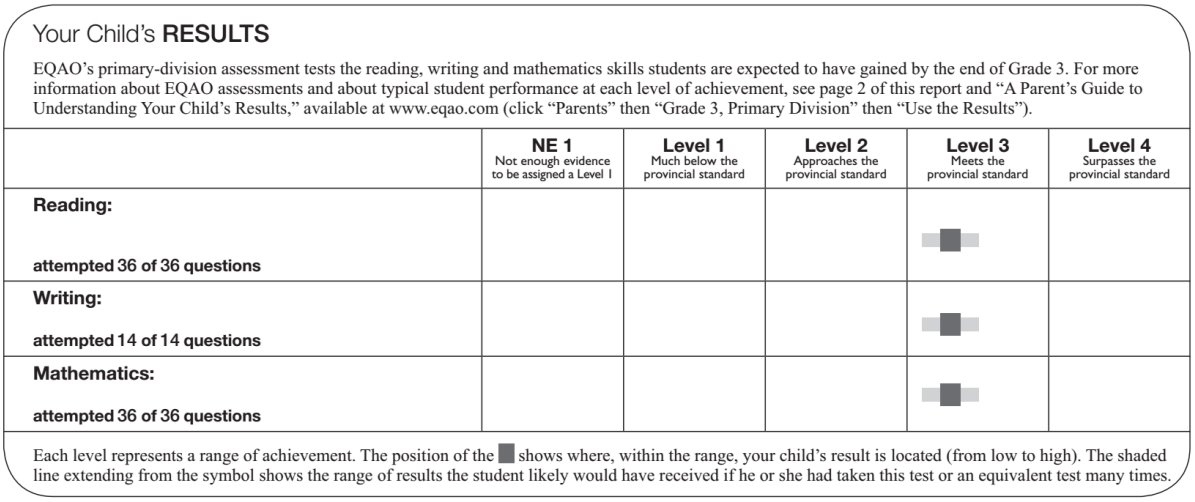
En el nivel provincial, se informan para propósitos de *accountability* sistémico: para informar acerca del estado de la escolarización en la provincia, en términos de describir los logros de los estudiantes en relación con el Currículum de Ontario. El informe provincial se hace anualmente, y se distribuye ampliamente al público y las autoridades del gobierno; está disponible en las páginas de internet de la EQAO y del Ministerio de Educación. El EQAO también tiene el encargo de difundir resultados para orientar políticas en formación docente y aquellas políticas orientadas a mejorar la implementación pedagógica del currículo provincial en las aulas.

La educación pública de Ontario está organizada en subsistemas que se llaman *School Boards* o "Juntas de Escuelas" que administran sistemas de escuelas correspondientes a subdivisiones geográficas de la provincia, aproximadamente correspondientes a comunas o municipios, y cada School Board recibe informes similares al informe provincial, también para propósitos de *accountability*. Cada escuela también recibe un informe.

Todo usuario puede acceder no solamente a los informes sino a los datos de las evaluaciones a través de la página de internet del EQAO, e inclusive llevar a cabo algunos análisis básicos de los datos utilizando una herramienta que provee la página web.

Las evaluaciones en Ontario son censales por el propósito expreso de proporcionar informes individuales para cada estudiante y su familia. Como en los otros informes mencionados, los resultados se expresan en términos de los estándares de desempeño. En la ilustración 1, se observa un ejemplo del informe individual para estudiantes, en donde se indica en que estándar de desempeño este se encuentra, y se indica el intervalo de confianza correspondiente a sus resultados. El informe también cuenta con una descripción breve de lo que significa alcanzar cada estándar, y sugerencias a los padres acerca de cómo apoyar a sus hijos e hijas en sus procesos de aprendizaje del currículum de Ontario.

Ilustración 1: Ejemplo de un informe para estudiantes en la provincia de Ontario



Your Child's **RESULTS**

EQAO's primary-division assessment tests the reading, writing and mathematics skills students are expected to have gained by the end of Grade 3. For more information about EQAO assessments and about typical student performance at each level of achievement, see page 2 of this report and "A Parent's Guide to Understanding Your Child's Results," available at www.eqao.com (click "Parents" then "Grade 3, Primary Division" then "Use the Results").

	NE 1 Not enough evidence to be assigned a Level 1	Level 1 Much below the provincial standard	Level 2 Approaches the provincial standard	Level 3 Meets the provincial standard	Level 4 Surpasses the provincial standard
Reading: **attempted 36 of 36 questions**					
Writing: **attempted 14 of 14 questions**					
Mathematics: **attempted 36 of 36 questions**					

Each level represents a range of achievement. The position of the ■ shows where, within the range, your child's result is located (from low to high). The shaded line extending from the symbol shows the range of results the student likely would have received if he or she had taken this test or an equivalent test many times.

5. USO ESTÁNDARES DE DESEMPEÑO

Los estándares de desempeño en Ontario están estrictamente vinculados al currículum provincial, como se ha indicado al describir la metodología. Las políticas de *accountability* de la provincia exigen al EQAO informar acerca del estado de la escolarización en la provincia en referencia al currículum, y los estándares de desempeño son las herramientas que se usan para cumplir ese objetivo. Además del Currículum y los resultados de las pruebas, se publican *Frameworks* ("marcos") que son documentos que explican en forma ejemplificada la relación entre los estándares de desempeño y el currículum provincial, a fines de comunicar especialmente a docentes, los propósitos de las pruebas.

De acuerdo a la información consultada, no se han realizado estudios que indaguen sistemáticamente acerca del impacto de los estándares de desempeño en la calidad educativa. Sí se hace una indagación periódica de las opiniones de distintas comunidades de interés acerca del impacto de los estándares en las pruebas, en la etapa de consulta en las evaluaciones periódicas que hemos descrito en el punto 2.3 .

6. OTRA INFORMACIÓN RELEVANTE

Es importante señalar que el EQAO, como es el caso del NAEP en EEUU, por ley, debe conducirse con total transparencia; no hay proceso técnico o dato (salvo la identificación de estudiantes y escuelas) que puede considerarse confidencial, inclusive los procesos que siguen las firmas subcontratistas. Todo está disponible al público en forma sencilla a través de la página de internet.

Es también importante señalar que el EQAO ahora enfrenta un desafío muy importante con respecto a los estándares de desempeño. Se ha anunciado que próximamente todas las evaluaciones serán en-línea, y ya se ha realizado un ensayo voluntario de la versión en-línea de la prueba de 10º grado. Todos los esfuerzos, hasta la fecha, de equiparar las pruebas escritas con las versiones en-línea indican que el comportamiento de ítemes en ambas modalidades es tan distinto que no se pueden considerar equivalentes. La transición a pruebas en-línea bien puede significar sacrificar la comparabilidad interanual de los estándares de desempeño. En la actualidad, el EQAO está llevando a cabo investigaciones (y ha encargado investigaciones a contratistas externos) para documentar mejor el problema y estudiar posibles soluciones.

7. REFERENCIAS:

Education Quality and Accountability Office (*Richard G. Wolfe, Ruth Childs and Susan Elgie, Principal Investigators for the Ontario Institute for Studies in Education, University of Toronto*). 2004. *Ensuring Quality Assessments: The Move Forward.* Toronto: Queen's Printer for Ontario.

Educational Quality and Accountability Office. 2013. *EQAO: Ontario's Provincial Assessment Program. Its History and Influence*. Toronto: Queen's Printer for Ontario.

Laveault, Dany, and Louise Bourgeois. 2014. "La Politique D'évaluation Du Rendement En Ontario: Un Alignement Qui Se Précise Dans La Persévérance et La Durée." Education et Francophonie 42 (3): 50–67.

Pinto, Laura Elizabeth. 2016. "Tensions and Fissures: The Politics of Standardised Testing and Accountability in Ontario, 1995–2015." The Curriculum Journal 27 (1): 95–112.

8. CONTACTO

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

(Entrevistas personales, entrevistas por Skype, y preguntas por correo electrónico.)

FICHA #6
NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP)
Evaluación Nacional del Progreso Educativo, EEUU

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la evaluación

El *National Assessment of Educational Progress* (NAEP)[21] – Evaluación Nacional del Progreso Educativo) es también conocido como el *Nation's Report Card* ("Reporte Nacional de Notas"). Los "Report Cards" comunican los hallazgos del NAEP y compara los resultados de la escolarización entre estados, distritos escolares (en el último censo se reporta la existencia de 13.506 distritos escolares públicos en EEUU y 178 sistemas escolares dependientes de gobiernos estatales), escuelas públicas y privadas, y distintos estratos socio demográficos.

NAEP se realiza con financiamiento del Departamento de Educación de los Estados Unidos y se ha realizado continuamente desde 1969.

1.2 Referente orientador de las evaluaciones.

Según la ley constitucional de los EEUU, toda función de gobierno no mencionado explícitamente en la Constitución de la nación, es autoridad de cada estado de la Unión. La educación es una de las funciones de gobierno que no se mencionan en la Constitución. De allí que no hay un referente nacional de evaluación, sino que el NAEP procura ser representativo (en general) del currículo de cada uno de los estados. No todos los estados tienen estándares de desempeño, programas de estudio o currículo estatal de diseño similar, y existen estándares de desempeño promovidos por organizaciones inter-estatales que han sido adoptados por varios estados (un proceso controversial incentivado por políticas federales recientes). Por tanto, el referente orientador del NAEP no es una política curricular específica, sino que son desarrollados por el propio NAEP. Se llaman los Marcos de NAEP (NAEP Frameworks)[22] que se desarrollan con la participación y consulta pública a educadores en ejercicio y oficiales de sistemas educativos, revisiones por comités conformados por decisores de política, educadores y miembros del público, inclusión de los supervisores de las distintas asignaturas en agencias educativas (de estados, distritos escolares, etc.), audiencias públicas requeridas por la ley que establece el NAEP, y la revisión 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.

[21] http://nces.ed.gov/nationsreportcard/
[22] https://nces.ed.gov/nationsreportcard/frameworks.aspx

1.3 Organismo responsable del programa de evaluación y del currículo.

Los Marcos de evaluación del NAEP deben ser aprobados por el organismo a cuyo cargo está el NAEP: el National Assessment Governing Board (NAGB, Junta de Gobierno de la Evaluación Nacional)[23].

La composición del NAGB se establece en la Ley Federal (PL 107-279 reautorizado en 2002) que establece que debe incluir dos gobernadores estatales (que no pueden ser del mismo partido político), dos legisladores estatales (que tampoco pueden ser del mismo partido político) y una serie de oficiales estatales y de distritos escolares, tres docentes de aula, un representante del mundo de los negocios o la industria, dos especialistas curriculares, tres expertos en medición educativa, un administrador o decisor de políticas de escuelas no-publicas, dos padres de familia, y dos representantes del público general.

El NAGB decide que áreas disciplinares serán evaluadas, aprueba marcos y niveles de desempeño, decide los objetivos de la evaluación, 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 ítemes que forman parte de las pruebas.

En EEUU el gobierno federal no tiene organismo responsable del currículo. Cada estado tiene distintas formas de autorizar y gobernar la política curricular.

1.4 Áreas disciplinares

El NAEP ejecuta evaluaciones periódicas en 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 para futuras evaluaciones en idiomas extranjeros.

1.5 Grados evaluados

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 referencia más viejos e informa sobre tendencias desde el año escolar de 1969-70, evalúa estudiantes a las edades de 9, 13 y 17.

1.6 Características de las pruebas

Los puntajes se calculan usando Teoría de Respuesta al Ítem (TRI) en una escala usualmente con un rango de 0-500, con Desviación Estándar de 35. Los resultados se reportan en tres niveles de desempeño en donde los puntos de corte están a una

[23] https://www.nagb.org/

desviación estándar de distancia en términos generales, pero no se hace un esfuerzo por estandarizar las distancias entre los niveles para que sean iguales en todos los grados y áreas disciplinares.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

EL NAGB está a cargo de establecer los estándares de desempeño del NAEP

2.2 Características de los estándares

Las evaluaciones de NAEP se reportan en niveles de desempeño ("achievement levels"), que son los estándares adoptados por el NAGB. Para cada nivel de desempeño existe una definición de general del nivel (que el NAGB denomina "policy definition"), una descripción de los desempeños que involucra, un conjunto de preguntas de evaluación ilustrativas, y un punto de corte. Las definiciones de política para cada estándar de desempeño son:

- Básico: Denota dominio parcial de conocimiento 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

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.

Las descripciones de los niveles de desempeño corresponden a las expectativas del marco de evaluación NAEP. Estas descripciones especifican las competencias, habilidades, contenidos, y situaciones en las que un estudiante debería poder demostrar lo que sabe y puede hacer. Las descripciones están estrictamente referidas al dominio disciplinar evaluado en las pruebas. Las descripciones se elaboran primero a partir del marco de evaluación, y luego se refinan tomando en consideración la evidencia empírica de las pruebas.

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

2.3 Historia

Los niveles de desempeño del NAEP, existen como respuesta a una demanda por parte de los Gobernadores de los Estados que en su reunión de 1986 llamaron por mejorías en la forma de reportar NAEP. Como resultado el entonces Secretario de Educación federal encargó a un panel de expertos y a la Academia Nacional de Educación dar recomendaciones que resultaron en enmiendas a la ley de NAEP (en 1988), incluyendo la controversial provisión de que el NAGB fijara estándares de desempeño para el NAEP. Como resultado, se adoptaron los tres niveles de desempeño que se usan en la actualidad.

En sus inicios, los niveles de desempeño fueron experimentales y no fueron la forma principal de reportar resultados. Esto debido no solo a lo controversial de establecer estándares a nivel federal, sino también porque dos evaluaciones externas, por parte de la Academia Nacional de Educación y la Contraloría (U.S. General Accounting Office) determinaron que los estándares aun no cumplían criterios técnicos suficientes. Tomó cerca de 20 años para que, según opinión de expertos, los procedimientos para establecer niveles de desempeño sean no solamente técnicamente suficiente, sino también reconocidos como la tecnología de punta mundial. No hay un calendario establecido de revisión, pero cada vez que se ha reautorizado la ley que establece el NAEP, se hace revisión de todos sus procesos incluídos el establecimiento de estándares de desempeño. Por tanto, se han realizado revisiones en 1988, 1994 y 2001. Estas revisiones son realizadas por la Academia Nacional de Educación y la Contraloría – siguiendo las recomendaciones de expertos – y sus resultados se toman en consideración en la reautorización de la ley.

La Academia Nacional de Education lidera el proceso, que incluye el mandato de:

1. Realizar una auditoria externa por parte de expertos de reputación internacional, de los procedimientos del NAGB para fijar los estándares de desempeño
2. Sitetizar investigaciones recientes relevante a la fijación de estándares de desempeño.
3. Examinar el uso que se les da a los informes del NAEP y la forma en que se comunican los estándares de rendimiento. Se enfoca en especial las interpretaciónes que se hacen de los informes, y la validez de esas interpretaciones.

Cada comité de revision externa de la Academia Nacional diseña su estudio y organiza su trabajo en forma autónoma después de consultar tanto con miembros del NAGB, el equipo técnico de NAEP y sus subcontratistas, y con el Centro Nacional de Estadistica Educativa de la Secretaria de Educacion (National Center for Education Statistics, Department of Education). Hay un equipo de evaluación externa con trabajo en curso, cuyo informe se anticipa para finales de la primavera de 2017.

Los estándares de desempeño se mantienen estables entre revisiones. Esto es, no se realizan cambios a las descripciones ni puntos de corte. Esto permite medir cambios en el tiempo del porcentaje de estudiantes que alcanza distintos niveles de desempeño. La necesidad de medir cambio en el tiempo conlleva a una renuencia considerable a cambiar los marcos y puntos de corte, ya que significaría sacrificar la serie de tiempo que va desde 1969. Inclusive eso es materia de discusión en torno a tener un NAEP totalmente en línea: todos los estudios realizados indican que la equiparación interanual del NAEP quedaría seriamente afectada. La forma en que esto se resuelve es que se siguen usando marcos de referencia y puntos de corte "antiguos" en algunas series de NAEP y nuevas definiciones en pruebas recientes que no cuentan con series en el tiempo tan largos. De este modo, para compatibilizar estabilidad y cambio, el NAEP realiza dos evaluaciones (estrategia dual):

1. El NAEP con resultados de largo plazo[24] (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)[25]. Reporta comparaciones interanuales dentro del período de vigencia de cada marco de evaluación.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

El NAGB es responsable por el desarrollo de estándares, diseña términos de referencia y luego organiza licitaciones públicas para contratistas que se encargan de la ejecución del trabajo. El NAGB tiene a su cargo el monitoreo y control de calidad del trabajo del contratista.

3.2 Metodología

Las metodologías para establecer estándares en NAEP son distintos según el área disciplinar y año en que se fijaron. A continuación, una tabla que resume las metodologías usadas por área disciplinar:

[24] https://nces.ed.gov/nationsreportcard/ltt/
[25] https://nces.ed.gov/nationsreportcard/subjectareas.aspx

Metodología	Áreas disciplinares y años.
Angoff Modificado	Matemáticas, Lectura y Escritura (1992) Geografía, Historia de los EEUU y Ciencias Naturales (1994; 1996)
Método ACT/NAGB	Educación cívica y escritura (1998)
Bookmark Modificado (Mapmark con Dominios)	Matemáticas de 12º grado (2005); Economía de 12º grado (2006); Tecnología y "Engineering Literacy" (2012)

Las metodologías usadas por NAEP en el establecimiento de estándares de desempeño han evolucionado de metodologías que podrían llamarse "de ítem en ítem" (item-by-item) a métodos más holísticos. Se partió por una metodología Angoff modificado en donde panelistas emiten juicios acerca de cómo los estudiantes en cada nivel de desempeño debían desempeñarse en cada ítem de la prueba.

Comenzando en 1994 se experimentó con el uso de mapas de ítemes para facilitar los juicos de los panelistas. Mapas de ítemes indican donde está localizado cada ítem en puntos de la escala NAEP y se propusieron para ayudar a los panelistas a entender como los puntos de corte indican lo que cada estudiante puede hacer, no puede hacer, o puede encontrar desafiante.

En 1998 se modificó el proceso aún más con la introducción de Diagramas de Reckase (Reckase Charts – inventados por Mark Reckase en ACT, una empresa contratista), para proporcionar realimentación a los panelistas. El experimento se juzgó exitoso y suficientemente diferente al método Angoff como para ameritar llamarse el método ACT/NAGB.

Desde 2005 se utiliza una modificación del método Bookmark que actualmente se llama "Mapmark con Dominios". Este método ofrece ventajas sobre Angoff: puede incluir tanto preguntas abiertas como de opción múltiple bajo el mismo proceso de juicios. Es más fácil de implementar y el cómputo de puntos de corte es más sencillo. La modificación Mapmark incluye además una dimensión espacial que permite a los panelistas juzgar las ubicaciones de los "bookmarks" que indican los límites de cada nivel de desempeño. Además, esta modificación incluye un aspecto holístico en vez de limitar los juicios a ítemes individuales[26].

El proceso de establecimiento de estándares actualmente incluye muchos pasos distintos, con retroalimentación después de cada oportunidad para emitir juicios. Los materiales de consulta más importante son los marcos NAEP, las definiciones de política de los niveles de desempeño, el banco de ítemes de las pruebas, los parámetros de los ítemes del banco, los Diagramas Reckase y los Cuadros Espaciales Mapmark. Los miembros de los paneles comienzan por tomar decisiones acerca del desempeño por ítemes en el banco basados en las definiciones de los niveles. Después de cada ronda de fijación de estándares (en total son típicamente 5) los panelistas reciben realimentación acerca de sus juicios y un análisis estadístico acerca de las consecuencias de sus decisiones (por ejemplo, el porcentaje de estudiantes que sería clasificado en cada nivel, nuevos Diagramas Reckase, etc.) Después de tres rondas de este tipo, se realizan evaluaciones comprensivas del procedimiento y sus consecuencias, y se proceda una recomendación de puntos de corte que son considerados en una cuarta ronda. La quinta ronda incluye la selección final de ítemes para la prueba y de ítemes ilustrativos para describir los niveles de desempeño. El proceso se evalúa en cuanto a fidelidad de cumplimiento de las tareas y confianza de los participantes en la validez procedimental.

Los paneles se componen siguiendo estrictamente las políticas establecidas por NAGB: 70% de los panelistas deben ser educadores y 30% no son educadores. La política de NAGB especifica que entre los grupos de educadores, 55% de los educadores deben esenar actualmente la asignatura en el grado evaludo, otro 15% deben ser educadores con conocimiento de la asigntura y el grado, pero que no enseñan en ese grado

---

[26] Para mayor detalle sobre Mapmark ver Schultz, E., & H. Mitzel (2005). The Mapmark Standard Setting Method. Paper presented to the National Assessment Governing Board for the National Assessment of Educational Progress, 2005. Disponible en http://files.eric.ed.gov/fulltext/ED490643.pdf.
Hay que tener en cuenta que todos estos trabajos lo realizan empresas subcontratantes de NAGB, y por tanto existen modificaciones en los procedimientos de empresa en empresa. El equipo técnico de NAEP evalúa los procedimientos en las propuestas que someten las empresas en el proceso de licitación pública y determina las propuestas técnica y economicamente más sólidas. Las políticas de NAGB solo establecen los porcentajes de grupos que deben estar en los paneles de jueces, por ejemplo: las empresas o agencias que concursan por los contratos proponen distintas maneras de cumplir con las estipulaciones, y en el proceso de adjudicación de los contratos, se evalúan esas propuestsa.

actualmente, y 30% deben ser no-educadores con conocimiento experto de la asignatura evaluada y de las población estudiantil a evaluar. Los educadores deben incluir docentes que enseñan el área curricular en el grado evaluado y otros tipos de educadores como directores/supervisores de currículo o especialistas curriculares de distritos o Estados. Pueden ser también académicos universitarios que imparten clases a futuros educadores para el área disciplinar y grado. Los no-educadores puede ser individuos preparados en el área disciplinar y con familiaridad con estudiantes del grado evaluado, pero sin experiencia significativa en la enseñanza. Los paneles además deben ser representativos en género, raza/etnicidad, y regiones geográficas.

Los cambios que se han ido dando en los procedimientos para establecer estándares de desempeño en NAEP se han hecho principalmente para mejorar la calidad y cantidad de información que reciben los miembros de los paneles de fijación de estándares. Estos métodos, principalmente gráficos, proporcionan en cada innovación, mejorías en la presentación de información acerca de las categorizaciones de los ítemes de la prueba, los datos acerca del desempeño de estudiantes en los ítemes categorizados, y los datos acerca de las consecuencias de distintas decisiones acerca de los estándares para el uso de jueces-panelistas. Por *datos de consecuencias*, el NAEP entiende datos acerca de los porcentajes de los estudiantes que alcanzarían los distintos niveles de desempeño, según distintas decisiones acerca de la categorización de ítemes en distintas posibles categorías de desempeño. El NAEP considera en especial el uso de datos de consecuencias una de las innovaciones más importantes en sus procedimientos de fijación de estándares. Se estima que el uso de estos datos mejora las decisiones finales sobre estándares y distintas metodologías sucesivas han tenido el fin de facilitar a los panelistas el análisis (especialmente visual) de los datos acerca de ítemes, categorías de desempeño, y las consecuencias de distintas definiciones de estándares de desempeño, para mejorar la calidad de las decisiones finales. Al concluir cada ronda de trabajo en los paneles de trabajo en fijación de estándares, se entrevistan y encuestan a los participantes y esta información se utiliza en la mejora continua de los procedimientos, con especial atención en encontrar mejores métodos de presentación de información para facilitar y mejorar los estándares de desempeño que resultan de su trabajo.

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

La estrategia de comunicación de los estándares al público es principalmente mediante la publicación del "Nation's Report Card" – y mediante la puesta al uso del publico de herramientas en internet para conducir análisis básicos propios en la base de datos, y otros exploradores de datos. Toda la información técnica, de procedimientos, y datos también están a la disposición del público. (Por disposición de la ley no se proporcionan datos para identificar escuelas o estudiantes individuales). Es responsabilidad de los estados difundir resultados del NAEP en su territorio. También se divulgan videos en internet, y presentaciones en diapositivas explicando los niveles de desempeño y los resultados de la evaluación. Los datos acerca de los niveles de desempeño los usan también las asociaciones profesionales de docentes en sus propios análisis e informes. La comunidad científica en medición y currículo también analiza los datos y procedimientos de NAEP. Los datos e informes de NAEP se utilizan en la formación pre servicio y en la formación continua de docentes.

Como hemos descrito arriba, en las evaluaciones periódicas realizadas por la Academia Nacional de Educación, se realizan estudios del impacto de los informes de NAEP. Es importante recordar que, en el contexto de los Estados Unidos, no se espera que NAEP tenga un impacto decisivo en las prácticas pedagógicas o en otros aspectos similares. Cada Estado de la Unión tiene autoridad sobre sus propios estándares y la Constitucion de los Estados Unidos no estipula autoridad federal en política curricular. La evaluación de los informes del NAEP se limitan a procurar que las conclusiones que se toman en base a sus informaciones son válidas y confiables.

5. USO ESTÁNDARES DE DESEMPEÑO

Los estándares de NAEP no son una política para establecer estándares nacionales y por consiguiente no hay responsabilidad por incentivar o monitorear su adopción en los estados. Sin embargo, son muy influyentes, y los procedimientos para establecer estándares de desempeño en las evaluaciones estatales y las del NAEP se han influido recíprocamente. La amplia divulgación de sus datos y documentos técnicos, su uso de auditorías externas, etc. han influido en los estándares profesionales técnicos y de transparencia de las evaluaciones estatales y distritales de EEUU

6. OTRA INFORMACIÓN RELEVANTE

Es importante señalar que el NAEP, por ley, debe conducirse con total transparencia – no hay proceso técnico o dato (salvo la identificación de estudiantes y escuelas) que puede considerarse confidencial. Todo está disponible al público. Inclusive los datos "confidenciales" identificando escuelas y estudiantes están disponibles a investigadores que se someten voluntariamente a normas legales y de ética profesional monitoreadas por NAGB.

NAGB conduce auditorías externas por comisión regularmente, pero también sus procesos están bajo el escrutinio permanente de la comunidad científica y el uso intensivo de sus datos han llevado a análisis críticos que contribuyen al mejoramiento continuo del NAEP.

7. REFERENCIAS:

Bourque, Mary Lyn. 2009. "A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989-2009." Paper Commissioned for the 20th Anniversary of the National Assessment Governing Board 1988–2008. Washington DC: National Assessment Governing Board.

Loomis, Susan Cooper, and Mary Lyn Bourque. 2001. "From Tradition to Innovation: Standard Setting on the National Assessment of Education Progress." In *Setting Performance Standards: Concepts, Methods, and Perspectives*, edited by Gregory J. Cizek, 175–218. Mahwah, NJ: Lawrence Erlbaum Associates.

NAGB. 2010. "Setting Standards on the National Assessment of Educational Progress in Reading and Mathematics for 12th Grade Preparedness." Design Document for 12th Grade NAEP Preparedness Research Judgmental Standard Setting Studies. Washington DC: American College Testing ACT Inc & National Assessment Governing Board.

Peterson, Christina Hamme, E. Matthew Schulz, and George Jr Engelhard. 2011. "Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress." *Educational Measurement: Issues & Practice* 30 (2): 3–14.

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

Teresa Neidorf fue entrevistada en persona en tres ocasiones.

ESTÁNDARES DE DESEMPEÑO
FICHA #7
ELA/Literacy and Mathematics Common Core tests – New York State
Tests de Aprendizajes Básicos Comunes para Lenguaje y Matemática – Estado de Nueva York

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la evaluación

Desde comienzos del año 2013, el Estado de Nueva York ha implementado una evaluación censal conocida como Tests de Aprendizajes Básicos Comunes para Lenguaje y Matemática (*ELA/Literacy and Mathematics Common Core tests*). Los propósitos de esta evaluación son, en primer lugar, determinar en qué medida los estudiantes alcanzan expectativas nacionales de aprendizaje (Common Core State Standards), y contribuir a las políticas de mejoramiento escolar y responsabilización por resultados de parte de los establecimientos escolares del estado.
Antes del año 2013, las evaluaciones estatales de aprendizaje estabán referidas a estándares definidos a nivel estatal y no nacional, como son actualmente los Common Core State Standards. Tal como se señala más adelante, el cambio en los estándares empleados requirió, de parte del Departamento de Educación, indicar algunas precauciones para la interpretación de resultados a través del tiempo.

1.2 Referente orientador de las evaluaciones.

Las evaluaciones censales aplicadas en el Estado de Nueva York están referidas al curriculum estatal de Nueva York (*New York State Learning Standards*) y a los Estándares Estatales de Aprendizajes Básicos Comunes (*Common Core State Standards*).

Los Estándares estatales de Nueva York se definen para las diferentes áreas del currículum, ampliando y especificando las expectativas establecidas en los Estándares Estatales de Aprendizajes Básicos Comunes (los cuales se describen más adelante). Estos últimos, establecen altas expectativas en Matemáticas y Lenguaje para estudiantes desde Kindergarten hasta el grado 12, definiendo lo que los estudiantes deben saber y poder hacer al final de cada uno de estos cursos. Su propósito es asegurar que todos los estudiantes finalicen su escolaridad con las habilidades y el conocimiento necesario para desenvolverse con éxito en la educación superior, como profesionales y ciudadanos, con independencia del lugar donde viven.

Los Estándares Estatales de Aprendizajes Básicos Comunes comenzaron a elaborarse en el año 2009. Dado que en EEUU no hay un curriculum nacional, cada estado desarrolla de manera independiente sus propios estándares de aprendizaje. La iniciativa de desarrollar estándares comunes surgió de los estados y autoridades agrupados en el Centro de Mejores Prácticas de la Asociación Nacional de Gobernadores (*National Governors Association Center for Best Practices*) y el Consejo de Funcionarios de Jefaturas Escolares Estatales (*Council of Chief State School Officers*). Esta iniciativa surge de la necesidad de tener estándares que sean consistentes y exigentes en todos los estados de la unión, y que preparen adecuadamente a todos los estudiantes para la educación terciaria o para ingresar a la fuerza laboral.

Los Estándares Estatales de Aprendizajes Básicos Comunes o *Common Core Standards* fueron implementados por este estado en enero del año 2011. Los Estándares Estatales de Aprendizajes Básicos Comunes se desarrollaron a partir del análisis de: las expectativas establecidas en los estándares estatales previamente elaborados en algunos estados; las expectativas de otros países de alto rendimiento en todo el mundo; y la investigación y literatura disponible sobre lo que los estudiantes necesitan saber y ser capaces de hacer para tener éxito en la universidad, su trabajo y la vida. Además, fueron sometidos a consulta con docentes, padres y madres a nivel nacional, para asegurar que se vincularan con la realidad de los estudiantes y fueran prácticos para los docentes.

Hasta el momento, 42 estados de los Estados Unidos de América, el distrito de Columbia, cuatro territorios y el Departamento de Defensa de la Educación han suscrito voluntariamente estos estándares. Los estándares pueden ser consultados en: http://www.corestandards.org/read-the-standards/

1.3 Organismo responsable del programa de evaluación y del currículo.

La Oficina de Evaluación Estatal (*Office of State Assessment - OSA*) es responsable de la coordinación, desarrollo e implementación de los tests censales de 3º a 8º grado, junto con otras evaluaciones estatales, tales como los Exámenes "Regents" (*Regents Examinations*), que forman parte del Programa de Evaluaciones del Estado de Nueva York (*New York Testing Program – NYSTP*)[27].
Los tests censales son elaborados por una agencia externa al Departamento de Educación del Estado de Nueva York. En el año 2016 esta agencia fue *Questar Assessment Inc* (antes, los tests habían sido encargados a la agencia *Pearson*).

[27] Más información en: http://www.p12.nysed.gov/assessment/testingprogram.html

Por su parte, el desarrollo del curriculum estatal está a cargo de la Oficina de Curriculum y Enseñanza (*Office of Curriculum and Instruction*).
Ambas instituciones dependen del Departamento de Educación del Estado de Nueva York.

1.4 Áreas disciplinares

Las evaluaciones estatales censales de Nueva York consideran las áreas de Lenguaje y Matemática.

1.5 Grados evaluados

Este programa se aplica a estudiantes de los grados 3º (8 a 9 años de edad) a 8º (13 a 14 años de edad) de establecimientos públicos, privados no subvencionados y privados subvencionados de todo el Estado.

1.6 Características de las pruebas

Las evaluaciones censales de Lenguaje y Matemática del estado de Nueva York son administradas al interior de cada establecimiento educacional en tres sesiones durante tres días consecutivos. La administración se realiza a través de docentes de la misma escuela. En las sesiones de administración de los tests, los estudiantes deben responder preguntas de selección múltiple de cuatro opciones y de respuesta abierta breve y también extensa.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

Del mismo modo que la evaluación estatal, el desarrollo de los estándares de desempeño está a cargo del Departamento de Educación del Estado de Nueva York.

2.2 Características

Los estándares de desempeño de los tests del estado de Nueva York se conocen como Descripciones de Niveles de Desempeño (*Performance Level Descriptions*) y, al igual que las evaluaciones del estado, están referidos a los Estándares Estatales de Aprendizajes Básicos Comunes y al curriculum de Nueva York.

Estos estándares de desempeño se materializan a través de la descripción de cuatro niveles de desempeño para cada grado evaluado, las que indican si un estudiante excede la expectativa planteada en los Estándares Estatales de Aprendizajes Básicos Comunes, si su desempeño coincide con esta expectativa, si están un poco bajo esta expectativa o muy lejos de alcanzar esta expectativa.

Estos estándares de desempeño tienen una descripción genérica similar para ambas asignaturas (Lenguaje y Matemática) y descripciones detalladas para diferentes elementos o "*clusters*" de cada disciplina.

En el caso de matemática, el siguiente cuadro muestra la descripción genérica de los estándares de desempeño:

NYS Level 4
Students performing at this level **excel** in standards for their grade**.** They demonstrate knowledge, skills, and practices embodied by the New York State P-12 Common Core Learning Standards for Mathematics that are considered **more than sufficient** for the expectations at this grade.

NYS Level 3
Students performing at this level are **proficient** in standards for their grade**.** They demonstrate knowledge, skills, and practices embodied by the New York State P-12 Common Core Learning Standards for Mathematics that are considered **sufficient** for the expectations at this grade.

NYS Level 2
Students performing at this level are partially proficient in standards for their grade. They demonstrate knowledge, skills, and practices embodied by the New York State P-12 Common Core Learning Standards for Mathematics that are considered partial but insufficient for the expectations at this grade. Students performing at Level 2 are considered on track to meet current New York high school graduation requirements but are not yet proficient on Common Core Learning Standards at this grade.

NYS Level 1
Students performing at this level are **well below proficient** in standards for their grade. They demonstrate **limited** knowledge, skills, and practices embodied by the New York State P-12 Common Core Learning Standards for Mathematics that are considered **insufficient** for the expectations at this grade.

El siguiente cuadro muestra la descripción más específica de los estándares de desempeño para uno de los elementos o clusters que se distinguen en Matemática para 3º grado:

Cluster	Performance Level 4	Performance Level 3	Performance Level 2	Performance Level 1
Students represent and solve problems involving multiplication and division. (3.OA.1-4)	Interpret and represent products and quotients of whole numbers.	Interpret products and quotients of whole numbers.	Interpret products of whole numbers.	Given visual models and/or manipulatives, interpret products of whole numbers with factors less than or equal to 5.
	Determine the unknown whole number in a multiplication and division problem by relating multiplication and division.	Determine the unknown whole number in a multiplication or division equation relating three whole numbers by relating multiplication and division. Factors are greater than 5 and less than 10.	Determine the unknown whole number in a multiplication equation by relating multiplication and division. Limit to factors less than or equal to 5.	Determine the product in a multiplication equation with whole number factors less than or equal to 5.
	Represent a multiplication or division situation as an equation.			
	Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area.	Use multiplication and division within 100 to solve word problems involving equal groups, arrays, area, and measurement quantities other than area. Both factors are less than or equal to 10.	Given visual models and/or manipulatives, use multiplication and division within 100 to solve word problems involving equal groups and arrays. Both factors are less than or equal to 10.	Given visual models and/or manipulatives, compute products within 25 in the context of word problems.
	Identify proper context given a numerical expression involving multiplication and division. Both factors are less than or equal to 10.			

2.3 Historia

Las Descripciones de Niveles de Desempeño comenzaron a elaborarse en el año 2013, luego que el Estado de Nueva York adoptara los *Common Core* standards para orientar sus evaluaciones. Antes de esto, las pruebas estaban referidas a otros estándares de desempeño cuyo nivel de exigencia respondía solo a expectativas estatales y no nacionales.

Cuando se cambiaron los estándares de desempeño, el Departamento de Educación debió aclarar que, dado que los nuevos tests miden habilidades complejas diferentes de las habilidades medidas por los tests estatales aplicados previamente, los puntajes de los estudiantes pueden parecer más bajos. El Departamento de Educación aclaró que esto no significaba necesariamente que los estudiantes lo estén haciendo peor, si no que la vara es más alta y los resultados no son comparables.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

El Departamento de Educación del Estado de Nueva York convocó a los Paneles Asesores de Contenido (*Content Advisory Panels*) para desarrollar los primeros

borradores de las descripciones de niveles de desempeño para los grados 3 a 8. Estos Paneles Asesores de Contenido estaban conformados por docentes, directores de escuelas, administradores educacionales distritales, especialistas en la enseñanza de estudiantes cuya primera lengua no es inglés, especialistas de en la enseñanza de estudiantes con discapacidades y miembros de facultades de educación del estado.

3.2 Metodología

Para el desarrollo de las Descripciones de Niveles de Desempeño los Paneles Asesores de Contenido consideraron tanto los Estándares estatales de Nueva York (*New York State Learning Standards*) como los Estándares Estatales de Aprendizajes Básicos Comunes (*Common Core State Standards*). Las descripciones de cada asignatura comenzaron a desarrollarse a partir del nivel 3, que es el que define los conocimientos y habilidades que se esperan de un estudiante que alcanza la expectativa estatal y nacional. Luego se definieron los niveles 4 (excede la expectativa) y 2 (logro parcial de la expectativa). Finalmente se describió el nivel 1, el que abarca un amplio rango de desempeño, desde los estudiantes que se ubican justo debajo de los requerimientos del nivel 2, hasta el desempeño de aquellos estudiantes que no lograron responder ninguna respuesta correctamente.

Luego de que los Paneles Asesores de Contenido realizaran un primer borrador de las Descripciones de Niveles de Desempeño, este fue sometido a rondas de revisión y edición a partir de comentarios de expertos en la disciplina y en evaluación, siempre bajo la supervisión del Departamento de Educación del Estado de Nueva York. Este proceso comenzó en 2010 con la adopción de los Common Core State Standards y, para el caso de Lenguaje y Matemática, finalizó en 2013 con la aplicación de las primeras pruebas referidas a ellos.

Estas descripciones fueron luego empleadas para establecer los puntajes de corte en las evaluaciones que permitirían clasificar el desempeño de los estudiantes en cada nivel. En este proceso, las Descripciones de Niveles de Desempeño fueron empleadas para describir los conocimientos y habilidades de un estudiante que, en las evaluaciones estatales, justo alcanza los requerimientos para ser clasificado en los niveles 2, 3 o 4.

Estas descripciones son luego usadas para fijar el puntaje que debe alcanzar un estudiante para que su desempeño sea ubicado en cada uno de los cuatro niveles. Esto se logra a través de una metodología diseñada e implementada por el Departamento de Educación de Nueva York junto con la colaboración de su Comité Técnico Asesor

(*Technical Advisory Committee*). Dicha metodología considera al método Bookmark como principal herramienta.

Para establecer los puntajes de corte por primera vez, en las áreas de Lenguaje y Matemática, se trabajó durante cinco días con cuatro grupos que incluían, en total, a 95 personas entre las que se cuentan docentes, administradores educacionales y académicos de instituciones de educación superior.

Resulta interesante destacar que un subgrupo de los participantes en el proceso de establecimiento de puntajes de corte (entre 17 y 18 personas por asignatura), fueron convocados en el quinto día a participar de un proceso denominado como "articulación vertical". El propósito de este procedimiento fue revisar los datos de impacto asociados con las puntuaciones de corte recomendadas en cada grado. Se pidió a los participantes que determinaran si las puntuaciones de corte recomendadas eran razonables dado el conjunto de expectativas esbozadas en cada grado, las características de los estudiantes, datos de otras evaluaciones como NAEP y el tipo de tareas incluidas en las evaluaciones.

Los pasos en el proceso de articulación vertical fueron los siguientes:

1. Los participantes revisaron las descripciones de los estándares de desempeño asociados con todos los grados evaluados.

2. Como grupo, los participantes discutieron la relación observada entre estas descripciones y el contenido evaluado en cada grado.

3. El grupo revisó los datos de impacto (cómo se distribuirían los estudiantes) asociados con las puntuaciones de corte recomendadas y luego discutió hasta qué punto los datos coincidían con sus expectativas.

4. Como grupo, el comité discutió si las puntuaciones de corte deberían ser ajustadas para ser más consistentes con sus expectativas.

5. Si se consideraron necesarios ajustes, se proporcionó a los participantes información estadística que limitara los ajustes posibles al puntaje de corte.

6. Los participantes formularon recomendaciones independientes sobre eventuales cambios a los puntajes de corte recomendados.

7. Las recomendaciones fueron procesadas y se presentaron nuevos resultados a los participantes.

8. Los participantes discutieron si estos nuevos resultados representaban mejor sus expectativas y realizaron una recomendación final.

Resulta interesante de este proceso que fue llevado a cabo con un subgrupo de participantes que revisaron los puntajes de corte para todos los grados en cada asignatura y que pudieron contar con información diferente a la entregada por las pruebas estatales, incluyendo NAEP, por ejemplo.

Luego de este proceso, los puntajes de corte fueron aprobados por las autoridades estatales.

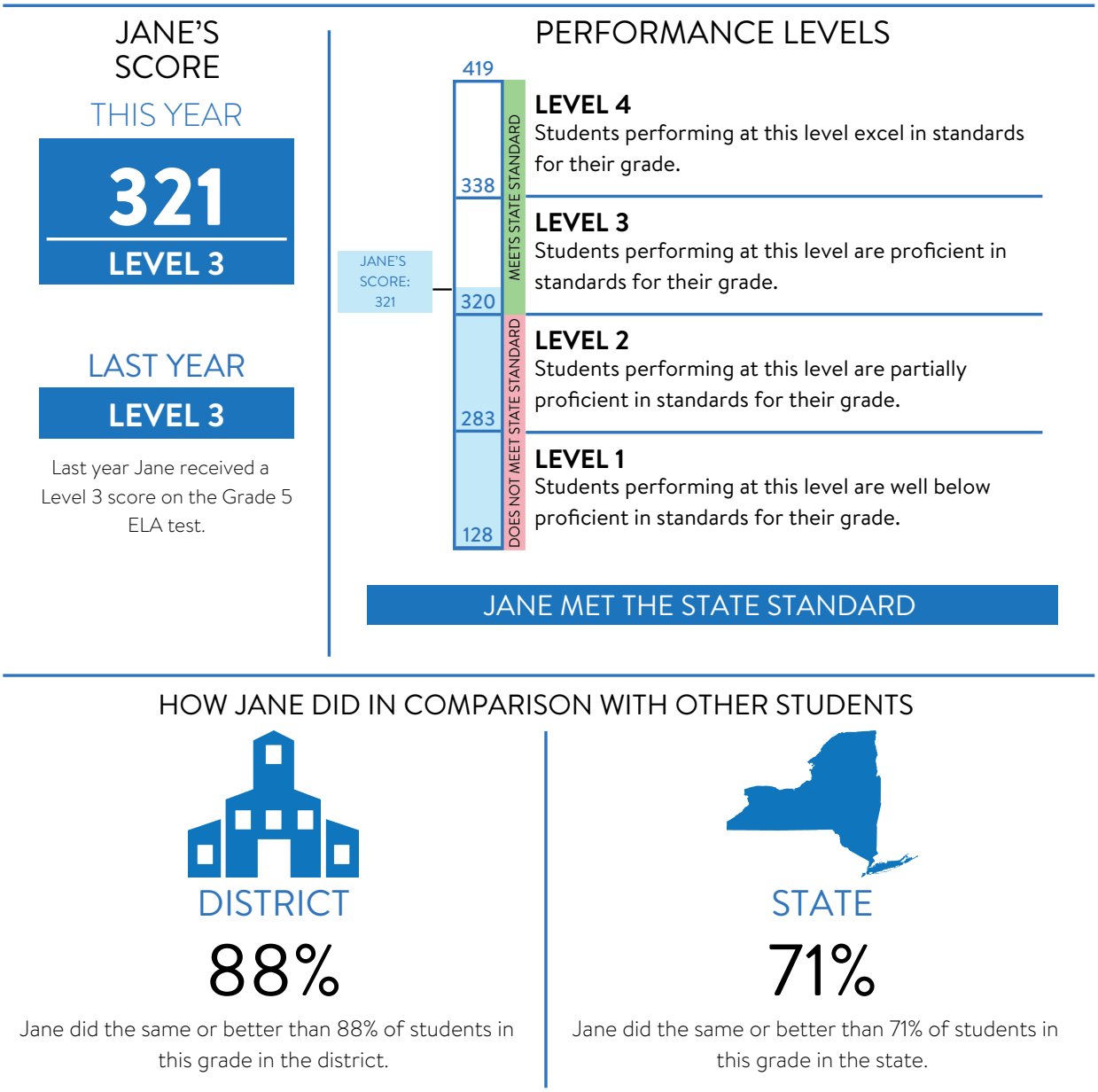
5. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

Las Descripciones de Niveles de Desempeño son comunicadas a partir del año 2013 a través de los reportes de las evaluaciones censales estatales y a través de un documento publicado en el sitio web "*EngageNY*[28]", el cual tiene como propósito apoyar y promover diversos procesos relacionados con reformas en el área de evaluación y responsabilización por resultados impulsadas por el Departamento de Educación del Estado de Nueva York.

Cada estudiante recibe un reporte de resultados individuales, donde se indica el nivel y puntaje alcanzado en las evaluaciones estatales del año y el nivel que alcanzó en la evaluación del año anterior. También se indica cómo se compara el desempeño demostrado por el estudiante con el de estudiantes del mismo distrito y del estado (se expresa indicando el porcentaje de estudiantes que tiene un puntaje inferior al del estudiante).

[28] https://www.engageny.org/

La siguiente imagen muestra un ejemplo de reporte de un estudiante:



A nivel de escuela, se reporta la cantidad y porcentaje de estudiantes que alcanza cada nivel, poniendo especial énfasis en la cantidad de estudiantes que supera el nivel 3 (competente). Esta información se desagrega para estudiantes con discapacidad; estudiantes de origen latino, blancos y multirracial; mujeres y hombres; estudiantes que no hablan inglés como primera lengua; estudiantes con desventajas económicas y sin estas desventajas; estudiantes migrantes y no migrantes.

5. USO ESTÁNDARES DE DESEMPEÑO

Mientras los resultados en las pruebas censales estatales son empleadas para responsabilizar a los establecimientos, las Descripciones de Niveles de Desempeño se emplean principalmente para comunicar a los estudiantes, sus familias y educadores y al público general en qué medida se están logrando las expectativas de aprendizaje del estado. Estos resultados pueden ser consultados en el sitio web: *data.nysed.gov*.

Complementariamente, el Departamento de Educación del estado espera que estas descripciones constituyan la base de las discusiones pedagógicas acerca de lo que necesitan los estudiantes para progresar en su aprendizaje y alcanzar la excelencia. Algunos ejemplos sugeridos por el Departamento de Educación para que los docentes usen las Descripciones de Niveles de Desempeño son:

- Diferenciar la enseñanza según el nivel alcanzado por cada estudiante para maximizar oportunidades de aprendizaje.
- Desarrollar evaluaciones de aula y rúbricas orientadas a identificar niveles de aprendizaje esperados para estudiantes o grupos de estudiantes.
- Monitorear el progreso de los estudiantes a través del continuo descrito por las Descripciones de Niveles de Desempeño.

Finalmente, las Descripciones de Niveles de Desempeño son empleadas para orientar el proceso de elaboración de la pruebas censales estatales, ya que estas deben ser capaces de diferenciar entre estudiantes que alcanzan o no un determinado nivel de desempeño.

De acuerdo a la información recopilada hasta ahora, no se han realizado estudios acerca del impacto de los estándares de desempeño de este estado en el aprendizaje de los estudiantes.

6. OTRA INFORMACIÓN RELEVANTE

No aplica.

7. REFERENCIAS:

THE STATE EDUCATION DEPARTMENT (2014). New York State Testing Program: Common Core English Language Arts Test, Performance Level Descriptions. Consultado en octubre 2016 en: http://documentslide.com/documents/performance-level-descriptions-grade-3.html

THE STATE EDUCATION DEPARTMENT (2014). New York State Testing Program: Common Core Mathematics Test, Performance Level Descriptions. Consultado en octubre 2016 en: http://documentslide.com/documents/performance-level-descriptions-grade-3.html

PTA (2016). Parents' Guide to Assessments in New York. Consultado en octubre 2016 en: http://s3.amazonaws.com/rdcms-pta/files/production/public/ptaupload/New%20York%20Assessment%20Guide%202016.pdf&sig2=oktA3tydjcIwdeMs-iOj6w&bvm=bv.140496471,d.Y2I

Pearson (2013). New York State Testing Program 2013: English Language Arts Mathematics Grades 3–8. Technical Report. Consultado en Noviembre 2016 en: http://www.p12.nysed.gov/assessment/reports/2013/ela-math-tr13.pdf

Sitios de interés:

Departamento de Educación del Estado de Nueva York: http://www.nysed.gov/
Engage Nueva York: https://www.engageny.org/

8. CONTACTO

No se contactó a ningún especialista para la elaboración de esta ficha.

FICHA #8
Standards of Learning (SOL) Tests – Virginia
Tests de Estándares de Aprendizaje – Estado de Virginia, Estados Unidos

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la evaluación

En el Estado de Virginia se aplican los Tests de Estándares de Aprendizaje (Standards of Learning (SOL) Tests). Estos se aplican desde el año 1998 y tienen por propósito informar a padres, madres y a la comunidad si cada estudiante y distintas agrupaciones de estudiantes están logrando las expectativas de aprendizaje estatales. Los estudiantes que no aprueban los tests estatales deben ingresar a un programa de apoyo remedial.

Por otro lado, estos tests permiten a la Junta de Educación estatal identificar a los establecimientos que necesitan ayuda. También permiten obtener información objetiva acerca de brechas en resultados de aprendizaje entre diferentes subgrupos de estudiantes y establecer metas a nivel de establecimiento, divisiones territoriales y del estado, para cerrar estas brechas.

En particular, los tests de educación secundaria determinan el ingreso a este nivel educacional (Lectura, Escritura y Álgebra II).

1.2 Referente orientador de las evaluaciones.

Los tests censales de Virginia están referidos a los Estándares de Aprendizaje (curriculum estatal) establecidos en el año 1995. Estos estándares representan un amplio consenso acerca de lo que padres, madres, docentes de aula, administradores escolares y líderes académicos, de negocios y de la comunidad, creen que debiera ser mínimamente enseñado y aprendido en las escuelas de Virginia.

Estos estándares se definen para cuatro asignaturas consideradas centrales: Inglés (Lenguaje), Matemática, Ciencias Naturales, e Historia y Ciencias Sociales. Para cada asignatura y en cada grado de la trayectoria escolar, se distinguen ejes disciplinarios y objetivos de aprendizaje para cada eje disciplinario.

Desde el año 2000, los Estándares de Aprendizaje son revisados considerando ciclos de siete años.

1.3 Organismo responsable del programa de evaluación y del currículo.

Tanto los Estándares de Aprendizaje como los tests referidos a ellos son responsabilidad del Departamento de Educación de Virginia (Virginia Department of Education).

El Departamento de Educación de Virginia trabaja en colaboración con la Junta de Educación, docentes, administradores escolares y académicos en el desarrollo de los Tests de Estándares de Aprendizaje. Esta colaboración se realiza a través de la formación de comités que revisan las pruebas para asegurar que miden el aprendizaje de los estudiantes de manera precisa y ecuánime. Todos los ítemes de estos tests son revisados al menos dos veces por comités de docentes de aula; solo los ítemes considerados por estos comités como justos y alineados a los estándares de aprendizaje pueden ser incluidos en las pruebas.

1.4 Áreas disciplinares

Los Tests de Estándares de Aprendizaje consideran las áreas de Inglés (Lectura y Escritura), Matemáticas (Matemáticas, Álgebra I, Geometría y Álgebra II), Ciencias Naturales (Ciencias Naturales, Ciencias de la Tierra, Biología y Química) e Historia y Ciencias Sociales (Estudios de Virginia, Educación Cívica y Economía, Historia Mundial y Geografía hasta el año 1500, Historia Mundial y Geografía desde el año 1500 al presente, Geografía Universal, e Historia de Virginia y de los Estados Unidos).

1.5 Grados evaluados

Si bien la evaluación estatal es censal, no todos los grados son evaluados en todas las asignaturas consideradas en los tests.

Entre los grados 3 y 8, los estudiantes son evaluados anualmente en lectura y matemáticas. Durante la educación secundaria los estudiantes son evaluados una vez en lectura y al finalizar los siguientes cursos: Álgebra I, Geometría y Álgebra II. Los estudiantes de los grados 5 y 8 son evaluados en Ciencias Naturales y al finalizar los siguientes cursos de educación secundaria: Ciencias de la Tierra, Biología y Química. Finalmente, los estudiantes son evaluados en Historia y Ciencias Sociales al concluir el curso de Estudios de Virginia en educación primaria y los siguientes cursos de educación secundaria: Educación Cívica y Economía, Historia Mundial y Geografía

hasta el año 1500, Historia Mundial y Geografía desde el año 1500 al presente, Geografía Universal, e Historia de Virginia y de los Estados Unidos.

En 2014, Virginia eliminó las siguientes cinco pruebas: Ciencias Naturales de Grado 3, Ciencias Sociales e Historia de Grado 3, Historia de Estados Unidos hasta 1865, Historia de Estados Unidos 1865 al Presente y Escritura de Grado 5. La eliminación de estas pruebas redujo el número total de evaluaciones de 34 a 29. La Junta de Educación y el Comité Consultivo de Innovación de los Estándares de Aprendizaje están estudiando incluir otras medidas para reducir la carga involucrada en la aplicación de pruebas, sin alterar condiciones óptimas para la rendición de cuentas.

1.6 Características de las pruebas

Cada forma de los Tests de Estándares de Aprendizaje está compuesto por entre 35 a 50 ítemes, principalmente de selección múltiple. En las áreas de Matemática, Inglés y Ciencias Naturales además se incluye "ítemes reforzados tecnológicamente" (technology-enhanced ítemes). Este último tipo de ítemes son de carácter interactivo y contemplan diversos tipos de formatos y modos de responder que requieren que el estudiante demuestre su capacidad de pensamiento crítico y de resolución de problemas; actualmente están disponibles para todos los tests, excepto para los de Historia y Ciencias Sociales. Los test de escritura administrados en el grado 8 y en educación secundaria incluyen la redacción de un ensayo u otro tipo de texto, además de preguntas de selección múltiple e ítemes reforzados tecnológicamente.

Complementariamente, el estado de Virginia agregó pruebas adaptativas computarizadas a sus evaluaciones estatales en matemáticas para grado 6 durante el año escolar 2014-2015. El Departamento de Educación de Virginia planea introducir este tipo de pruebas para las evaluaciones de matemáticas y lectura en los grados 3 a 8 durante los próximos años.

Los estudiantes de educación primaria que fallan en los tests por un margen estrecho de puntaje o debido a situaciones especiales, pueden volver a rendir el tests durante el año en curso (excepto para estudiantes de grado 8 que rinden el test de Escritura).

El desempeño de los estudiantes es calificado en una escala de 0 a 600, considerándose 400 como el puntaje mínimo para considerar el desempeño del estudiante como competente y 500 como el puntaje mínimo para un desempeño avanzado (el establecimiento de estos puntajes se explica más adelante en esta ficha).

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

Los estándares de desempeño o "Descripciones de Niveles de Desempeño" (Performance Level Descriptors) asociados a los Tests de Estándares de Aprendizaje del estado de Virginia, son responsabilidad del Departamento de Educación y la junta de Educación, quienes trabajan con docentes de aula para el desarrollo de estas descripciones.

2.2 Características

Los estándares de desempeño de los Tests de Estándares de Aprendizaje describen los conocimientos y habilidades que se requieren para que el desempeño de un estudiante pueda ser clasificado en un determinado nivel y determinar si su desempeño cumple o no con las expectativas estatales.

Estos estándares de desempeño se describen para los tests de Lectura, Matemática, Ciencias Naturales, e Historia y Ciencias Sociales. En los tests para los grados 3 a 8 de Lectura y Matemática se definen cuatro niveles: Avanzado (aprueba), Competente (aprueba), Básico (reprueba) y Bajo Básico (reprueba). En todos los tests de Ciencias Naturales, Historia y Ciencias Sociales y de tests de los cursos de Inglés y Matemática (excepto Álgebra II) de educación secundaria, los estándares de desempeño describen tres niveles: Avanzado (aprueba), Competente (aprueba) y No cumple (reprueba). En Álgebra II se describen solo dos niveles: Avanzado (ruta hacia educación superior) y No cumple (reprueba).

A continuación, se presenta un ejemplo de estándares de desempeño para Lectura en el grado 3:

Virginia Standards of Learning Assessment
Grade 3 Reading Performance Level Descriptors

Fail/Below Basic	Fail/Basic	Pass/Proficient	Pass/Advanced
A student performing at this level should be able to: • Identify meaning of words when clearly evident in reading materials. • Locate information in fiction, poetry, and nonfiction texts to answer literal questions. • Identify word-reference sources.	A student performing at this level should be able to: • Use language structure or word relationships, such as common roots, affixes, synonyms and antonyms to determine meanings of words. • Demonstrate comprehension of fiction, poetry, and nonfiction texts by identifying explicitly stated main ideas, answering literal questions, and identifying author's purpose when explicitly stated. • Obtain information using word-reference sources.	A student performing at this level should be able to: • Use word-analysis and vocabulary acquisition skills when reading to derive meaning from unfamiliar words, including vocabulary from other content areas. • Demonstrate comprehension of fiction, poetry, and nonfiction texts by identifying main idea and supporting details, summarizing text and drawing conclusions, making predictions, and identifying author's purpose. • Interpret information from word-reference sources.	A student performing at this level should be able to: • Apply word-analysis and vocabulary acquisition skills, such as knowledge of word structure, homophones, roots, affixes, synonyms/antonyms, and context clues when reading. • Demonstrate comprehension of fiction, poetry, and nonfiction texts by identifying implied main ideas, summarizing text, drawing conclusions based on a passage as a whole, making predictions, and analyzing how vocabulary choice affects the author's purpose • Evaluate information from word-reference sources.

Las descripciones de los estándares de desempeño para otras áreas y grados pueden ser consultadas en:
http://www.doe.virginia.gov/testing/scoring/performance_level_descriptors/

Tal como se señaló anteriormente, en todas las pruebas se requiere un puntaje mínimo de 400 para que el desempeño del estudiante sea considerado como competente o que ha aprobado el test. Del mismo modo, en cualquier prueba se deben obtener 500 puntos para considerar el desempeño del estudiante como de nivel como Avanzado. Estos puntajes de corte se mantienen inalterados durante los años en los que se aplican los tests.

2.3 Historia

Desde el año 1998 los resultados de las evaluaciones estatales de Virginia son comunicadas de manera similar, con un foco en estándares de desempeño que indican qué proporción de estudiantes cumple o no con las expectativas estatales (pass/fail).

De 1998 a 2005, los resultados de las pruebas estatales se informaban para las asignaturas de Lenguaje, Matemática, Historia/Ciencias Sociales y Ciencias para los grados 3, 5 y 8 de manera separada. A partir de 2006, los resultados de las pruebas estatales son reportados como resultados combinados - de todas las evaluaciones en los grados 3, 4, 5, 6, 7 y 8 y asignaturas.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

Las descripciones de los estándares de desempeño y sus respectivos puntajes de corte son desarrolladas por el Departamento y la Junta de Educación de Virginia en colaboración con docentes y otros especialistas.

Para el establecimiento de los puntajes de corte se conformaron comités. La conformación de estos comités estuvo a cargo del Departamento de Educación de Virginia, el cual estableció los siguientes criterios para seleccionar a los miembros de los comités:

- Capacitación en temas de enseñanza y experiencia significativa en la asignatura.
- Conocimiento profundo de los Estándares de Aprendizaje de Virginia.
- Experiencia con estudiantes con discapacidad o necesidades educativas especiales.
- Representación geográfica balanceada.

Adicionalmente, se reclutaron académicos de instituciones de educación superior para aquellos tests de educación secundaria que determinan el ingreso a este nivel educacional (Lectura, Escritura y Álgebra II).

Puntajes de corte independientes son fijados por el Departamento de Educación para estudiantes con necesidades educativas especiales permanentes en las pruebas de grado 8 de Lectura y Matemáticas.

3.2 Metodología

Las descripciones cualitativas de los estándares de desempeño son desarrolladas por comités de docentes y académicos. Estas descripciones son revisadas en ciclos de 7 años, coincidiendo con ciclos de revisión de siete años de los Estándares de Aprendizaje.

En el caso de los puntajes de corte, estos se establecieron a través de una metodología predefinida basada en el método *Angoff*, incluyendo la variación conocida como el "Método Sí/No" (*Yes/No Method*).

Los panelistas de los comités[29] comenzaron consensuando una descripción del desempeño de un estudiante que "justo alcanza" un determinado nivel de desempeño. Luego, para los ítemes de selección múltiple, los panelistas revisaron cada uno de ellos y evaluaron si un estudiante que cumple con la descripción del desempeño "que justo alcanza" para el nivel Avanzado y el Competente sería capaz de responder correctamente (al menos 2 de 3 veces). Cada ítem es puntuado con 1 ó 0 dependiendo si la estimación del panelista fue que el estudiante sí lo respondería correctamente (1) o no (0); a continuación, se suman las puntaciones de cada panelista.

Para las preguntas abiertas de escritura se empleó otra variación de Angoff conocida como "Puntuación de Tarea Esperada" (*Expected Task Score*). Según este método, los panelistas evalúan si los estudiantes que "justo alcanza" un nivel podría (2 de 3 veces) alcanzar los distintos puntajes definidos en la rúbrica (1, 2, 3 ó 4) para cada uno de sus dominios.

Tanto para los ítemes de selección múltiple como para las preguntas abiertas, la recomendación final del punto de corte se obtiene después de tres rondas de discusión donde se agregan los puntajes de todos los panelistas.

Luego de obtener la recomendación proveniente de la tercera ronda de discusión, un grupo pequeño de panelistas la revisa para todos los grados dentro de una misma asignatura a la luz de los resultados esperados (distribución porcentual de estudiantes por niveles de desempeño), pudiendo modificar la recomendación inicial. La recomendación acordada por este grupo pequeño de panelistas es enviada a la Junta de Educación para su aprobación.

Los puntajes de corte se expresan como puntaje bruto, es decir cantidad de respuestas

[29] En los informes técnicos de las evaluaciones de este estado no se especifica más información que la descrita en esta ficha.

correctas que se requiere para que el desempeño de un estudiante sea clasificado en un determinado nivel. Tal como se mencionó anteriormente en esta ficha, este puntaje es luego transformado a un puntaje estandarizado, siendo 400 el corte para el nivel Competente y 500 para el Avanzado. La siguiente tabla muestra los puntajes de corte brutos aprobados para todos los Tests de Estándares de Aprendizaje.

Grades 3 through 8 SOL Tests	Fail/Basic	Pass/Proficient	Pass/Advanced	Adoption Date
Grade 3 Mathematics* (2009 Standards)	Administered as a computer adaptive test as of Spring 2016.**			
Grade 4 Mathematics* (2009 Standards)	17 out of 50 items	31 out of 50 items	45 out of 50 items	Mar 22, 2012
Grade 5 Mathematics* (2009 Standards)	18 out of 50 items	31 out of 50 items	45 out of 50 items	Mar 22, 2012
Grade 6 Mathematics* (2009 Standards)	Administered as a computer adaptive test as of Fall 2014.**			
Grade 7 Mathematics* (2009 Standards)	Administered as a computer adaptive test as of Fall 2015.**			
Grade 8 Mathematics* (2009 Standards)	Administered as a computer adaptive test as of Fall 2015.**			
Grade 3 Reading (2010 Standards)	13 out of 40 items	25 out of 40 items	35 out of 40 items	Mar 28, 2013
Grade 4 Reading (2010 Standards)	12 out of 40 items	25 out of 40 items	35 out of 40 items	Mar 28, 2013
Grade 5 Reading (2010 Standards)	11 out of 40 items	25 out of 40 items	35 out of 40 items	Mar 28, 2013
Grade 6 Reading (2010 Standards)	14 out of 45 items	28 out of 45 items	40 out of 45 items	Mar 28, 2013
Grade 7 Reading (2010 Standards)	14 out of 45 items	28 out of 45 items	40 out of 45 items	Mar 28, 2013
Grade 8 Reading (2010 Standards)	14 out of 45 items	28 out of 45 items	40 out of 45 items	Mar 28, 2013
Grade 5 Science (2010 Standards)	NA	24 out of 40 items	35 out of 40 items	April 25, 2013
Grade 8 Science (2010 Standards)	NA	27 out of 50 items	45 out of 50 items	April 25, 2013
Grade 8 Writing (2010 Standards)	NA	31 out of 48 items	41 out of 48 items	April 25, 2013
Content Specific History SOL Tests	**Fail/Basic**	**Pass/Proficient**	**Pass/Advanced**	**Adoption Date**
Virginia Studies (2008 Standards)	NA	21 out of 40 items	32 out of 40 items	Mar 24, 2011
Civics and Economics (2008 Standards)	NA	21 out of 40 items	34 out of 40 items	Mar 24, 2011
End-of-Course SOL Tests	**Fail/Basic**	**Pass/Proficient**	**Pass/Advanced**	**Adoption Date**
English: Reading (2002 Standards)	NA	28 out of 50 items	42 out of 50 items	Nov 30, 2005
English: Writing (2002 Standards)	NA	37 out of 54 items	49 out of 54 items	October 1998
Earth Science (2003 Standards)	NA	30 out of 50 items	45 out of 50 items	October 1998
Earth Science (2010 Standards)	NA	25 out of 50 items	45 out of 50 items	Jan 10, 2013
Biology (2003 Standards)	NA	26 out of 50 items	45 out of 50 items	October 1998
Biology (2010 Standards)	NA	27 out of 50 items	45 out of 50 items	Jan 10, 2013
Chemistry (2003 Standards)	NA	27 out of 50 items	45 out of 50 items	October 1998
Chemistry (2010 Standards)	NA	25 out of 50 items	44 out of 50 items	Jan 10, 2013
World History (I) to 1500 A.D. (2001 Standards)	NA	30 out of 60 items	50 out of 60 items	Nov 19, 2003
World History (I) to 1500 A.D. (2008 Standards)	NA	31 out of 60 items	53 out of 60 items	Jan 13, 2011
World History (II) from 1500 A.D. to the Present (2001 Standards)	NA	30 out of 60 items	50 out of 60 items	Nov 19, 2003
World History (II) from 1500 A.D. to the Present (2008 Standards)	NA	31 out of 60 items	52 out of 60 items	Jan 13, 2011
Virginia & U.S. History (2001 Standards)	NA	30 out of 60 items	51 out of 60 items	Nov 19, 2003
Virginia & U.S. History (2008 Standards)	NA	30 out of 60 items	53 out of 60 items	Jan 13, 2011
World Geography (2001 Standards)	NA	33 out of 60 items	50 out of 60 items	Nov 19, 2003
World Geography (2008 Standards)	NA	33 out of 60 items	54 out of 60 items	Jan 13, 2011
Algebra I* (2001 Standards)	NA	27 out of 50 items	45 out of 50 items	October 1998
Algebra I* (2009 Standards)	NA	25 out of 50 items	45 out of 50 items	Jan 12, 2012
Geometry (2001 Standards)	NA	27 out of 45 items	41 out of 45 items	October 1998
Geometry (2009 Standards)	NA	25 out of 50 items	44 out of 50 items	Jan 12, 2012
Algebra II (2001 Standards Revised)	NA	30 out of 50 items	45 out of 50 items	Nov 30, 2005
	Fail/Basic	**Pass/Proficient**	**Advanced/College Path**	**Adoption Date**
End-of-Course Reading (2010 Standards)	NA	31 out of 55 items	49 out of 55 items	Jan 10, 2013
End-of-Course Writing (2010 Standards)	NA	33 out of 54 items	46 out of 54 items	April 25, 2013
Algebra II (2009 Standards)	NA	27 out of 50 items	43 out of 50 items	Jan 12, 2012

* Includes the Plain English version of the test.

** A computer adaptive test is an online test that is customized for each student taking the test. A student's score is determined by the number of questions a student answe correctly and the relative difficulty of the questions answered. For more information about computer adaptive tests, go to:
http://www.doe.virginia.gov/testing/test_administration/cat/index.shtml

Los puntajes de corte, al igual que las descripciones de los estándares de desempeño, se ciñen a un ciclo de revisión de 7 año, la cual coincide con los ciclos de revisión de los Estándares de Aprendizaje (estándares de contenido). La siguiente tabla ilustra cuándo se fijó el punto de corte actualmente vigente (Puntaje de corte previo) y cuándo corresponde revisarlo (Puntaje de corte próximo):

Asignatura	Puntaje de corte previo	Puntaje de corte próximo
Historia	2010–2011	2017–2018
Matemática	2011–2012	2018–2019
Lectura y Escritura	2012–2013	2019–2020
Ciencias Naturales	2012–2013	2019–2020

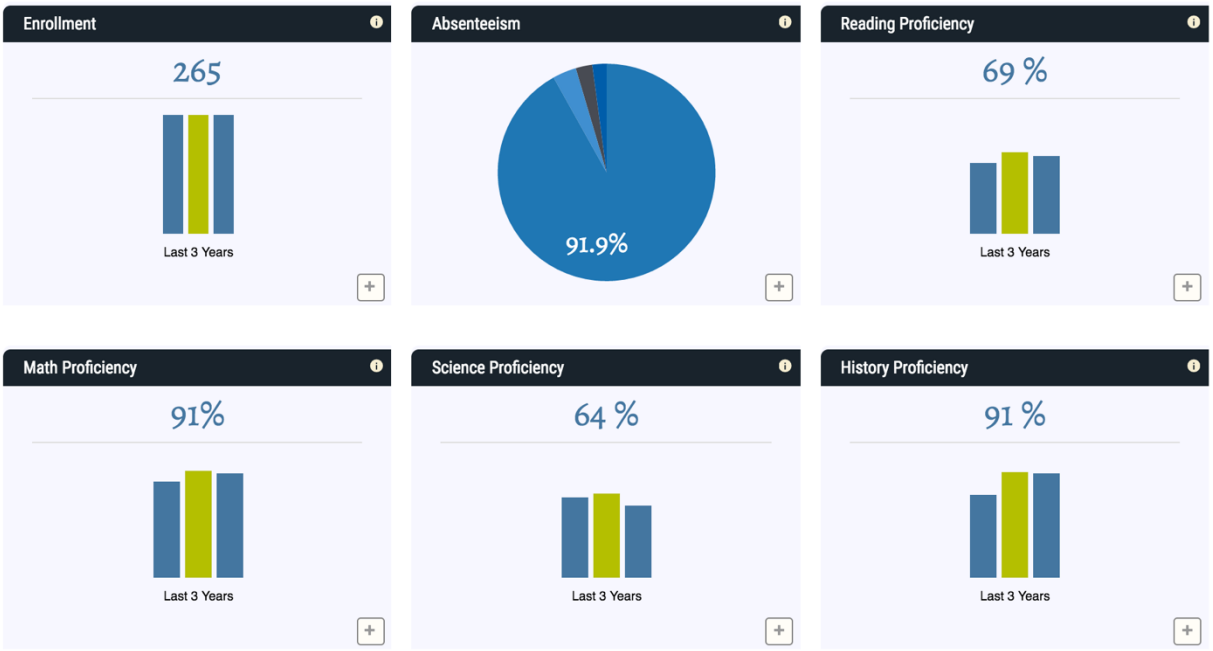
Por ejemplo, en el año 2016 se promulgaron nuevos Estándares de Aprendizaje (estándares de contenido) para Matemática. Estos son comunicados ese mismo año junto con indicaciones para vincular los estándares de contenido previos (promulgados en 2009) con los nuevos. Se espera que estos nuevos Estándares de Aprendizaje sean implementados en las salas de clase entre los años 2017 y 2018. Las evaluaciones estatales aplicadas en esos años estarán referidas a los estándares de contenido promulgados en 2009. Desde fines del año 2018 las evaluaciones estatales y sus respectivos estándares de desempeño deben estar completamente alineados a los Estándares de Aprendizaje promulgados en 2016[30].

[30] El Departamento de Educación de Virginia no ha comunicado, hasta el momento, indicaciones para la interpretación de resultados de acuerdo a la fijación de nuevos puntajes de corte.

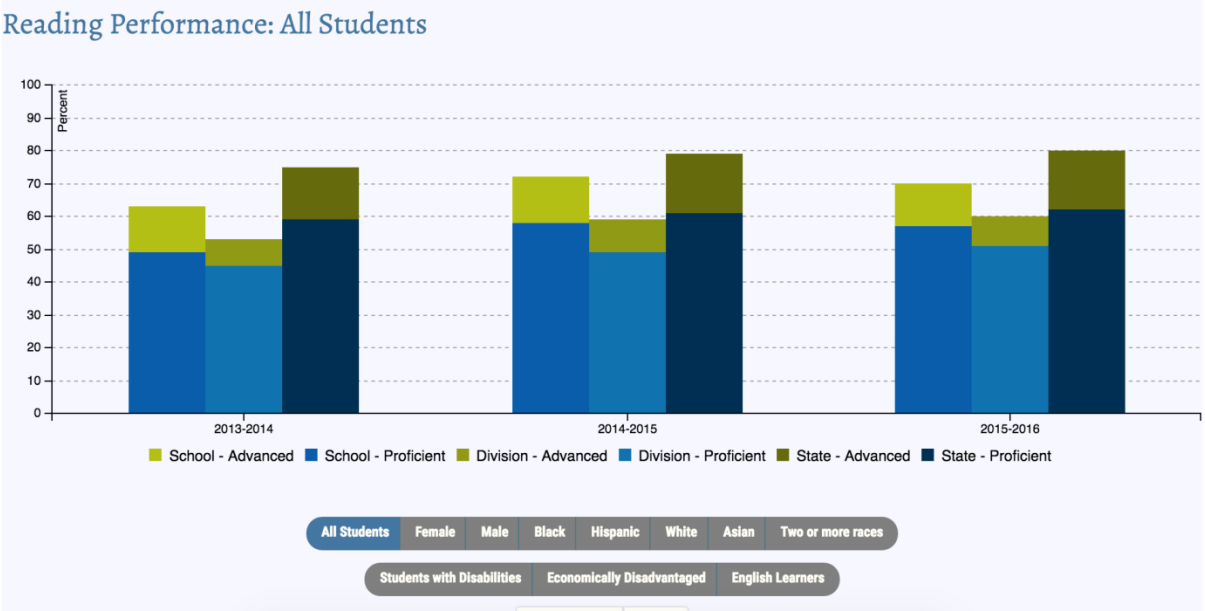
4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

Los resultados referidos a los niveles de desempeño por establecimiento escolar, condado y a nivel estatal pueden ser consultados a través de una plataforma interactiva disponible en: http://schoolquality.virginia.gov/

Para cada establecimiento se entrega un "snapshot" que sintetiza el desempeño en distintos ámbitos, incluyendo el porcentaje de estudiantes que alcanza o supera el nivel Competente:



También se entregan datos desagregados para cada prueba, comparando los datos del establecimiento con los de la división territorial y con los del estado, y pudiendo desagregar resultados para distintos subgrupos:



Overall Student Performance: Reading Performance	2013-2014				2014-2015				2015-2016			
Student Subgroup	Advanced	Passed	Proficient	Failed	Advanced	Passed	Proficient	Failed	Advanced	Passed	Proficient	Failed
All Students	14	63	49	38	14	72	58	28	13	69	57	31
Female	15	65	50	35	16	76	60	24	20	77	57	23
Male	13	61	48	39	13	69	56	31	6	62	56	38
Asian	<	100	<	0					<	<	<	<
Black	13	60	47	40	14	71	57	29	12	69	57	31
White	18	82	64	18	<	<	<	<	20	70	50	30
Students with Disabilities	16	53	37	47	13	56	44	44	-	86	86	14
Economically Disadvantaged	8	58	50	42	13	71	58	29	8	66	58	34
English Learners	<	100	<	0					<	<	<	<
LEGEND	< = A group below state definition for personally identifiable results - = No data for group * = Data not yet available Unduplicated = Students are able to be in two gap groups											

Los resultados por estudiante son entregados solo a sus padres, madres y establecimientos escolares.

5. USO ESTÁNDARES DE DESEMPEÑO

Los estándares de desempeño de las pruebas del estado de Virginia permiten monitorear el progreso de los estudiantes y las necesidades de los establecimientos escolares. También permiten entregar información a la comunidad sobre el cumplimiento de las expectativas de aprendizaje.

Del mismo modo, son empleados como medida de reponsabilización por resultados en el contexto de la política "Ningún Niño se Queda Atrás" (*No Child Left Behind*) y como indicador para la obtención del diploma de educación secundaria y para el ingreso a estudios superiores.

6. OTRA INFORMACIÓN RELEVANTE

Resulta interesante destacar que en el ámbito de la evaluación, además del desarrollo de pruebas y estándares de desempeño, el Departamento de Educación del Virginia contrató a la *Virginia Commonwealth University* para conducir una revisión externa del alineamiento entre las evaluaciones y los Estándares de Aprendizaje (estándares de contenido en este caso). Esta revisión se realizó utilizando los procedimientos desarrollados por Norman Webb (Método Webb), enfocándose en cuatro criterios de alineamiento: (i) concurrencia de categorías, (ii) consistencia en la profundidad del conocimiento, (iii) correspondencia del rango de conocimiento y (iv) balance de la representación.

Los resultados de este estudio indican que las pruebas están bien alineadas a los Estándares de Aprendizaje (estándares de contenido). Algunas discrepancias detectadas se relacionan con ítemes que requirieron un menor nivel de desarrollo del conocimiento que lo esperado en los Estándares de Aprendizaje.

El estudio permite concluir que las pruebas aplicadas por el Departamento de Educación son un buen reflejo de las expectativas establecidas en los Estándares de Aprendizaje.

7. REFERENCIAS:

Virginia Department of Education (2015). *Virginia Standards of Learning Assessments Technical Report 2014–2015 Administration Cycle*. Consultado en diciembre de

2016 en:
http://www.doe.virginia.gov/testing/test_administration/technical_reports/sol_technical_report_2014-15_administration_cycle.pdf

Virginia Department of Education (2015). *Frequently Asked Questions about SOL Testing*. Consultado en diciembre 2016 en:
http://www.doe.virginia.gov/testing/sol_faq.pdf

http://www.doe.virginia.gov/

http://schoolquality.virginia.gov/

8. CONTACTO:

No se contactó a ningún especialista para la elaboración de esta ficha.

FICHA #9
PLAN NACIONAL PARA LA EVALUACIÓN DE LOS APRENDIZAJES (PLANEA), MÉXICO.

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la Evaluación

Los exámenes de Plan Nacional para la Evaluación de los Aprendizajes (Planea) son pruebas llevadas a cabo 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 currículo nacional. También tiene la misión de identificar los factores asociados a las diferencias entre los niveles de logro. Se realiza en una muestra nacional de escuelas – con representatividad de cada uno de los Estados Mexicanos – cada cuatro años.

1.2 Referente orientador de las evaluaciones

En términos generales, el referente orientador de Planea son los currículos nacionales de las materias escolares evaluadas; en México no se considera un solo documento como la especificación única del currículo 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 Subsecretaría de Desarrollo Curricular de la Subsecretaría de Educación Básica de la Secretaria de Educación Pública, o de las Direcciones Generales correspondientes a los distintos planes de estudio en la Subsecretaría de Educación Media.

1.3 Organismo responsable del programa de evaluación y del currículo.

Como se menciona anteriormente, el currículo nacional es responsabilidad de distintas direcciones generales en las subsecretarías de educación básica y media en la Secretaría de Educación Pública federal de México. Los Planea son 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 gobernada 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.

1.4 Áreas disciplinares

Las áreas disciplinares que evaluá Planea son matemáticas, español, ciencias naturales, ciencias sociales y humanidades.

1.5 Grados evaluados.

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

Tabla 4 Niveles, grados y áreas disciplinares de los Planea.

Nivel	Grado	Áreas Disciplinares
Preescolar	3º[31]	Razonamiento Numérico
		Razonamiento Verbal
Primaria	3º	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales
	6º	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales
Secundaria	3º[32]	Matemáticas
		Español
		Ciencias Naturales
		Ciencias Sociales

1.6 Características de las pruebas.

Los Planea, a juicio del INEE, tienen tres características distintivas: son referidos a criterios o criteriales, están alineados al currículo y son matriciales.

Son criteriales porque se diseñan para evaluar el dominio que alcanzan los estudiantes del sistema educativo de una disciplina en particular. Su enfoque principal es identificar el nivel de logro que alcanzan los estudiantes como resultado de su escolarización formal. Cada área curricular se mide a profundidad, y se hace un esfuerzo para incluir todos los conocimientos y habilidades de importancia para la disciplina y grado escolar.

[31] 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.

[32] 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.

Están alineados al currículo porque su propósito es evaluar los objetivos del currículo intencional de México, principalmente los planes y programas de estudio nacionales.

Planea tiene el objetivo de evaluar todos los contenidos curriculares importantes. Por tanto, cuenta con un diseño matricial en el cual los ítemes que conforman una prueba se agrupan en bloques para ser distribuidos entre los alumnos; no todos contestan las mismas preguntas, pero con las repuestas de todos se obtienen resultados del examen en su conjunto. En los Planea, la calificación individual del estudiante no se reporta, sino que se reportan los resultados agregados a nivel de entidad federativa y modalidad educativa, dado que lo que se busca es evaluar al sistema educativo en su conjunto.

Las pruebas están compuestas principalmente de ítemes de opción múltiple. Se utiliza equiparación (*equating*) y calibración con Teoría de Respuesta al Ítem (TRI) para garantizar la comparabilidad interanual, y en el cálculo de puntajes de corte para los estándares de desempeño (niveles de logro). La media se fija en 500 unidades y la desviación estándar en 100 unidades.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO.

2.1 Organismo a cargo

El INEE está a cargo de la definición de los estándares de desempeño, y trabaja en forma colegiada con psicometristas y otros expertos miembros del personal del Instituto, y con docentes de aula, autores de libros de texto, y otros expertos externos tanto en el Comité de Niveles de Logro, como en el Comité de identificación de puntajes de corte.

2.2 Características

Los estándares de desempeño en Planea, se llaman niveles de logro y son cuatro. Los niveles de desempeño y sus definiciones se presentan en la Tabla 2.

Tabla 5: 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 currículo.
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 currículo.

Las descripciones de los niveles de logro son el resultado de trabajo sucesivos de comités colegiados encargados de determinar etiquetas y establecer definiciones e identificar ítemes "ilustrativos" correspondientes a los niveles. El trabajo se realiza comenzando por un comité colegiado que revisa únicamente documentos curriculares, sin referirse a evidencia de las pruebas. Otros comités se encargan de la determinación de puntos de cortes, siguiendo una adaptación del método Bookmark[33] en donde el trabajo se realiza en base a evidencias de las pruebas

2.3 Historia

No ha habido cambios en los estándares de desempeño en Planea. 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 (que es una de las pruebas ELSEN que se van liberando) que es co-aplicada por el INEE y la Secretaria de Educación Pública[34]. Los procedimientos de establecimiento de estándares se siguieron de igual forma en cada prueba, y los procedimientos de equiparación y calibración se siguen para asegurar comparabilidad interanual.

3. DESARROLLO DE LOS ESTÁNDARES

3.1 Instituciones

El INEE es la institución encargada.

3.2 Metodología

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. Los puntajes

[33] El procedimiento especifico se describe en la sección 3.2 de esta ficha, y se resume en la Ilustración 1.

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

correspondientes a los puntos de corte se establecen después de aproximadamente tres rondas de consulta en un comité de jueces expertos y se fijan en la mediana del puntaje TRI correspondiente al nivel de habilidad de los ítemes identificados como "marcadores", es decir, en los puntos de inflexión de los niveles de logro.

A continuación, se decribe con mayor detalle el procedimiento para establecer los puntajes corto el método es el siguiente:

El Comité de Puntuaciones de Corte (CPC)[35] recibe entrenamiento sobre 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 luego en forma conjunta para lograr consenso o congruencia entre jueces con respecto a los ítemes que marcan los puntos de inflexión entre niveles. Estos ítemes – llamados ítemes marcadores – se identifican comenzando por las 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.

[35] La documentación técnica disponible al público con respecto a estos procesos es extensa, sin embargo, no hay información acerca de los términos de referencia de los miembros del Comité, del número de integrantes, o acerca del periodo de tiempo que toma su trabajo. Estas preguntas no fueron respondidas en las entrevistas llevadas a cabo con los expertos del INEE, sí se nos indicó que había considerable variación en estos aspectos en cada asignatura, grado y en distintos año, sin especifiar la naturaleza de esas variaciones.

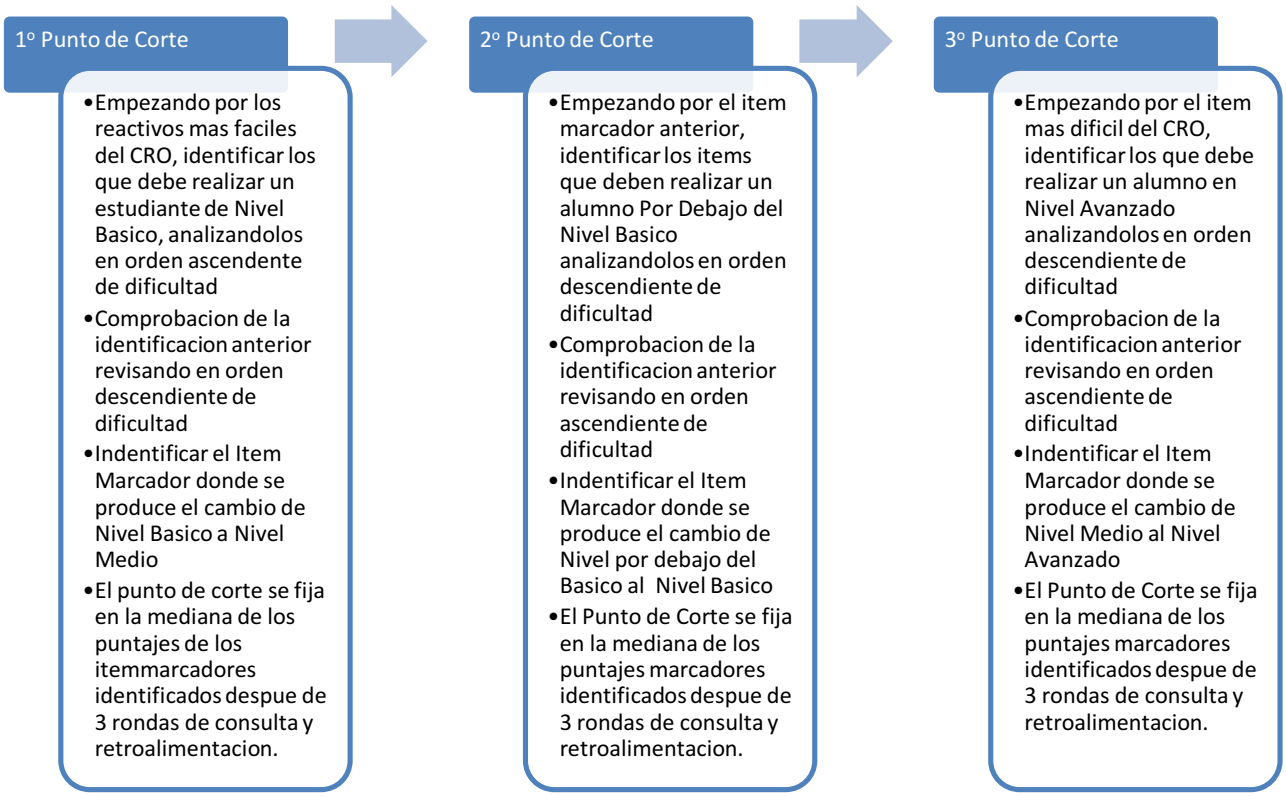


Ilustración 2: Esquema del proceso de establecimiento de puntos de corte Planea

4. Comunicación de Estándares de Desempeño

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"[36]. Los informes del INEE y su sitio de internet difunden todo el material 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 el informe nacional anual, denominado "Panorama Educativo de México".
Los Planea no tienen consecuencias para escuelas, estudiantes o docentes; tampoco se consideran en la evaluación de docentes, que tiene sus propios instrumentos. Tampoco tienen consecuencias para los estudiantes. La intención es que las escuelas y los docentes retomen la información de la evaluación diagnóstica para elaborar rutas de mejora de la escuela y para los docentes.

[36] Artículo 12, Inciso IV

5. USO ESTÁNDARES DE DESEMPEÑO

Los estándares de desempeño son herramientas de evaluación. No son elementos propios del currículo nacional, sino que son definiciones operacionales de esos objetivos realizados por comités técnicos colegiados con fines de traducir el currículo nación en términos mensurables. Se usan, en la actualidad, únicamente en la evaluación y en los informes y bases de datos de la evaluación.

Hasta la fecha, en Mexico, los estándares no se promulgan o promueven en forma independiente como parte de la política curricular. El currículo nacional, compuesto por planes y programas de estudio, libros de texto, y otros instrumentos constituyen el curriculom oficial. Los estándares de desempeño se consideran herramientas técnicas para la operacionalización del currículo nacional para propósitos de llevar a cabo evaluaciones.

6. OTRA INFORMACION RELEVANTE

Los cambios en evaluación que comienzan con la reforma del año 2013 son importantes en cuanto establecen el nuevo marco legal e institucional del INEE, elimina las pruebas desarrolladas por la Secretaría de Educación Pública, y establece dos tipos de pruebas Planea: la evaluación Planea-ELSEN totalmente a cargo del INEE, y la evaluación censal Planea que es desarrollada por el INEE, pero co-aplicada por el INEE y la Secretaría de Educación Pública.
Hasta la fecha los cambios son únicamente logísticos y administrativos, no ha habido cambios en los parámetros técnicos de las pruebas, en los estándares o en la metodología para analizar y reportar resultados.

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 Martinez 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

Ambos profesionales fueron contactados vía email.

FICHA #10
PERÚ: EVALUACIÓN CENSAL DE ESTUDIANTES (ECE)

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la Evaluación

En el año 2006, el Ministerio de Educación de Perú (Minedu) tomó la decisión de llevar a cabo una evaluación de carácter censal a los estudiantes de primaria de todo el país. Ésta es conocida como "Evaluación Censal de Estudiantes" (ECE) y se realiza con el propósito de monitorear el desarrollo de las habilidades fundamentales de los estudiantes para que continúen aprendiendo a lo largo del ciclo escolar. En primaria se evalúa el aprendizaje de la lectoescritura y el dominio básico de algunos conceptos matemáticos fundamentales, como la estructura aditiva y la comprensión del sistema de numeración decimal; en secundaria se evalúa la lectura, escritura y habilidades matemáticas.
Los instrumentos de la ECE son construidos por los especialistas de Matemática, Comunicación y Educación Intercultural Bilingüe del equipo de evaluación de la Oficina de Medición de la Calidad de los Aprendizajes (UMC). Estos instrumentos son pruebas de ítemes de opción múltiple y de respuesta construida o de ensayo (estos últimos solo para el segundo grado de secundaria).

1.2 Referente orientador de las evaluaciones.

El Curriculum Nacional de Educación Básica fue recientemente ajustado (2015) y se estructura en torno a competencias acompañadas de estándares de aprendizaje que definen el nivel que se espera puedan alcanzar todos los estudiantes al finalizar los ciclos de la Educación Básica. Los estándares de aprendizaje del curriculum tienen por propósito ser los referentes para la evaluación de los aprendizajes tanto a nivel de aula como a nivel de sistema (evaluaciones nacionales, muestrales o censales). No obstante, las evaluaciones censales evalúan algunos desempeños de las competencias, pero no pueden ni pretenden dar cuenta de toda la competencia.
Los estándares de aprendizaje descritos en el curriculum nacional son descripciones del desarrollo de la competencia en niveles de creciente complejidad, desde el inicio hasta el fin de la Educación Básica, de acuerdo a la secuencia que sigue la mayoría de estudiantes que progresan en una competencia determinada. Estos estándares se describen a través de ocho niveles transversales a la Educación Primaria para diferentes ejes de las asignaturas que componen el currículo nacional. Dado que este modo de organizar el curriculum nacional es relativamente reciente, no se dispone de información detallada acerca de cómo se vinculan los estándares de aprendizaje curriculares con los estándares de desempeño de las ECE (estos últimos se describen más adelante).

Las pruebas están referidas a tablas de especificaciones basadas en el curriculum nacional. Después de que los equipos de Matemática, de Comunicación y de Programa de Educación Intercultural Bilingüe han construido los ítemes, estos son revisados por los expertos de la UMC (quienes evalúan aspectos como la calidad, vigencia y veracidad de la información según cada disciplina científica, la correspondencia con la tabla de especificaciones, la adecuación de la complejidad del ítem a la población evaluada, y la construcción del enunciado y las alternativas, tanto en lo formal como en su eficacia para la medición del constructo).

1.3 Organismo responsable del programa de evaluación y del currículo.

Las ECE están a cargo de la Oficina de Medición de la Calidad de los Aprendizajes (UMC), dependiente del Ministerio de Educación de Perú.
Por su parte, el currículo nacional de Perú está a cargo directamente del Ministerio de Educación. De modo más específico, los estándares de aprendizaje del currículo fueron elaborados inicialmente por el Instituto Peruano de Evaluación, Acreditación y Certificación de la Educación Básica (IPEBA) en coordinación con el Ministerio de Educación. El equipo de la UMC ha colaborado en las áreas que regularmente se evalúan a nivel de sistema.

1.4 Áreas disciplinares

De manera censal se evalúan las áreas de Matemática, Lectura y Escritura.

1.5 Grados evaluados

La ECE evalúa a los estudiantes de segundo grado de primaria y, en caso de que en las instituciones educativas se aplique el Programa de Educación Intercultural Bilingüe, evalúa a los estudiantes de cuarto grado de primaria. A partir del 2015 se inicia la evaluación censal en segundo grado de secundaria.

1.6 Características de las pruebas

Los instrumentos de las ECE son pruebas conformadas por ítemes de opción múltiple y de respuesta construida (estos últimos solo para el segundo grado de secundaria).

Según se declara en documentos técnicos de la ECE 2015, "la construcción de los ítemes y de las pruebas sigue los principios de validez, confiabilidad y diseño universal de evaluación, que establecen que los instrumentos de evaluación deben recoger información de los estudiantes de tal manera que se pueda estimar de forma fiable su nivel de aprendizaje y que dicha información pueda ser usada para los fines propios del proceso educativo (AERA, APA y NCME, 2014). Asimismo, debe reflejar una concepción inclusiva de la educación, conforme a los lineamientos de la política educativa nacional" (p.6).

El análisis psicométrico de las pruebas aplicadas en la ECE se basa en el modelo Rasch para ítemes dicotómicos.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

Los estándares de desempeño se describen a partir de los resultados de las pruebas ECE y, por tanto, están a cargo de la misma institución que las administra: la Oficina de Medición de la Calidad de los Aprendizajes (UMC).

2.2 Características de los estándares

La interpretación de los resultados de la ECE está referida a niveles de desempeño (o de logro; se les llama de ambos modos en los documentos oficiales) que se describen para Matemática, Lectura y Escritura.

De acuerdo con su puntaje individual, el desempeño de los estudiantes de secundaria puede ser ubicados en cuatro niveles de logro: *Satisfactorio*, *En proceso*, *En inicio* y *Previo al inicio*. Por su parte, el desempeño de los estudiantes de primaria puede ser ubicado en tres niveles de logro: *Satisfactorio*, *En proceso* y *En inicio*.

Cada uno de estos niveles describe lo que sabe y puede hacer un estudiante cuyo puntaje está dentro de un determinado rango de habilidad. Los niveles de logro son inclusivos, esto significa, por ejemplo, que los estudiantes ubicados en el nivel Satisfactorio tienen alta probabilidad de responder adecuadamente las preguntas del nivel Satisfactorio y las preguntas de los niveles En proceso y En inicio.

A continuación, se presenta una descripción genérica de los niveles de desempeño de secundaria (los puntajes de corte corresponden a los de la prueba de Lectura):

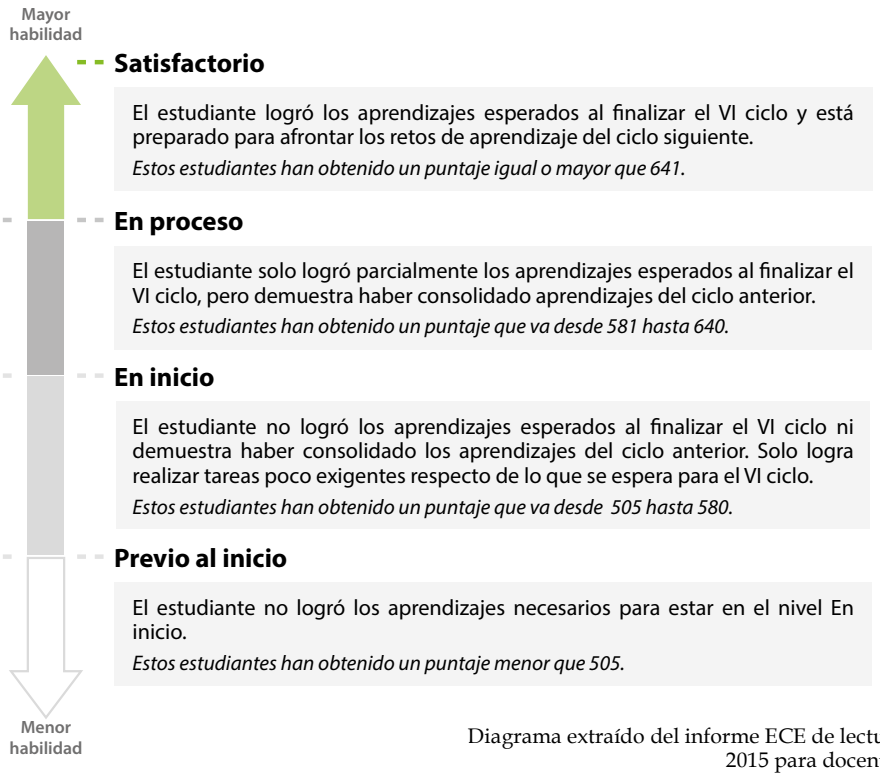


Diagrama extraído del informe ECE de lectura 2015 para docentes

En un documento especialmente dirigido a docentes para que reflexionen sobre los resultados obtenidos por sus estudiantes en la ECE, se entregan descripciones más detalladas sobre los niveles de logros, acompañadas por ejemplos de preguntas que ilustran el desempeño en cada nivel. A continuación, las descripciones detallas y ejemplificadas para segundo grado de primaria:

Nivel Satisfactorio — Logró los aprendizajes esperados

Este nivel agrupa a los estudiantes que demostraron en la ECE 2015 un manejo adecuado de las capacidades evaluadas, según lo esperado para el grado. Sobre esto, resulta necesario mencionar que el nivel de estos estudiantes no es avanzado o destacado: su nivel de logro es adecuado para el grado. Adicionalmente, conviene recordar que los niveles de logro son inclusivos. Esto significa que los niños y niñas que han alcanzado el nivel Satisfactorio pueden realizar las tareas propias de este nivel y las del nivel En proceso.

En Lectura, los estudiantes de este nivel pueden ubicar información que no se encuentra tan fácilmente en el texto. Además, pueden deducir ideas que les permiten comprender algunas partes específicas del texto, así como entenderlo en su conjunto. Asimismo, reflexionan sobre el contenido, para aplicarlo a situaciones externas al texto, y sobre la forma del texto, para lo cual se apoyan en su conocimiento cotidiano. Estas tareas las realizan en diversos tipos de textos, de estructura simple, de extensión media y complejidad adecuada para el grado.

A continuación, se presentan dos textos y preguntas representativas que logran resolver los estudiantes de este nivel.

Lee este texto.

Un monito muy especial

El tití pigmeo es famoso por ser el mono más pequeño del mundo. Por eso, también es conocido como "mono de bolsillo". Es tan pequeño que cabe en la mano de una persona.

El tití pigmeo vive en la selva del Perú. Hace sus nidos en las partes más altas de los árboles. Allí, entre las hojas de los árboles, se protege de águilas, halcones y otros animales que se lo pueden comer.

Su cuerpo está cubierto con pelos suaves y esponjosos. Además, tiene una cola delgada y larga.

Sus dedos son delgados y terminan en garras muy pequeñas con las que trepa hasta lo alto de los árboles. De esa manera, alcanza las hojas más tiernas, que son sus favoritas. También se alimenta de insectos, frutas y de la savia, un líquido que se encuentra dentro de las plantas.

Este animalito se encuentra en peligro de desaparecer. Muchas personas están cortando los árboles de la selva. Pronto, el tití pigmeo no tendrá dónde vivir.

- Según el texto, ¿por qué podría desaparecer el tití pigmeo?

  Deduce relaciones de causa o finalidad.

- ¿Para qué se escribió este texto?

  Deduce el propósito de un texto.

- ¿De qué trata principalmente el texto?

  Deduce el tema central del texto.

Lee este cuento.

En un establo, vivían un burrito trabajador llamado Sancho y un caballo de paso llamado Sipán. El caballo Sipán bailaba muy bien la marinera junto con los niños Joaquín y Micaela. Se iban a presentar al Concurso Nacional de Marinera.

El burrito Sancho también quería bailar marinera, pero creía que nadie le iba a hacer caso. ¡Los burros no bailan marinera!

Una mañana, Sancho oyó música y se puso a bailar. Joaquín lo vio y se burló:
—¡Miren a Sancho! ¡Cree que puede bailar marinera!

Pero en vez de desanimarse por las burlas, el burrito no se dio por vencido. Siguió bailando todos los días hasta hacerlo mejor.

Un día antes del concurso, Sipán empezó a quejarse. Se había lastimado la pata practicando y no podría participar en el concurso. Los niños, preocupados, lo atendieron con cariño.

Al ver que el caballo sentía mucho dolor, Sancho quiso alegrarlo bailando marinera. ¡Había practicado muchísimo y se sabía todos los pasos! Los niños se quedaron asombrados.

El día del concurso, Joaquín y Micaela se presentaron con Sancho. El burrito bailaba tan bonito que todos aplaudieron y exclamaron:
—¡Viva el burrito que sabe bailar marinera!

Y así, Sancho, muy feliz, vio que su esfuerzo valió la pena.

- El cuento dice que Sancho "no se dio por vencido". ¿Qué significa esto?

Deduce el significado de palabras o expresiones usando información del texto.

- Según el cuento, ¿cómo era Sancho?

Deduce cualidades o defectos de los personajes de un texto.

- En el cuento, ¿en qué se parecen Sancho y Sipán?

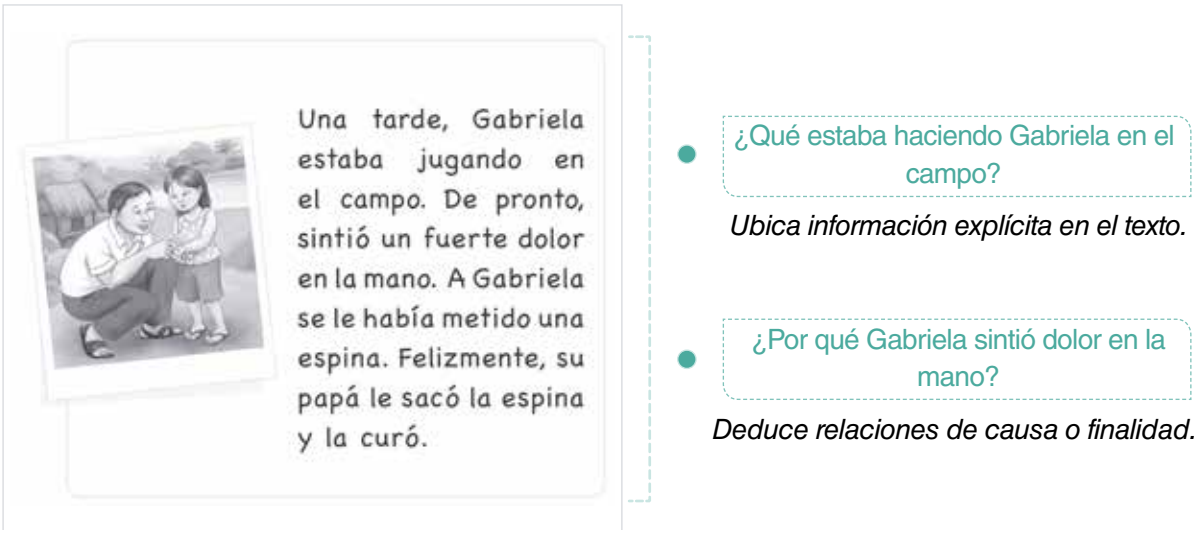
Establece semejanzas o diferencias entre diferentes elementos del texto.

Nivel En proceso — No logró los aprendizajes esperados

Este nivel agrupa a un conjunto de estudiantes cuyos logros, si bien no les permiten alcanzar el nivel Satisfactorio, proporcionan evidencia de que están en camino de alcanzarlo. Estos estudiantes muestran algunos logros fundamentales para el desarrollo de la competencia lectora.

Los estudiantes de este nivel comprenden solo textos breves y sencillos. Cuando se enfrentan a textos de extensión media y complejidad adecuada para el grado, únicamente ubican información que se puede encontrar fácilmente y realizan deducciones sencillas.

A continuación, se presenta un texto y dos preguntas representativas que logran resolver los estudiantes de este nivel.



Nivel En inicio **No logró los aprendizajes esperados**

Este nivel agrupa a los estudiantes que se alejan considerablemente de los aprendizajes esperados para el grado. Estos estudiantes solo leen oraciones y responden preguntas muy sencillas.

2.3 Historia

Esta información no está disponible.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

Para el establecimiento de puntos de cortes en la ECE, la UMC convocó a un conjunto de expertos en diversos aspectos del ámbito educativo (investigadores, curriculistas, especialistas y docentes de diferentes regiones del país, quienes han obtenido excelentes puntajes en las evaluaciones para la Carrera Pública Magisterial o han sido ganadores de los Concursos Nacionales de Buenas Practicas Docentes) a un taller de dos días para aplicar el procedimiento Bookmark con los resultados definitivos de la ECE.
En Lectura y Matemática de secundaria se contó con grupos de entre 18 y 19 jueces, mientras que para las áreas de primaria se contó con grupos de 24 jueces, compuestos en su mayor parte por docentes de aula de escuelas públicas y privadas. Asimismo, se convocó a expertos en evaluación en las áreas evaluadas, investigadores, elaboradores de textos escolares y formadores de formadores. También se procuró que los participantes provengan de todas las regiones del país.

3.2 Metodología

Tal como se mencionó anteriormente, estos niveles se establecieron a través de método Bookmark, el cual se realiza por única vez al inicio de un ciclo de evaluaciones, con la intención de sostener los mismos puntos de corte en las ediciones posteriores y asegurar que los resultados sean comparables en el tiempo.

La aplicación de Bookmark es descrita en los documentos de técnicos de la ECE de modo bastante apegado al procedimiento tradicional. Así, es método de puntajes de corte aplicado consiste en colocar marcas (tantas como cortes se hayan preestablecido) en un cuadernillo de ítems ordenado por dificultad. La pregunta típica que guía el establecimiento de cortes es la siguiente: "¿Hasta qué ítem debe ser capaz de resolver un estudiante, como mínimo, para ser considerado parte del nivel?". El procedimiento establece que los jueces, organizados en grupos pequeños, determinan, en tres rondas, los cortes para cada nivel de desempeño. De no llegar a un acuerdo, se aplican procedimientos estadísticos para resolver la discrepancia.
El juicio sobre los cortes considera tanto ítemes de selección múltiple como ítems de respuesta construida y con créditos parciales, dado que estos se incluyen en la misma métrica que los de opción múltiple.

El taller se condujo en tres rondas, tal como establece el procedimiento estándar. En la primera ronda, los participantes leyeron las descripciones de los niveles de logro elaboradas previamente por el equipo de la UMC, resolvieron todos los ítems de la prueba y analizaron las razones por las cuales un ítem era más difícil que el anterior. La primera ronda concluyó con un primer establecimiento individual de cortes. En la segunda ronda, los participantes expusieron, en los subgrupos las razones que los motivaron a colocar sus cortes. Asimismo, se les entregó un reporte de discrepancias donde señalaron qué tan distintos habían sido sus juicios con respecto de los demás grupos. La segunda ronda finalizó con un segundo establecimiento individual de cortes. En la tercera ronda, los participantes tuvieron acceso al datos de impacto, es decir, a la distribución de personas en los distintos niveles de desempeño, si los resultados de la segunda ronda hubiesen sido los definitivos. Con esta información, los jueces emitieron su tercer y último corte.

A fin de que los jueces cuenten con las condiciones para un juicio informado, tuvieron a su disposición los currículos vigentes, los textos escolares, las rutas de aprendizaje, entre otros documentos.

Un estudiante está en un nivel de logro si tiene una probabilidad mayor o igual a 0,62 para responder correctamente el ítem que marca el corte entre dos niveles de logro consecutivos (Minedu, 2005).

Es importante señalar que, antes del taller para establecer puntajes de corte, la UMC tomó las siguientes decisiones:

• Las escalas de Lectura y Matemática tendrían tres cortes y, por lo tanto, generarían cuatro grupos de personas: Satisfactorio, En Proceso, En Inicio y Previo al Inicio. La escala de Escritura tendría dos cortes (y tres grupos de personas). Esta decisión se tomó en función de la dispersión de la escala y el número de ítems, criterios que permitirían compatibilizar una descripción adecuada de los niveles y minimizar el error de clasificación.

• En las escalas de Lectura y Matemática, el corte del nivel En Proceso correspondería a la medida del corte Satisfactorio en la Evaluación Muestral (EM) 2013. Esto se pudo definir porque la prueba de segundo grado de secundaria fue equiparada con la de sexto grado de primaria.

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

Los niveles de logro se comunican junto con la entrega de los resultados de las pruebas nacionales censales a través de informes de resultados a nivel nacional, regional, para directivos (institución educativa), docentes y padres o apoderados.

Mientras en el reporte para padres, madres o apoderados se presenta una versión genérica de los niveles de logro, en el reporte a la institución educativa se presentan una descripción abreviada de estos niveles y en el reporte para docentes una versión más extendida que incluye ejemplos de preguntas para cada nivel.

Junto con los informes de resultados destinados a los establecimientos escolares, se difunde un documento con un taller para realizar en el contexto de una jornada de reflexión nacional sobre los resultados de la ECE (dicha jornada está mandatada por el Ministerio de Educación a través de un decreto). Este taller contiene orientaciones para trabajar con los docentes y también con padres, madres y apoderados.

5. USO ESTÁNDARES DE DESEMPEÑO

Si bien la evaluación no tiene consecuencias directas en las escuelas, los estándares de desempeño indican con claridad que los logros ahí descritos debieran ser demostrados por todos los niños y niñas del nivel al finalizar el año escolar.

Durante el año 2014 se realizó una encuesta sobre la recepción y uso de los informes de resultados de la Evaluación Censal (ECE), pero esta no profundiza en el tema de los estándares de desempeño[37].

[37] Se trata de una encuesta realizada por el Ministerio de Educación a todas las escuelas del país que indaga respecto de diferentes temas vinculados al sistema educativo. Respecto de la ECE, esta encuesta indaga, a través de indicadores cuantitativos, acerca de los usos que hacen directores y docentes de los informes de resultados (si los recibieron o no, si los leyeron o no, si realizaron o no una jornada de análisis, su nivel de comprensión, etc.).

6. REFERENCIAS

Ministerio de Educación del Perú (2014). *Evaluación Censal de Estudiantes (ECE) Segundo grado de primaria y Cuarto grado de primaria de IE EIB Marco de Trabajo.* Recuperado en Agosto de 2016 de: http://umc.minedu.gob.pe/wp-content/uploads/2014/07/Marco_de_Trabajo_ECE.pdf

Ministerio de Educación del Perú (2014). *Marco de fundamentación de las Pruebas de rendimiento de la Evaluación Censal De Estudiantes De 2.° De Secundaria 2015*. Recuperado en Agosto de 2016 de: http://umc.minedu.gob.pe/wp-content/uploads/2015/08/Marco-de-la-ECE-2%C2%BA.-de-secundaria.pdf

Ministerio de Educación del Perú (2014). *Reporte técnico de la Evaluación Censal de Estudiantes (ECE 2015). Segundo y Cuarto (EIB) de Primaria y Segundo de Secundaria.* Recuperado en Agosto de 2016 de: http://umc.minedu.gob.pe/wp-content/uploads/2016/07/Reporte-Tecnico-ECE-2015.pdf

Ministerio de Educación del Perú (2014). Reportes generales, para instituciones educativas, docentes y padres. Recuperados en Agosto de 2016 de: http://umc.minedu.gob.pe/evaluacion-censal-de-estudiantes-ece-2015/

Ministerio de Educación del Perú (2015). *Currículo nacional de la Educación Básica.* Recuperado en Julio de 2016 de: http://www.minedu.gob.pe/curriculo/pdf/curriculo-nacional-2016-2.pdf

Ministerio de Educación del Perú (2015). *Jornada de Reflexión: "Resultados de la ECE: Una oportunidad para reflexionar sobre el aprendizaje de TODOS los estudiantes de nuestra IE y no solo del grado evaluado"* ECE 2015 2.° Primaria. Recuperado en Agosto de 2016 de: http://umc.minedu.gob.pe/wp-content/uploads/2016/03/jornada-de-reflexion-2015_primaria.pdf

7. CONTACTO

Liliana Miranda Molina
Jefa de la Oficina de Medición de la Calidad de los Aprendizajes
lmiranda@minedu.gob.pe

[Nuestro contacto, Liliana Miranda Molina, fue nombrada Viceministra de Gestión Pedagógica en el trascurso de nuestro estudio (26 diciembre 2016). Desde el 12 enero 2017, el nuevo Jefe de la OMCE es Humberto Hildebrando Perez León Ibáñez – hperez@minedu.gob.pe]. Se le solicitó información vía email.

FICHA #11
National Assessment Program – Literacy and Numeracy (NAPLAN) - Australia
Programa Nacional de Evaluación – Alfabetización y Matemática - Australia

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la evaluación

Los estándares de desempeño australianos son de carácter nacional y son evaluados a través del Programa Nacional de Evaluación de Alfabetización y Matemática (*National Assessment Program – Literacy and Numeracy*; NAPLAN es su sigla en inglés). NAPLAN es una evaluación nacional de carácter censal que se aplica anualmente desde el año 2008, sustituyendo a las evaluaciones censales estatales.

Como complemento a la evaluación censal, el Programa Nacional de Evaluación (NAP), incluye, además, pruebas trienales de alfabetización en tecnologías de la información y la comunicación (años 6º y 10º); alfabetización científica (6º); y educación cívica y ciudadana (6º y 10º). El NAP incluye también la participación en PISA y TIMSS (Estudio de Tendencias Internacionales en Matemáticas y Ciencia), lo que permite el monitoreo del progreso de los estudiantes tomando como referencia los estándares internacionales.

1.2 Referente orientador de las evaluaciones.

Las pruebas de NAPLAN están alineadas con el curriculum nacional australiano según lo que se señala para las asignaturas de Inglés y Matemática.

1.3 Organismo responsable del programa de evaluación y del currículo.

La Autoridad Australiana de Currículo, Evaluación y Reporte (ACARA, *Australian Curriculum, Assessment and Reporting Authority*) está a cargo del NAPLAN.
Este organismo está a cargo, además, del desarrollo del curriculum nacional escolar, la realización de estudios sobre el desempeño escolar y políticas de responsabilización por resultados y asignación de recursos.

1.4 Áreas disciplinares

NAPLAN evalúa lectura, escritura, convenciones del lenguaje (ortografía, gramática, puntuación) y matemática. Tal como se mencionó anteriormente, estas áreas disciplinarias son consideradas en esta prueba de acuerdo a los énfasis del curriculum nacional.

1.5 Grados evaluados

Este programa se aplica a estudiantes de los años 3º (8 a 9 años de edad), 5º (10 a 11 años de edad), 7º (12 a 13 años de edad) y 9º (14 a 15 años de edad) de establecimientos escolares tanto públicos como privados.

1.6 Características de las pruebas

Las pruebas de NAPLAN consideran ítemes de selección múltiple y de respuesta abierta. Los ítemes de selección múltiple requieren que el estudiante escoja la opción correcta entre cuatro opciones posibles; los ítemes de respuesta abierta pueden ser respondidos a través de un número, una palabra o una oración breve. Todos los ítemes son puntuados de manera dicotómica (correcto o incorrecto). Las pruebas de aritmética para los años 7 y 9 están compuestas por dos formas. En la primera se permite el uso de calculadora, pero no en la segunda.
El análisis de las pruebas se lleva a cabo en dos etapas. La primera etapa involucra el análisis de los ítemes y del test en general, y la construcción de las escalas y tablas de equivalencias. Las escalas de reporte verticales se equiparan a las escalas históricas para asegurar comparabilidad desde la serie de tests aplicadas en 2008 (la comparabilidad se logra aplicando una prueba de *equating* alineada a los datos históricos de NAPLAN a una muestra de estudiantes). La segunda etapa de análisis incluye el análisis de toda la cohorte evaluada para dar cuenta de los resultados a nivel nacional.

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismo a cargo

Al igual que NAPLAN, los estándares de desempeño están a cargo de la Autoridad Australiana de Currículo, Evaluación y Reporte (ACARA).

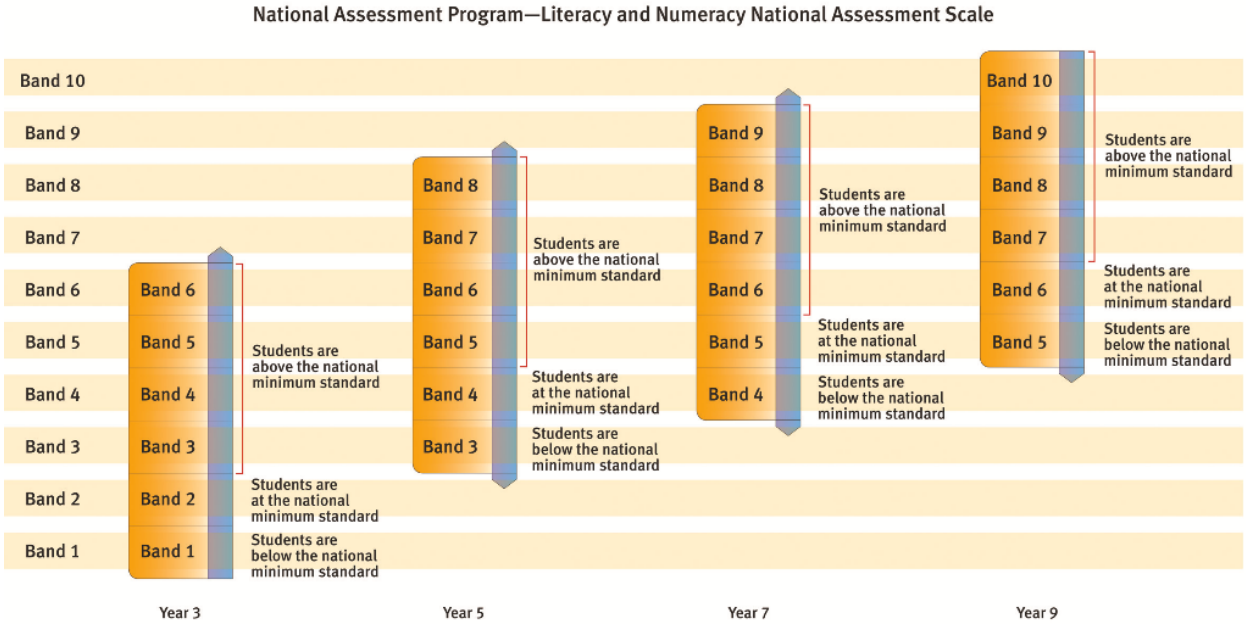
2.2 Características

Los estándares de desempeño australianos empleados en las pruebas NAPLAN son conocidos como "Estándares Mínimos" (*Minimun Standards*) y se describen para cada grado (3, 5, 7 y 9) y área evaluada (lectura, escritura, convenciones del lenguaje y matemática). Describen las habilidades y conocimientos que los estudiantes de un determinado grado típicamente pueden demostrar en un dominio específico.

Al analizarlos en conjunto, los estándares mínimos nacionales para los años 3, 5, 7 y 9 muestran una progresión en la complejidad de las habilidades y conocimientos que un estudiante debe demostrar a medida que avanza en su escolaridad. Los estudiantes que no alcanzan el estándar mínimo no han logrado los resultados esperados para su grado y están en riesgo de no poder seguir progresando adecuadamente en su escolaridad;

requieren apoyo para poder progresar adecuadamente. Los estándares mínimos nacionales también permiten monitorear el desempeño de los estudiantes en el tiempo, ya que se reportan a través de escalas de puntaje comparables.

Los estándares mínimos nacionales se ubican en un continuo demarcado por 10 bandas de desempeño que describen el progreso de los estudiantes a lo largo de su escolaridad en las áreas evaluadas por las pruebas NAPLAN. Los estándares mínimos nacionales se ubican en la banda que corresponda al grado evaluado. De este modo, en la banda 2 se ubica el estándar mínimo para el año 3; en la banda 4, el estándar mínimo para el año 5; en la banda 5, el estándar mínimo para el año 7; y en la banda 6, el estándar mínimo para el año 9. Es importante señalar que no todas las bandas se reportan para todos los años o grados evaluados. Para el Año 3 se reportan resultados empleando las bandas 1 a 6; para el Año 5, las bandas 3 a 8; para el Año 7, las bandas 4 a 9; y para el Año 9, las bandas 5 a 10. La siguiente imagen ilustra con claridad la relación entre años escolares, bandas y estándares mínimos nacionales:



Los estándares mínimos nacionales, junto con asociarse a una de las bandas de puntaje de NAPLAN, están asociados a una descripción que muestra los desempeños y conocimientos que puede demostrar un estudiante que alcanza el estándar. Esta descripción se compone de un párrafo general, además de indicadores que ilustran el desempeño esperado de modo más específico. En el caso de lectura, se distinguen indicadores según distintos tipos de textos. Los estándares de convenciones del lenguaje distinguen indicadores para puntuación y gramática. Finalmente, los estándares de matemática distinguen indicadores para número, geometría, álgebra, medición, probabilidades y datos, y razonamiento matemático. No se distinguen áreas más específicas en los estándares de escritura.

A continuación, se presenta como ejemplo el estándar mínimo nacional en Lectura para estudiantes de año 3:

Year 3

In Year 3, reading texts tend to have predictable text and sentence structures. Words that may be unfamiliar are explained in the writing or through the accompanying illustrations. Typically, these texts use familiar, everyday language.

At the minimum standard, Year 3 students generally make some meaning from short texts, such as stories and simple reports, which have some visual support. They make connections between directly stated information and between text and pictures.

When reading simple imaginative texts, students can:

- find directly stated information
- connect ideas across sentences and paragraphs
- interpret ideas, including some expressed in complex sentences
- identify a sequence of events
- infer the writer's feelings.

When reading simple information texts, students can:

- find directly stated information
- connect an illustration with ideas in the text
- locate a detail in the text
- identify the meaning of a word in context
- connect ideas within a sentence and across the text
- identify the purpose of the text
- identify conventions such as lists and those conventions used in a letter.

2.3 Historia

Previo al año 2016, los tests estaban referidos a los "*Statements of learning for English*" y los "*Statements of learning for mathematics*.". Dado que estos "statements" son coherentes con el curriculum nacional, las pruebas de NAPLAN no cambiaron significativamente, sin embargo, se debieron realizar algunas modificaciones en las especificaciones de las pruebas de NAPLAN. Por ejemplo, en el test de matemática se alinea la cantidad de ítemes para cada eje de la disciplina con los énfasis del curriculum nacional, lo que tiene como resultado que en la prueba de 2016 se incluyan menos ítemes del eje espacio (el que ahora se denomina como "geometría") y más ítemes de los ejes de números, estadísticas y probabilidades.

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

El desarrollo de los estándares está a cargo de los equipos técnicos de la Autoridad Australiana de Currículo, Evaluación y Reporte (ACARA).

3.2 Metodología

Tal como se señaló anteriormente, los estándares mínimos nacionales están asociados a las 10 bandas de puntaje con las que se reportan los resultados de las pruebas NAPLAN. Los puntajes de corte de estas bandas se establecen *a priori* a través de la determinación de intervalos de puntaje equivalentes entre bandas. Estos intervalos se fijaron en el año 2008 y permiten la comparación en el tiempo de las bandas reportadas.

La escala para cada área evaluada es equivalente y tiene la misma cantidad de bandas y puntajes de corte. Para indicar que el desempeño de un estudiante se ubica en una banda en particular, se espera que responda correctamente al menos 50% de las preguntas del test que corresponden a esa banda.

La siguiente tabla presenta los puntajes de corte para las bandas de desempeño de cada dominio evaluado. Así, por ejemplo, el punto de corte que separa las bandas 1 y 2 tiene un valor de 270 puntos en la escala de puntajes vertical, equivalentes a -2.257 puntos en la escala IRT normalizada (que varía de -3 a +3, aproximadamente).

		Logits				
Band Cut Score	**Scale Score**	**Reading**	**Writing**	**Spelling**	**Grammar and Punctuation**	**Numeracy**
9/10	686	3.928	7.380	5.821	3.783	3.907
8/9	634	3.155	5.629	4.457	3.076	3.042
7/8	582	2.382	3.878	3.092	2.369	2.176
6/7	530	1.609	2.126	1.728	1.661	1.310
5/6	478	0.836	0.375	0.363	0.954	0.444
4/5	426	0.063	-1.376	-1.001	0.246	-0.422
3/4	374	-0.710	-3.128	-2.366	-0.461	-1.288
2/3	322	-1.483	-4.879	-3.730	-1.169	-2.154
1/2	270	-2.257	-6.630	-5.095	-1.876	-3.020

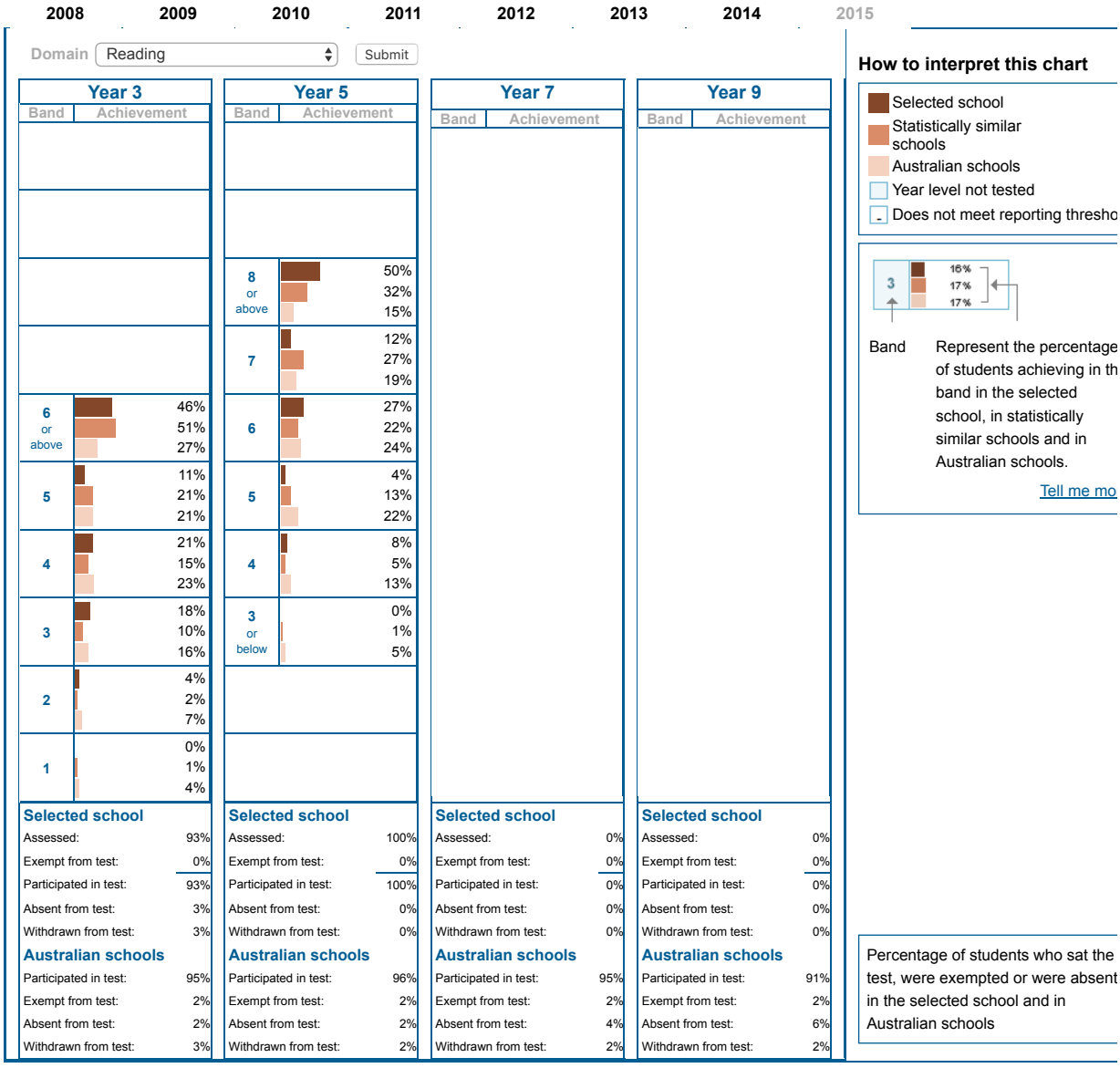
4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

Los estándares mínimos australianos son reportados a los establecimientos educacionales junto con los resultados obtenidos por sus estudiantes en las pruebas NAPLAN. Los establecimientos educacionales reciben los resultados de cada estudiante individualizados, además de análisis del conjunto de los estudiantes, por franjas y en niveles. A continuación, un ejemplo de la información que reciben los establecimientos escolares:

St John's Primary School, Scarborough, WA

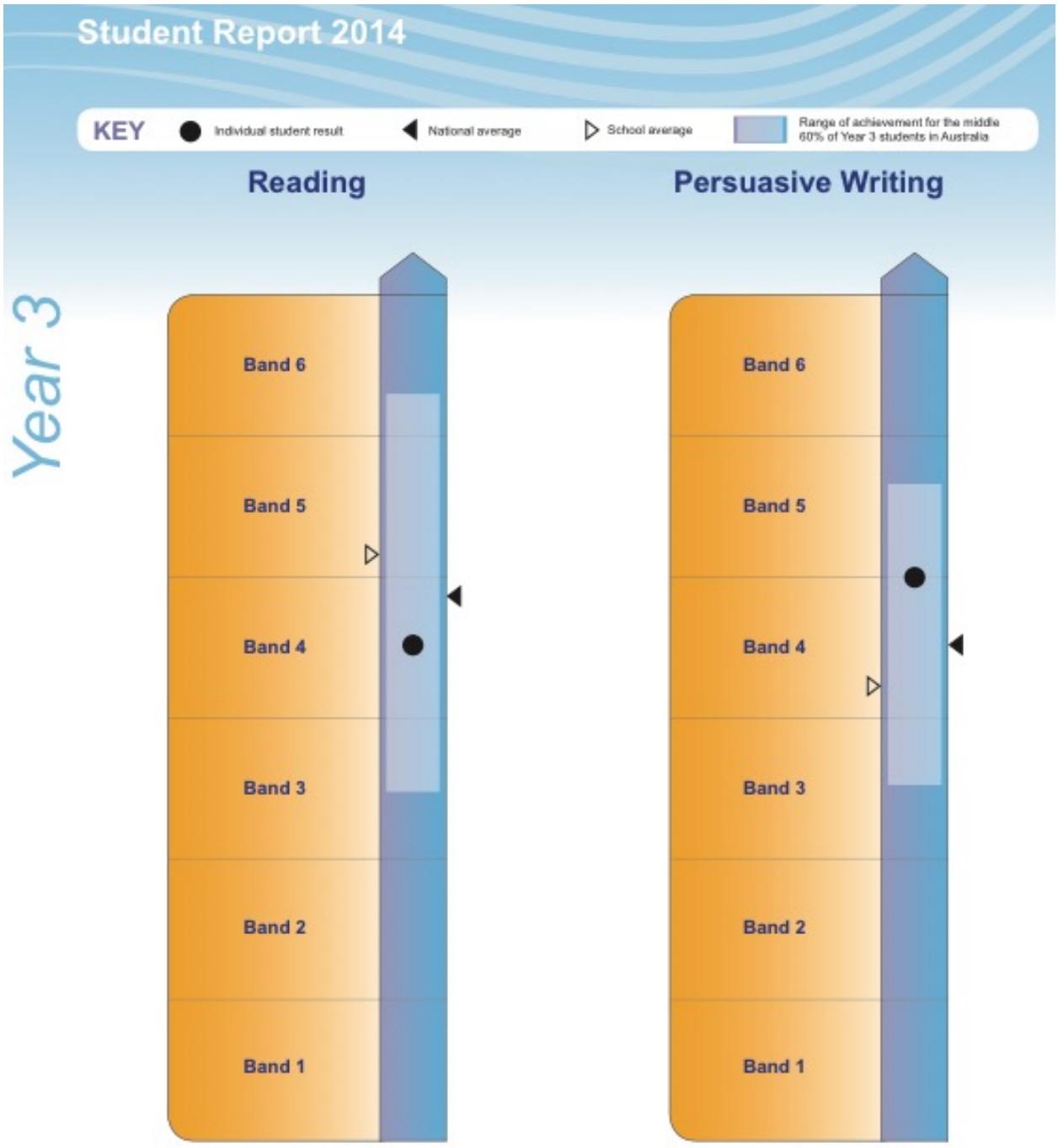
NAPLAN results are reported in bands. For more information, visit the NAPLAN website.

The chart below shows the school's results for the five domains at each year level. It displays the percentage of students achieving in each band, as well as the of students in statistically similar schools and the percentage of students in Australian schools achieving in each band.



Además, se dispone de un reporte por estudiante, el que es enviado a cada familia. En este reporte se señala, para cada área medida, el promedio, el estándar nacional mínimo y la franja en que cae el 60% de los alumnos en su grado. Se indica a su vez si el alumno alcanzó o no el estándar nacional mínimo. El reporte incluye una tabla con una breve descripción de las habilidades típicamente demostradas por alumnos en cada nivel. En algunos estados se incluye el promedio de su grado y centro educativo. Otros incluyen los ítemes respondidos correcta e incorrectamente en la prueba.

En la página siguiente se presenta un ejemplo de cómo se presentan los resultados de cada estudiante en el informe para padres de NAPLAN:



Students read a range of factual and non-factual texts with supporting pictures and diagrams. Students were assessed on aspects of reading that included:

- finding information that is clearly stated
- connecting ideas and drawing conclusions
- recognising how a character acts and thinks
- recognising a sequence of events
- recognising different opinions
- identifying the main idea or purpose of a text.

Students wrote a persuasive text and were assessed on aspects that included:

- supporting the reader and understanding the purpose of their writing
- structuring a persuasive text, developing ideas and points of argument, and making effective word choices
- using the conventions of written language such as grammar, punctuation, spelling and paragraphs.

5. USO ESTÁNDARES DE DESEMPEÑO

Desde el nivel central se incentiva el uso de los resultados asociados a estándares de desempeño de modo diagnóstico, para identificar estudiantes que necesiten ayuda. Por su parte, cada región tiene un modo diferente de apoyar a los centros educativos con herramientas para interpretación y análisis de resultados.

Los resultados por institución escolar son publicados en *My School*[38], un sitio web administrado por ACARA. Además de la consulta de establecimientos escolares en particular, este sitio web busca fomentar el uso de esta información para el monitoreo del desempeño a nivel nacional o jurisdiccional y la toma de decisiones sobre políticas educativas.

6. OTRA INFORMACIÓN RELEVANTE

No aplica.

7. REFERENCIAS:

Australian Curriculum, Assessment and Reporting Authority (2015). *National Assessment Program – Literacy and Numeracy 2014: Technical Report*, ACARA, Sydney.

Australian Curriculum, Assessment and Reporting Authority (2015). *Measurement Framework for Schooling in Australia* 2015. ACARA, Sydney.

Ejemplos de reportes de resultados consultados en: www.nap.edu.au y http://www.myschool.edu.au/

8. CONTACTO:

Dr. Stanley Rabinowitz

Administrador general de Departamento de Evaluación y Reportes - ACARA

stanley.rabinowitz@acara.edu.au

Stanley RAbinowitz fue contactado vía email.

[38] http://www.myschool.edu.au/

FICHA #12
National Monitoring Study of Student Achievement (NMSSA)
Estudio Nacional para Monitorear el Rendimiento de los Estudiantes, Nueva Zelanda

1. DESCRIPCIÓN DEL SISTEMA DE EVALUACIÓN

1.1 Nombre de la evaluación

El *National Monitoring Study of Student Achievement* (NMSSA)[39] --Estudio Nacional para Monitorear el Rendimiento de los Estudiantes— tiene como propósito informar sobre el rendimiento global de los estudiantes en Nueva Zelanda, con el fin principal de mejorar las prácticas pedagógicas y los aprendizajes en las escuelas. Para ello: (a) informa sobre tendencias en el tiempo del rendimiento de los estudiantes; (b) da cuenta de factores asociados al rendimiento educativo; (c) provee de información útil para diseñar políticas educativas, para la planificación curricular, y para la práctica educativa; y (d) provee de información al público en general.

NMSSA existe como tal desde 2012, siendo esta evaluación la sucesora del *National Education Monitoring Project (NEMP)*[40] –Proyecto Nacional de Monitoreo Educativo-- que existió entre 1995 y 2010. El Ministerio de Educación licita la realización de estas evaluaciones por 5 o más años. En 2010 se acabó el contrato entre el Ministerio y la Universidad de Otago (ver sección "Organismo responsable"). Este contrato recién pudo ser renovado en 2012.

1.2 Referente orientador de las evaluaciones

El NMSSA evalúa distintas áreas disciplinarias del curriculum nacional de Nueva Zelanda. Éstas son: lectura, escritura, matemáticas, ciencias naturales, ciencias sociales, salud, y educación física. Cada año se evalúan dos áreas, las que se van rotando de modo tal que todas las áreas son evaluadas en un ciclo de cuatro años. Así, por ejemplo, en 2014 se evaluaron las áreas de lectura y ciencias sociales.

El curriculum hace referencia a la visión, principios y valores en los que se quiere formar a los estudiantes. Por ejemplo, que los estudiantes sean seguros de sí mismos, y tengan capacidad de aprender a lo largo de toda la vida. También refiere a competencias claves (*key competencies*) comunes a distintos grados y áreas disciplinarias. Estas competencias aluden tanto a habilidades cognitivas (ej., capacidad de pensar, de resolver problemas) como al desarrollo socio-emocional (ej., capacidad de relacionarse con otros) de los estudiantes.

Además del curriculum, las pruebas están también alineadas con las progresiones de aprendizaje (*Literacy Learning Progression*[41]). Este documento apoya la implementación

[39] http://nmssa.otago.ac.nz/
[40] http://nemp.otago.ac.nz/
[41] http://www.literacyprogressions.tki.org.nz/

curricular, precisando el tipo de habilidades que se espera que los estudiantes puedan utilizar al final del 4o y 8o grado, en distintas áreas disciplinarias.

A partir del curriculum y de las progresiones de aprendizaje, se elabora un marco de evaluación conceptual (*Conceptual Assessment Framework*) para cada área disciplinaria evaluada. El marco de evaluación es la pieza clave para asegurar el alineamiento entre las pruebas y el curriculum. El marco de evaluación conceptual especifica las competencias claves (ej., resolución de prblemas), los objetivos conceptuales de la prueba (ej., identificar, interpretar), habilidades específicas (ej, identificar información, ideas; cuándo y cómo hacer inferencias), tipos de tareas o ítemes (ej., selección múltiple, desarrollo), tipos de texto (ej., poesía, ficción, no ficción), entre otros. A partir del marco de evaluación de crean las especificaciones para el desarrollo de ítemes. Las especificaciones dan cuenta de la cantidad de ítemes por objetivo conceptual, por tipo o formato de pregunta, entre otros.

1.7 Organismo responsable del programa de evaluación y del curriculum.

El NMSSA es posible gracias a la colaboración entre el Ministerio de Educación, el Consejo de Investigación Educativa (NZCER--*New Zealand Council for Educational Research*), y la unidad de investigación en evaluación educativa (EARU--*Educational Assessment Research Unit*) de la Universidad de Otago. El Ministerio de Educación externaliza la evaluación a estas dos instituciones, las que tienen a cargo distintos aspectos de la evaluación. El contrato actual entre con la Universidad de Otago es por cinco años.

No hay un marco legal que respalde al NMSSA ni los estándares de desempeño.

El Ministerio de Educación es el encargado de desarrollar el curriculum nacional. El NMSSA está principalmente a cargo del EARU, quien trabaja en forma coordinada con el Ministerio de Educación. El NZCER participa en diferentes instancias de la evaluación, teniendo una responsabilidad principal en las tareas de administración grupal.

El EARU presenta al Ministerio los instrumentos y procedimientos a utilizar en las evaluaciones, incluyendo los procedimientos para desarrollar estándares de desempeño alineados con el curriculum. El Ministerio no debe aprobar formalmente estos procedimientos, pero se trabaja en una lógica de consenso, negociando y llegando a común acuerdo.

1.8 Áreas disciplinarias evaluadas

Lectura, Escritura, Matemáticas, Ciencias Naturales, Ciencias Sociales, Salud y Educación Física.

1.9 Grados Evaluados

4º grado (Year 4: estudiantes de 8-9 años)

8º grado (Year 8: estudiantes de 12-13 años)

1.10 Características de las pruebas

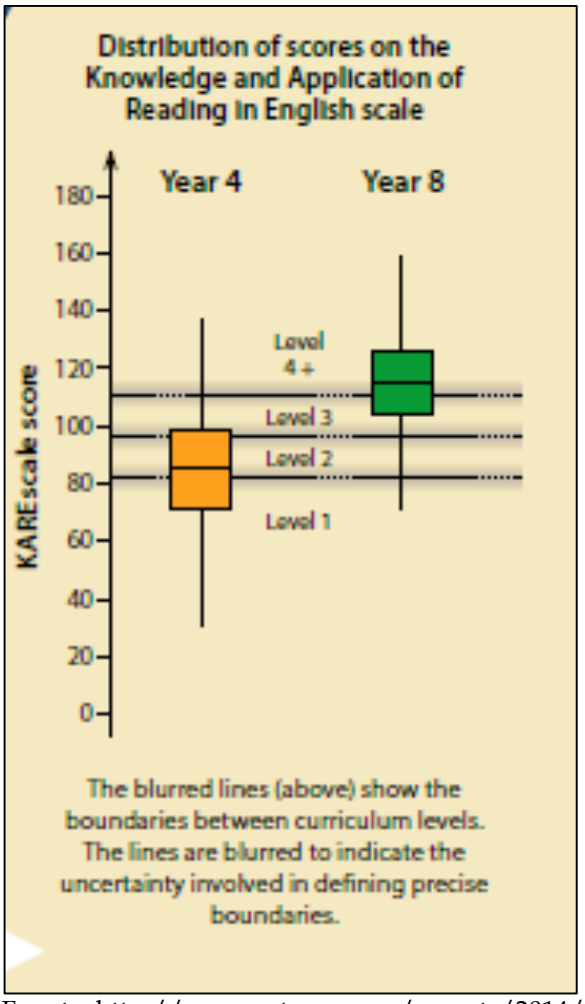
NMSSA se administra anualmente a muestras representativas a nivel nacional de estudiantes en 4º (Year 4, estudiantes de 8-9 años) y 8º (Year 8, estudiantes de 12-13 años) grados, que asisten a escuelas que usan el inglés como medio de instrucción. El 4º grado fue selecionado por ser el primero en que los estudiantes pueden participar en pruebas estandarizadas con una variedad de formatos. El 8º grado fue seleccionado por ser el último grado de la educación primaria (Flockton, 2012). El tamaño muestral es de 2.000 a 4.000 estudiantes por grado, repartidos en 100 escuelas, aproximadamente.

NMSSA posee una amplia cobertura curricular, asegurando así un fuerte alineamiento entre el curriculum nacional y la evaluación. Este alineamiento es posible gracias al uso de una variedad de formatos de administración de las pruebas, y de formatos de tareas (ítemes o preguntas) dentro de cada prueba. La evaluación usa dos formatos de administración: grupal e individual. La administración grupal utiliza pruebas de lápiz y papel. La administración individual abarca una muestra reducida de estudiantes (600-800).

La evaluación incluye tareas que son motivantes y relevantes en la vida de los estudiantes (criterio de autenticidad), tales como entrevistas, videos, realización de experimentos, realización de proyectos, lectura de un libro, escribir una carta, uso de marionetas, realización de una obra de arte, cantar y bailar, y actividades de educación física (Flockton, 2012).

Los resultados del NMSSA se reportan en una escala de puntajes IRT (modelo Rasch) que es común a los dos grados evaluados. Esta escala vertical tiene la ventaja de permitir cuantificar el progreso anual y entre los grados evaluados. La escala de puntajes fue estandarizada de modo tal que la media de todos los estudiantes evaluados (de 4º y 8º grado juntos) es igual a 100 puntos y la desviación estándar es igual a 20 puntos; con puntaje mínimos de 0 y máximos de 180 puntos, aproximadamente (Figura 1).

Figura 1. Escala vertical de puntajes de la prueba de Comprensión de Lectura (Escala KARE --Knowledge and Application of Reading in English)



Fuente: http://nmssa.otago.ac.nz/reports/2014/Reading_SOF.pdf

2. DESCRIPCIÓN DE LOS ESTÁNDARES DE DESEMPEÑO

2.1 Organismos a cargo

La unidad de investigación en evaluación educativa (*Educational Assessment Research Unit,* EARU) de la Universidad de Otago está a cargo de establecer los estándares de desempeño del NMSSA. El Consejo de Investigación Educativa de Nueva Zelanda (NZCER--*New Zealand Council for Educational Research*) colabora en esta tarea, aportando con profesionales en distintas etapas del proceso. El Ministerio de Educación participa revisando y consensuando criterios y procedimientos.

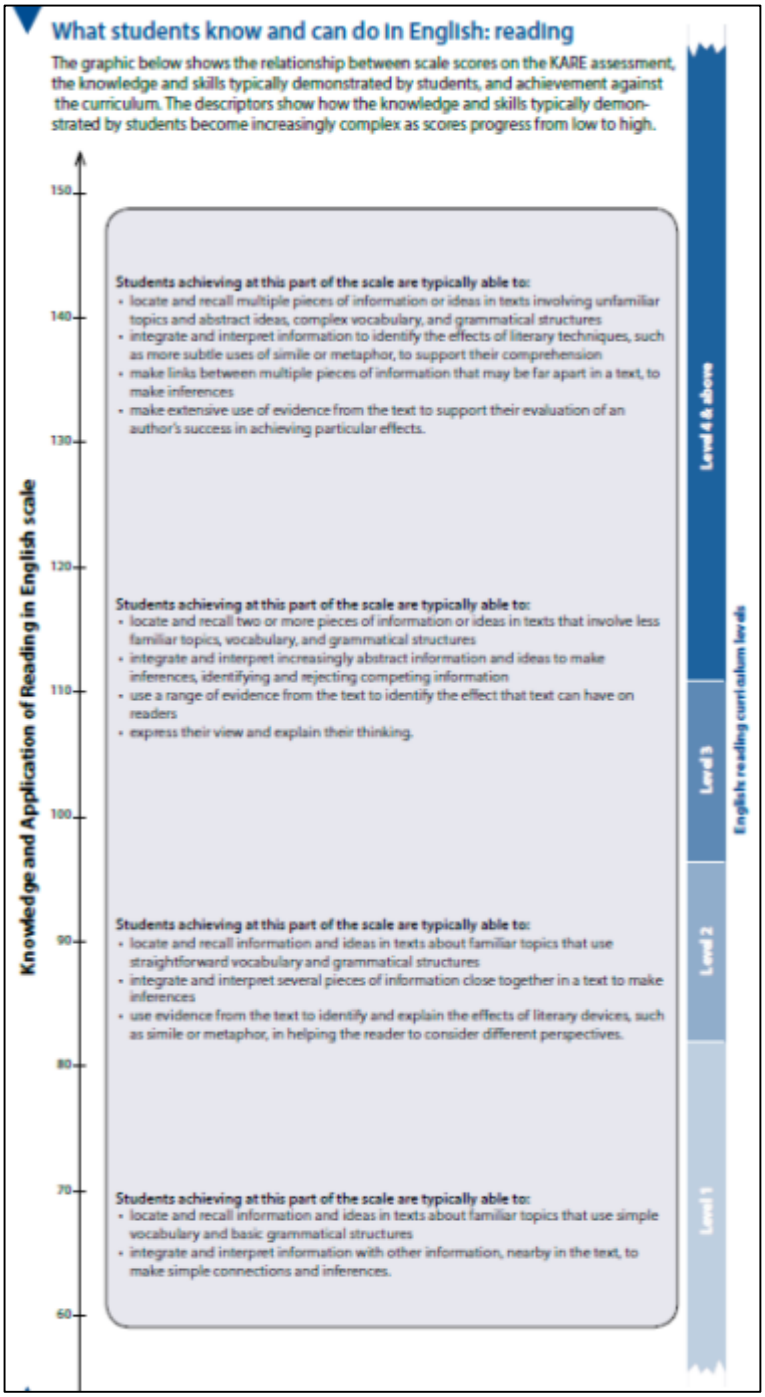
2.2 Características

Para dar un sentido pedagógico a la escala de puntajes de las pruebas, el NMSSA reporta resultados asociados a cuatro estándares de desempeño o Niveles Curriculares

("Curricular Levels"), que están anclados en la escala de puntajes. Los niveles describen lo que los estudiantes saben y son capaces de hacer en distintos puntos de la escala de puntajes (Figura 2).

Los niveles de desempeño tienen dos componentes: uno cuantitativo y uno cualitativo. El componente cuantitativo corresponde a los puntos de corte. Éstos refieren al valor de la escala de puntajes que diferencia entre los estudiantes que alcanzan un estándar desempeño y otro. El componente cualitativo corresponde a las descripciones asociadas a cada nivel. Estas descripciones dan cuenta de las habilidades, competencias, contenidos, y contextos en que se espera que los estudiantes demuestren lo que saben y pueden hacer.

Figura 2. Estándares de Desempeño de Comprensión Lectora.



Fuente: http://nmssa.otago.ac.nz/reports/2014/Reading_SOF.pdf

3. DESARROLLO ESTÁNDARES DE DESEMPEÑO

3.1 Instituciones y profesionales involucrados

- Unidad de investigación en evaluación educativa (*Educational Assessment Research Unit,* EARU) de la Universidad de Otago. Sus profesionales del área de evaluación y psicometría tienen la primera responsabilidad en el desarrollo de estándares.
- New Zealand Council for Educational Research. Sus profesionales del área de psicometría lideran parte del proceso.
- Ministerio de Educación. Participa revisando y consensuando procedimientos.
- Panelistas externos con experiencia en el área disciplinar evaluada, incluyendo docentes y curriculistas.

3.2 Metodología

El informe técnico y el informe de resultados del NMSSA 2014 (NMSSA, 2014a, 2014b) presentan información detallada sobre los procedimientos para desarrollar los estándares de desempeño. No hay otros protocolos o manuales disponibles. Estos procedimientos se refieren a: (a) Metodología para definir los puntos de corte, y (b) Metodología para desarrollar las descripciones.

A. Metodología para definir los puntos de corte (componente cuantitativo)

En Nueva Zelanda, el procedimiento para definir los puntos de corte asociados a los niveles de desempeño ("*Curricular Levels*") recibe en nombre de alineamiento curricular ("*curriculum alignment*").

Si bien hay algunas variaciones metodológicas según el área curricular evaluada, en general se distinguen las siguientes etapas y procedimientos, propios del método Bookmark (NMSSA, 2014a, Apendices 5 y 8; NMSSA, 2014b, Cap. 2):

1. Se invita a especialistas participar en un panel de alineamiento ("alignment panel"). El panel está compuesto por 6-9 personas. Profesionales del área de evaluación y psicometría del AERU y del NZCED lideran los paneles. Panelistas externos traen consigo experiencia en el área evaluada en la sala de clases, en docencia y curriculum. Los panelistas externos son propuestos por el AERU y elegidos con la aprobación de un Curriculum Advisory Panel y del Ministerio de Educación.
2. Profesionales del AERU presentan al panel información detallada sobre el marco de evaluación del área curricular evaluada. Es necesario que los panelistas tengan una acabada comprensión del marco de evaluación y de la escala de

puntajes de la prueba antes de hacer juicios sobre los puntos de cortes. Esto es así dado que, en rigor, están haciendo juicios sobre el alineamiento entre los niveles curriculares (estándares de desempeño) y la escala de puntajes.

3. Los ítemes (tareas, preguntas) de las pruebas de lápiz y papel[42] son presentadas a los panelistas y revisadas una a una. Se presentan los ítemes, textos y cada una de sus preguntas, criterios de corrección, y puntajes asociados; se incluyen notas adicionales, y ejemplos de respuestas de los estudiantes. Se discute sobre las habilidades en juego y etapas necesarias para responder a cada pregunta, y se aclaran dudas. El propósito es asegurar que los panelistas tengan una cabal comprensión de la prueba.
4. Los panelistas también discuten sobre las condiciones en la que los estudiantes rindieron las pruebas, y como éstas podrían afectar negativamente el desempeño de los estudiantes ypor ende, el nivel de dificultad de las preguntas. Por ejemplo, el que los estudiantes deban responder a la prueba en forma individual, sin ningún apoyo de sus docentes; el que haya un límite de tiempo para responder la prueba; y efectos de formato de las preguntas.
5. Los panelistas reciben cuadernillos con las preguntas de la prueba, ordenadas de más fácil a la más dificil. Las preguntas que pueden recibir 1 ó 2 puntos se presentan en dos ocaciones en los cuadernillos: la primera corresponde a la dificultad de dar una respuesta de un punto, la segunda corresponde a la dificultad dar una respuesta de dos puntos.
6. Se pide a los panelistas imaginar un grupo de 100 estudiantes con las competencias mínimas ("*minimal competence*") para estar clasificados en el Nivel 2. Se discute en profundida sobre el significado de competencia mínima (ej., un estudiante con competencia mínima justo alcanza las expectativas curriculares del nivel), hasta llegar a un consenso sobre su significado.
7. Se pide a los panelistas que, en forma individual, indiquen en su cuadernillo la primera página con una pregunta en donde los estudiantes mínimamente competentes del Nivel 2 tendrían una probabilidad menor a 0,70 de responder correctamente. Luego los panelistas discuten dónde pusieron sus marcas, y tienen la opción de moverlas. No es requisito que los panelistas lleguen a un total consenso donde poner las marcas. Sin embargo, en los casos con marcas más discrepantes, se pone más énfasis en que los panelistas justifiquen sus marcas. Esto usualmente los lleva a cambiar sus marcas y a acercarse al consenso.

[42] Las tareas o preguntas administradas individualmente en pruebas de desempeño (ej. hacer un proyecto) no son reportadas en la escala cuantitativa de puntajes, y por lo tanto no son consideradas en el ejercicio de puntos de corte. Estas tareas se reportan en forma cualitativa, con descripciones detalladas de las competencias que los estudiantes deben utilizar para llevar a cabo la tarea.

8. Se calcula el punto de corte asociado a la marca de cada uno de los panelistas. Esto se hace promediando el puntaje asociado a la pregunta marcada y el puntaje asociado a la pregunta inmediatamente anterior a dicha pregunta en el cuadernillo.
9. Se calcula el punto de corte del Nivel 2 promediando los puntos de corte de todos los panelistas.
10. Se repiten los pasos 6 a 9, ahora con un foco en el Nivel 4.
11. El punto de corte asociado al Nivel 3 se calcula promediando los puntos de corte asociados a los Niveles 2 y 4.

B. Metodología para elaborar las descripciones (componente cualitativo)

Las descripciones asociadas a los estándares de desempeño (Niveles Curriculares) especifican lo que un estudiante mínimamente competente debería poder hacer en cada Nivel Curricular. Cada nivel corresponde a dos años de aprendizaje en la escuela, equivalentes a dos grados en el curriculum nacional, aproximadamente. Así, las descripciones del Nivel 2 indican lo que un estudiante mínimamente competente debería saber y poder hacer en el 4o grado (*Year 4),* y las del Nivel 4 indican lo que un estudiante mínimamente competente debería saber y poder hacer en el 8o grado (*Year 8)* (NMSSA, 2014b, Cap. 2).

Las descripciones asociadas a los puntos de corte se basan estríctamente en las preguntas evaluadas. Dan cuenta de los objetivos conceptuales evaluados, junto con los contenidos, habilidades, y contextos en los que los estudiantes deben demostrar lo que saben y pueden hacer, en distintos puntos de la escala de puntajes. Esta base empírica aporta la evidencia necesaria para afirmar lo que los estudiantes son capaces de hacer o no, en distintos puntos de la escala de puntajes. Sin embargo, esto tiene el costo de que las descripciones no están estrictamente alineadas con el curriculum (NMSSA result report, p. 13).

Las descripciones son elaboradas por los mismos profesionales que elaboran las pruebas y por curriculistas, a partir del análisis de las preguntas que anclan en torno a los puntos de corte. Para ello, se ordenan las preguntas según su nivel de dificultad empírica. Se identifican aquellas que tienen un puntaje asociado cercano al punto de corte ($p = .70$). Se identifican las habilidades, contenidos y contextos asociados a cada una de las preguntas, y se escriben las descripciones asociadas.

También hay descripciones asociadas a los estudiantes en el Nivel 1 (es decir, bajo el punto de corte más bajo), y a los que están en el Nivel 4 (es decir, sobre el punto de corte más alto). Estas descripciones se hacen a partir del juicio profesional y no a partir de los ítemes que anclan en determinados puntos de corte.

Una vez desarrollados los estándares de desempeño, estos son presentados al Ministerio de Educación. Se recogen comentarios y, de haber discrepancias, se hacen los ajustes necesarios. No hay una validación o revisión con otras personas o instituciones.

Los estándares de desempeño son relativamente nuevos en Nueva Zelanda. Los procedimientos arriba descritos corresponden a la primera evaluación del NMSSA en donde se desarrollaron estándares de desempeño. Es decir, aún no se administra una evaluación en un área disciplinaria donde los estándares de desempeño ya hayan sido desarrollados. A la fecha, no hay una visión clara sobre cómo y cuando se harán futuras revisiones a los estándares de desempeño. Lo más probable es que se haga de nuevo el proceso de desarrollo de estándares, para verificar la convergencia de resultados entre dos procesos independientes.

4. COMUNICACIÓN ESTÁNDARES DE DESEMPEÑO

EL NMSSA tiene un impacto importante entre docentes y otros profesionales de la educación. Esto es posible gracias al uso de una variada gama de productos comunicacionales, que incluyen sitio web, informes de resultados, informes técnicos, y folletos de difusión.

Los informes de resultados enfatizan la comunicación de resultados en relación al curriculum y la progresión de aprendizajes entre ambos grados evaluados. Los resultados incluyen el porcentaje de estudiantes según niveles curriculares y según el logro de las expectativas curriculares para su grado. Los resultados muestran claramente la amplia distribución de puntajes en cada grado, y el traslape que existe entre ambas distribuciones (ver Figura 1 de esta ficha). También se da cuenta del progreso anual en los aprendizajes de los estudiantes (NMSSA, 2014b). Los resultados incluyen información detallada para cada una de las tareas evaluadas.

Los resultados de las evaluaciones se publican a nivel nacional, y para distintos grupos de estudiantes (e.g., hombres/mujeres, por etnias, por grupo socioeconómico). No se reportan por escuelas, y no hay consecuencias asociadas para éstas.

Es un desafío pendiente generar más espacios para la discusión de resultados de las evaluaciones entre docentes y educadores. En su versión anterior (NEMP), se hacían foros nacionales de educadores para discutir los resultados. Estos foros contribuían a formar una cultura de evaluación entre educadores, y promover cambios en las formas de enseñar en las escuelas.

5. USO ESTÁNDARES DE DESEMPEÑO

En el NMSSA hay una especial preocupación asegurar el alineamiento entre las pruebas y las prácticas docentes. Esto es clave para potenciar el aprendizaje de los estudiantes. Por ello, hay una especial preocupación por involucrar a los docentes en la evaluación.

Esto facilita que los docentes comprendan, valoren, y usen las evaluaciones para la mejora.

El NMSSA usa dos estrategias principales para asegurar el alineamiento entre la evaluación y las prácticas docentes. Primero, ofrece a los docentes oportunidades de perfeccionamiento y de participación en el diseño de las pruebas (ej., elaboración de tareas y de pautas de corrección alineadas con el curriculum), administración de las pruebas (en escuelas distintas a las escuelas donde trabajan), y corrección (aplicando pautas de corrección). Segundo, publica alrededor de 2/3 de las tareas utilizadas en las pruebas cada año. De esta manera, los docentes pueden utilizar dichas tareas en su trabajo cotidiano con sus estudiantes (Flockton, 2012).

6. REFERENCIAS:

NMSSA. 2014a. Ministry of Education, New Zealand, New Zealand Council for Educational Research, and the Educational Assessment Research Unit (EARU) of the University of Otago. National Monitoring Study of Student Achievement, English: Reading, 2014 Overview. NMSSA Report 5.1.
http://nmssa.otago.ac.nz/reports/2014/Reading_OverviewPDF.php

NMSSA. 2014b. New Zealand Council for Educational Research and Educational Assessment Research Unit, (EARU) University of Otago. Technical Information 2014, Social Studies, English: Reading. NMSSA Report 7.
http://nmssa.otago.ac.nz/reports/2014/Technical_MOE_PDF.php

Flockton, L. (2012). *The development of the student assessment system in New Zealand.* Systems Approach for Better Education Results – Student Assessment (SABER-SA). Washington DC: World Bank.

Sitios web:

Literacy Learning Progressions: http://www.literacyprogressions.tki.org.nz/
NMSSA: National Monitoring Study of Student Achievement:
http://nmssa.otago.ac.nz/
NEMP: National Education Monitoring Project 1995-2010: http://nemp.otago.ac.nz/
New Zealand Council for Educational Research:
http://www.nzcer.org.nz/research/national-monitoring-study-student-achievement-wanangatia-te-putanga-tauira

7. CONTACTO

Alison Gilmore
Co-Directora
Unidad de investigación en evaluación educativa (EARU -- Educational Assessment Research Unit) de la Universidad de Otago, Nueva Zelanda
http://www.otago.ac.nz/education/staff/alisongilmore.html
Email: alison.gilmore@otago.ac.nz

Alison Gilmore fue entrevistada en videoconferencia por Skype.

5. CUADRO RESUMEN DE LA INFORMACIÓN

En la página siguiente se presenta una tabla que sintetiza la información más relevante de cada uno de los sistemas educativos considerados en esta revisión.

Tabla: Resumen de información sobre estándares de desempeño para cada uno de los sistemas educativos revisados

Pais*	**CHL**	**AUS**	**CAN**	**ESP**	**GBR**	**GBR**	**MEX**	**NLD**	**NZL**	**PER**	**USA**	**USA**	**USA**
Jurisdiccion			**ONT**		**ENG**	**SCT**						**NY**	**VA**
1. Características del sistema de evaluación[43]													
1.1.Tipo de Muestra (Censo: C, Muestra:M	C	C	M	M	C	M	C y M	C y M	M	C	M	C	C
1.2.Tipo de consequencias (C/C: con consecuencias desde publicación de resultados a cierre de escuelas; S/C: sin consecuencias)	C/C	C/C	S/C	S/C	C/C	S/C	S/C	C/C	S/C	S/C	S/C	C/C	C/C
2.Características generales de los Estándares de Desempeño[44]													
2.1.Se elaboran tomando como primera referencia el currículum oficial.	Si	Si	Si	Si	Si	Si	No	Si	Si	Si	No	Si	Si
2.2 Se establecen basados en niveles de aprendizaje establecidos en el curriculum oficial.	No	Si	No	No	Si	Si	No	Si	SI	No	No	No	No
2.3.Todos los contenidos y habilidades están especificados en el currículum del grado evaluado.	Si	Si	Si	Si	Si	Si	Si	No	Si	Si	No	Si	Si
2.4.Están estrictamente referidos a los contenidos y habilidades evaluados en las pruebas	No	Si	Si	Si	Si	Si	Si	Si	SI	SI	SI	SI	SI
2.5 Se establecen operacionalizando el currículum formal del sistema.	Si	Si	Si	Si	Si	Si	No	Si	Si	SI	No	SI	SI

[43] CAN (ONT): Las pruebas son censales, pero solo tiene consecuencias (certificación de estudios secundarios) la prueba de 10o grado. Las otras pruebas son para diagnóstico del sistema y devolución de información diagnóstica a docentes, estudiantes y familias. MEX: Hay dos versiones de Planea: una versión para diagnóstico del sistema que es una prueba muestral, otra versión censal para devolver información, sin consecuencias, a las escuelas. USA: No existen políticas nacionales de accountability. La prueba NAEP es para informar a nivel de la nación. GBR (ENG): Las pruebas aportan información a los resultados que considera la Inspección de escuelas, que finalmente clasifica a todas las escuelas en categorías asociadas a consecuencias, que pueden llegar a la intervención o eliminación de la escuela del registro público.

[44] CAN (ONT): Los estándares usan el curriculum como punto de partida, pero no forman parte formal del curriculum provincial. ESP: La evaluación a nivel país se implementó una sola vez, por lo que no se alcanzaron a hacer revisiones. MEX: los estándares 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. NZL: Los estándares de desempeño se establecen en referencia al curriculum nacional y a los "Literacy Learning Progressions", que operacionalizan el curriculum y describen expectativas de aprendizaje. En NZL no se han hecho ajustes a los estándares de desempeño dado que son relativamente nuevos. USA: Dado que en EEUU no hay un curriculum nacional oficial, los estándares de desempeño se elaboran a partir de las expectativas de aprendizaje de los marcos de evaluación. Los estándares de desempeño en NAEP solo los utiliza NAEP y sus informes nacionales.

Pais*	CHL	AUS	CAN	ESP	GBR	GBR	MEX	NLD	NZL	PER	USA	USA	USA
Jurisdiccion			ONT		ENG	SCT						NY	VA
2.6 Son referidos a criterios	Si	Si	Si	Si	Si	Si	Si	No	Si	SI	Si	SI	SI
2.7.Las descripciones se ajustan cuando hay cambios en el currículum oficial	SI	Si	Si	N/A	Si	S/I	S/I	S/I	N/A	S/I	Si	S/I	SI
2.8.Las descripciones se ajustan o revisan según un calendario preestablecido	SI	S/I	No	N/A	S/I	S/I	No	No	N/A	S/I	No	S/I	SI
3.Metodología del establecimiento de los Estándares de Desempeño[45]													
3.1.Están a cargo del Ministerio de Educación (o su equivalente)	Si	Si	No	Si	Si	Si	No	No	No	Si	No	Si	Si
3.2.Las categorías o etiquetas se establecen bajo responsabilidad conjunta de autoridades del currículum y las encargadas de las pruebas.	No	Si	No	Si	N/A	Si	No	N/A	No	Si	No	Si	Si
3.3.Las categorías o etiquetas se establecen con participación amplia de actores del sistema educativo.	P	No	Si	P	N/A	S/I	P	S/I	P	P	Si	P	Si
3.4.Los puntajes de corte o intervalos de puntajes se establecen con participación amplia de actores del sistema educativo.	P	No	Si	P	P	S/I	Si	S/I	P	Si	Si	Si	Si
3.5.Se utilizan resultados de las pruebas, o pilotos en establecimiento de puntajes de corte o intervalos de puntaje para categorizar resultados.	Si	Si	Si	Si	S/I	S/I	Si	Si	Si	Si	Si	Si	Si
3.6.Se siguen metodologías de equiparación (equating) psicométricas para garantizar comparabilidad interanual en los puntos de corte.	Si	Si	Si	N/A	S/I	Si	Si	S/I	N/A	Si	Si	Si	Si
4.Transparencia y control de calidad													

[45] CAN (ONT): En distintas etapas del proceso de establecimiento de estándares se involucra docentes, administradores y el publico interesado. ESP: Estándares de desempeño a cargo del INEE (Instituto Nacional de Evaluación Educativa) del Ministerio de Educación. Los estándares de desempeño (puntos de corte y descripciones) son definidos por los mismos especialistas que elaboraron las pruebas, con la colaboración de especialistas externos. La evaluación a nivel de país se realizó una sola vez por grado, por lo que no fue necesario hacer equating para comparaciones interanuales. MEX: Un panel de expertos tomo la decision acerca de las características generales de los estándares (4 niveles, y sus nombres) y luego paneles con participación amplia de actores trabajaron en su operacionalización. NZL: Los estándares de desempeño fueron introducidas recientemente, por lo que aún no reportan comparaciones interanuales. USA: En USA no hay un curriculum oficial nacional. Se han seguido distintas metodologías históricamente para fijar los puntos de corte, siempre con participación amplia de distintos tipos de actores, siempre a cargo de la agencia autónoma que gobierna la prueba.

Pais*	CHL	AUS	CAN	ESP	GBR	GBR	MEX	NLD	NZL	PER	USA	USA	USA
Jurisdiccion			ONT		ENG	SCT						NY	VA
4.1 Informes técnicos están disponibles para el público.	Si	Si	Si	No	P	P	Si	P	Si	Si	Si	Si	Si
4.2.Existe procedimiento formal para evaluacion independiente del desarrollo de estándares. Resultados de estas evaluaciones son públicos.	No	S/I	Si	No	S/I	S/I	No	No	No	S/I	Si	S/I	S/I
5.Comunicación y uso de los Estándares de Desempeño[46]													
5.1.Existen documentos que explican la relación de estándares de desempeño y el currículum formal	P	P	Si	No	Si	Si	Si	S/I	Si	No	No	P	P
5.2.Existen documentos sobre relación de estándares y la prueba con experiencias de aprendizaje de estudiantes en sus escuelas.	No	S/I	Si	No	S/I	Si	No	Si	No	S/I	No	S/I	S/I
5.3.Se siguen estrategias para comunicar y entrenar usuarios clave en la interpretación y el uso de estándares de desempeño.	No	Si	Si	No	S/I	Si	No	S/I	No	Si	Si	Si	Si
5.4.Existe política de monitoreo de estudiantes, centros educativos, o del sistema que usa los estándares de desempeño en forma explícita.	Si	S/I	Si	Si	Si	Si	Si	Si	No	S/I	Si	S/I	S/I

**Codigos ISO -Alpha-3: CHL=Chile; AUS=Australia; CAN -ONT= Canada, Ontario; ESP = Espana; GBR = Gran Bretana; GBR -ENG = Gran Bretana, Inglaterra; GBR -SCT = Gran Bretana, Escocia; MEX = Mexico; NLD = Holanda; NZL = Nueva Zelandia; PER = Peru; USA = Estados Unidos; USA - NY = Estados Unidos, Nueva York; USA -VA = Estados Unidos, Virginia.*

P: cumplimiento parcial (implica que el equipo a cargo de esta investigación considera que el criterio está presente, pero de manera intermitente o incompleta

N/A: no aplica

S/I: sin información (esta información no existe, no es pública, o el contacto no tenía acceso a la información)

[46] CAN (ONT): Existe un documento "Understanding Levels of Achievement" que explica y ejemplifica los estándares de desempeño en términos de actividades de aprendizaje y practicas instruccionales, preparado para docentes con participación de docentes en su preparación. ESP: La evaluación a nivel de país alcanzó a ser administrada una vez en cada grado solamente. El plan era que, a partir de los resultados, se elaboran compromisos de revisión y mejora educativa. Sin embargo, solo se publicaron informes de resultados.

6. ANÁLISIS DE LA INFORMACIÓN ENCONTRADA

Tal como se señaló al inicio de este informe, este estudio busca cumplir con el siguiente objetivo general y sus respectivos objetivos específicos:

Objetivo general: describir y analizar el "proceso de elaboración, seguimiento y evaluación de estándares de desempeño en distintos sistemas educativos a nivel internacional":

Objetivo específico 1: Caracterizar y describir sistemas educativos que elaboran, implementan y evalúan estándares de desempeño en su jurisdicción.

Objetivo específico 2: Caracterizar y describir el proceso de elaboración, implementación y evaluación de estándares de desempeño, incluyendo la evaluación que se realiza respecto al uso de los estándares y las evaluaciones del impacto del uso de estándares en el aprendizaje, en los sistemas educativos seleccionados.

Objetivo específico 3: Analizar comparativamente las características de cada uno de los sistemas educacionales analizados y sus procesos de elaboración de estándares, identificando tanto elementos comunes como particularidades relevantes.

Objetivo específico 4: Generar recomendaciones para el sistema educacional chileno respecto de metodologías de elaboración, seguimiento y evaluación de estándares de desempeño, considerando el contexto y las particularidades del país."

Para cumplir con el objetivo general y dar luego curso al cumplimiento de los objetivos específicos, nuestra revisión de prácticas en el establecimiento de estándares de desempeño incluyó un catastro general de información básica con respecto a los sistemas de evaluación de países de la OCDE, Perú y México y algunos países más de Asia.
Para cumplir con los objetivos específicos 1, 2 y 3, llevamos a cabo un estudio de mayor profundidad de 12 sistemas educativos que emplean estándares de desempeño en la evaluación educativa. En el análisis comparativo de estos sistemas educativos, enfocamos aspectos generales del sistema de evaluación (objetivo específico 1), características generales de los estándares, metodologías del establecimiento de estándares, aspectos de transparencia y control de calidad, y aspectos de comunicación y de uso de los estándares (objetivo específico 2). Estos temas fueron explorados en profundidad en las fichas que se encuentran en la sección 4 de este documento y hemos resumido las características de cada sistema en Cuadro Resumen en la tabla que se presenta en la sección 5. Cabe señalar que no siempre fue posible encontrar información que abarcara los mismos elementos o con una misma profundidad en cada uno de los casos, debido a que las fuentes consultadas no tenían esta información disponible o porque el nivel de desarrollo del sistema no contaba con esta información de manera pública o suficientemente sistematizada..

Es importante primero notar que las características generales de los sistemas de evaluación varían según las consecuencias que se espera que tengan. De tal modo, como es natural, se realizan evaluaciones censales principalmente en sistemas que buscan que las evaluaciones de estándares de desempeño tengan consecuencias para estudiantes o centros educativos individuales, aunque hay algunas excepciones (por ejemplo: Ontario y México).

En los sistemas educativos que estudiamos en este informe, encontramos que los estándares de desempeño comúnmente se definen en referencia al currículum oficial[47]. A menudo, sin embargo, el curriculum oficial no hace prescripciones acerca de los rendimientos en términos medibles u operacionalizables. Una de las primeras labores técnicas es derivar estándares de desempeño medibles mediante un análisis del curriculum oficial y el uso de opinión de expertos en curriculum y en medición educativa. Esto llevaría, en la mayoría de las ocasiones, a centrar la formulación de los estándares en estricta referencia al instrumento de medición: la prueba y sus ítemes. En otras ocasiones, menos frecuente en los países examinados (Chile por ejemplo) los estándares incluyen habilidades que no se miden en las pruebas.

En todos los casos, los sistemas educativos confrontan el problema de cómo abordar las consecuencias de cambios o reformas en el curriculum oficial, a fin de asegurar el continuo alineamiento de este con respecto a los estándares de desempeño y las pruebas que los miden. Pocos de los sistemas de evaluación que estudiamos tienen políticas explícitas formuladas al respecto, aunque en entrevistas y documentos es claro que en la mayoría existe preocupación por guardar, por un lado, el alineamiento al día con el curriculum oficial, y por otro, mantener la posibilidad de comparar válidamente los desempeños de estudiantes a lo largo de una serie de años para monitorear mejoría, deterioro o estabilidad en los resultados de la escolarización en el sistema educativo. Un cambio en los estándares de desempeño después de una reforma o cambio del curriculum significa sacrificar la comparabilidad de los estándares en el tiempo, mientras que no hacer el cambio puede conservar la comparabilidad en el tiempo, pero sacrificar una alineación óptima entre los estándares y el currículum. En casos donde se presentan estos cambios, esto se informa explícitamente al sistema escolar alertando acerca de la comparabilidad de las mediciones (Estado de Nueva York, por ejemplo) o acerca de las áreas que se han incluido, excluído o cuyo peso en la prueba ha cambiado (Australia, por ejemplo).

Es claro también que existen dos formas de pensar en el ámbito internacional acerca del papel que debe jugar el Ministerio de Educación (o su equivalente) en el establecimiento de estándares[48]. En algunos países (ej., Chile, Australia, España y otros) el Ministerio de Educación los establece y un sistema de evaluación nacional o estatal los operacionaliza. En otros sistemas (ej., Ontario, México, Holanda y otros), una agencia distinta del Ministerio de Educación se encarga de establecer los estándares. En la primera forma de proceder, prevalece la opinión de que los expertos en curriculum del Ministerio están en mejor posición para velar por el alineamiento apropiado entre el curriculum oficial (que también está a su cargo) y los estándares de desempeño. En la segunda forma, prevalece una opinión de que los estándares

[47] Ver encabezado 2 de la Tabla Resumen y las fichas correspondientes a cada país.
[48] Ver encabezado 3 de la Tabla Resumen y las fichas correspondientes a cada país.

de desempeño son los criterios de calidad contra los cuales se verificaran los logros del sistema educativo, y que por consiguiente una agencia independiente del Ministerio está en mejor posición de garantizar objetividad y transparencia en los juicios basados en los estándares.

En ambos casos, sin embargo, se requiere contar con evidencia empírica para apoyar conclusiones acerca de la alineación de los instrumentos. Estas evidencias son cruciales para juzgar cuán bien fundamentadas son las conclusiones acerca de la calidad de los sistemas educativos, los establecimientos escolares o de los logros de estudiantes individuales. Hay mucha varianza en los criterios de evidencia para apoyar inferencias acerca de la alineación o para fundamentar juicios acerca de los niveles de aprendizaje entre los distintos países. En casi todos los sistemas educativos estudiados en profundidad encontramos criterios de evidencia claros y procedimientos de verificación en curso, e informes disponibles al público.

Hay importantes diferencias en cuanto a la participación o no de actores claves del sistema educativo en el establecimiento de estándares de desempeño. Ontario es un ejemplo de un sistema educativo que incluye un espectro amplio de actores, incluyendo docentes, administradores de centros educativos e inclusive miembros del público interesado en los paneles para fijar estándares. En cambio, hay otros países en los que hay una participación más restringida de actores, como es el caso en Chile, México, Nueva Zelandia y otros países. En Chile, quienes participan por lo general todos docentes y especialistas de las disciplina de la Región Metropolitana. En otros países, el establecimiento de estándares vista como una tarea técnica que debe estar a cargo únicamente de expertos, como ocurre en Australia.

Hay poco consenso con respecto a los criterios para verificar la calidad del trabajo que se hace para establecer y analizar los propios estándares de desempeño[49]. Entre los expertos en medición existe un consenso de que los procedimientos de establecimiento, análisis y comunicación de los estándares deben también estar sujetos a la auditoria externa y que estas auditorías deben estar disponibles al público. Sin embargo, son muy pocos los sistemas educativos que realizan estas auditorías o comparten sus resultados, los encontramos solo en Canadá y EEUU (ver fichas 5 y 6 presentadas en la sección 4 de este informe para más información).

No hay consenso con respecto a si los estándares de desempeño deben ser herramientas de referencias para las practicas pedagógicas[50]. Internacionalmente, los estándares de desempeño se usan principalmente como parte de una política formal de monitoreo de estudiantes, centros educativos o del sistema en su conjunto. En algunos países es solo un instrumento de monitoreo formal del sistema educativo, y no se hacen esfuerzos por traducir los estándares en términos apropiados para que docentes o estudiantes tengan la oportunidad de entender su relación con las actividades de enseñanza y aprendizaje que se viven cotidianamente en las aulas. Este es el caso de NAEP en los Estados Unidos. En otros sistemas (ej., Ontario, Escocia, Holanda), existe un interés importante por parte de las autoridades por el uso de los estándares

[49] Ver encabezado 4 de la Tabla Resumen y las fichas correspondientes a cada país.
[50] Ver encabezado 5 de la Tabla Resumen y las fichas correspondientes a cada país

de desempeño por parte de escuelas, docentes y otros actores del sistema como instrumentos para mejorar su trabajo. En estos sistemas existen documentos, videos, formación en servicio y otros mecanismos para comunicar los estándares y procurar que estas tengan influencia en el trabajo de las aulas. La estrategia en esos países es vincular los estándares de desempeño con las experiencias de aprendizaje de los estudiantes en sus escuelas, a fin de servir como instrumento de mejoría. Es decir, la intención es que los estándares de desempeño en esos sistemas tengan impacto directo en la calidad de las oportunidades educativas que se ofrecen en las aulas.

7. RECOMENDACIONES PARA EL SISTEMA EDUCACIONAL CHILENO

El objetivo de los estándares de desempeño, es hacer posible decisiones categóricas acerca del desempeño de estudiantes con el mayor rigor, confiabilidad y validez posibles. Los estándares se evalúan para el monitoreo general del sistema educativo y a veces también para tomar decisiones con consecuencias para estudiantes, sus docentes o los establecimientos escolares. Estas decisiones tienen consecuencias para las oportunidades de vida de estudiantes y docentes, y las oportunidades institucionales y del personal de establecimientos escolares.

A los sistemas educativos que usan los estándares de desempeño, les interesa determinar por tanto en todos los casos, qué categorías de desempeño alcanzan sus estudiantes en determinados hitos de su trayectoria escolar y justificar las decisiones que se hacen a partir de ello. Estos pueden ser decisiones con relativamente leves consecuencias para individuos o establecimientos como, por ejemplo, decidir que el sistema educativo está logrando o no criterios aceptables de calidad. O bien pueden impactar fuertemente a estudiantes o establecimiento, otorgándoles (o negándoles) a los primeros credenciales u oportunidades educativas específicas o para justificar incentivos o sanciones asignados a las escuelas.

Debido a la importancia social de estas decisiones, las categorizaciones del desempeño de estudiantes deben ser independientemente verificables y seguir criterios sólidos de evidencia. Los sistemas de evaluación en los países que estudiamos en este informe, son un muestrario importante del estado actual de las técnicas empleadas para la definición, análisis, verificación y comunicación de estándares de desempeño. En todos ellos, los estándares de desempeño – como también lo indica la literatura científica al respecto – tienen dos características básicas: están fundamentados en evidencia empírica y son verificables mediante la evidencia empírica.
Esta sección delinea algunas propuestas de prioridades orientadoras para futuros esfuerzos en el ámbito de estándares de desempeño en Chile. Estas se derivan de nuestro análisis de las prácticas en los países estudiados, y en consideración de las características del caso chileno. La intención es que adelanten conversaciones acerca de futuras direcciones, estrategias, y procedimientos en estándares derivados de la experiencia internacional

Recomendación 1: Se debe guardar tanto la comparabilidad en el tiempo de los estándares, como su óptimo alineamiento con una política curricular dinámica, mediante una estrategia dual.

> Es desafiante formular una respuesta técnica a la necesidad de contar con estándares de desempeño de alta calidad en un contexto de una política curricular dinámica, que responda a nuevas ideas y necesidades refinando, extendiendo o inclusive reformulando sus contenidos, habilidades prioritarias, y otras características. Es también cierto que todo sistema educativo tiene interés en contar con la posibilidad de comparar el logro de estándares de año en año, y mantener series temporales largas de datos acerca de la

calidad de la escolarización. Encontramos que la estrategia dual (utilizada en NAEP de EEUU) de medir y reportar estándares invariantes de desempeño a lo largo del tiempo, y otros estándares puestos al día y alineados a la política curricular más reciente, ofrece el mejor balance entre la necesidad de alinear los instrumentos de política curricular apropiadamente, y la necesidad de contar con la posibilidad de aprovechar largas series temporales para monitorear el progreso del sistema educativo.

En el caso de Chile, se podrían mantener estándares de desempeño como aquellos que son requeridos por la Ley de Aseguramiento de la Calidad para el monitoreo y clasificación de las escuelas y otro tipo de estándares o expectativas más generales que no varíen en el tiempo o lo hagan de modo menos frecuente. Estos últimos pueden basarse en el aprendizaje descrito tanto marcos curricualres pasados como en expectativas fijadas en otros países o a través de pruebas internacionales, de manera de recoger expectativas que de manera más o menos estable han sido consideradas como relevantes para los estudiantes. Esto requeriría de analizar si es necesario o no enriquecer las pruebas para dar cuenta de este tipo de estándares y establecer metodologías para fijar puntos de corte adecuadas (no necesariamente debiera seguir la misma metodología empleada para los otros estándares).

Una ventaja de implementar este tipo de estándares es que se podrá tener una idea nítida y a largo plazo acerca de cómo varían los logros académicos de los estudiantes en el tiempo.

Recomendación 2: Usar niveles de desempeño preliminares para informar el desarrollo de ítemes de la prueba.

En Chile no se usan niveles de desempeño definidos previamente a partir del curriculum para informar el desarrollo de ítemes de la prueba. En consecuencia, no se vela por desarrollar una cantidad suficiente de ítemes que describan dichos niveles. En México y Ontario, se desarrollan niveles de desempeño preliminares, a partir del curriculum, para informar la elaboración de ítemes. 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 ítemes representen una muestra sólida de todos los aspectos incluidos en 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 unas de las prácticas más usadas para velar por esa confiabilidad.

Recomendación 3: Se requiere de una política que permita la verificación y auditoría pública de la calidad del trabajo en los estándares y su evaluación.

Asegurar que un estudiante, una escuela, o un sistema educativo ha alcanzado un estándar de desempeño determinado es una aseveración que debe ser auditable – una aseveración que debe ser independientemente verificable, para justificar las decisiones que se toman a partir de ese juicio.

Así como la evaluación de los estándares de desempeño son una herramienta para verificar la eficacia del sistema educativo, los procedimientos de establecer y evaluar estándares de desempeño también deben estar sujetos a verificación y auditoria externa. Los sistemas de establecimiento de estándares, y los sistemas de evaluación de los mismos, no solo deben evaluar internamente la calidad de sus procedimientos y resultados poniéndolos luego a disposición del público interesado, sino que también deben de establecer una política de auditoria externa periódica que también debe ser transparente a la ciudadanía. Ver casos de Ontario y EE. UU (criterio 4.2 en Cuadro Resumen).

Recomendación 4: Se requiere de la participación amplia de actores dentro y fuera del sistema educativo en los procesos de establecimiento de estándares.

Categorizar a estudiantes, escuelas o subsistemas educativos según niveles de desempeño, implica hacer juicios de valor acerca del aprendizaje que se espera todos los estudiantes alcancen. En estos juicios deben participar un amplio espectro de la ciudanía e institucionalidad de una nación o jurisdicción. Inclusive, en procedimientos de carácter exclusivamente técnicos, participan docentes de aula, administradores de establecimientos escolares y miembros del público interesado en la definición de estándares y el establecimiento de puntos de corte (ver caso de Ontario). Esta participación aumenta el conocimiento y compromiso con los estándares de desempeño, y contribuye a su legitimación social y “face validity” o validez aparente (Criterios 3.3 y 3.4 en Cuadro Resumen).

En el caso de Chile, se sugiere evaluar la posibilidad de incluir a actores más diversos (no solo docentes y especialistas en las disciplinas) en el proceso de elaboración de los estándares, ya sea en la elaboración de las descripciones, selección de ejemplos, establecimiento de puntajes de corte o estrategias para comunicar los estándares. De modo especial, se sugiere ampliar el rango geográfico de los actores consultados ya que, de acuerdo a la información disponible en los documentos facilitados por Mineduc, los particpantes son, en su mayoría, de la Región Metropolitana. En un país altamente centralizado como Chile, explorar el uso de metodologías online ofrece una oportunidad para involucrar a un amplio espectro de la cuidadanía. Esto permite darle mayor validez a los estándares, involucrando, por ejemplo, a panelistas de las distintas regiones del país.

Recomendación 5: Los estándares de desempeño referidos a criterios deben contar con herramientas que guíen a estudiantes, docentes, familias y el público interesado a entender como verificar su cumplimiento en las aulas y las escuelas del sistema educativo

Los sistemas educativos establecen y miden estándares de desempeño porque le atribuyen a estos un impacto positivo en el aprendizaje. Los aprendizajes son producto principalmente de las oportunidades que se ofrecen en las aulas, y se debe de contar con descripciones y evidencia de validez instruccional de los estándares, a fin de que docentes, estudiantes y sus familias cuenten con los elementos para orientar su trabajo y sus objetivos. En otras palabras, la magnitud de impacto depende de acciones concretas que permitan relacionar los estándares con aquello que el profesor observa en su sala de clases (criterios 5.1 y 5.2 en Cuadro Resumen).

Al respecto, se sugiere que una institución externa evalue el impacto que han tenido los estándares de desempeño en las prácticas docentes, para verificar, entre otros efectos, que los docentes procuran que todos sus estudiantes logran las expectativas nacionales a través de estrategias que no involucren prácticas no deseadas tales como el estrechamiento curricular o de las estrategias de evaluación empleadas en el aula. Del mismo modo, se sugiere investigar prácticas efectivas y pedagógicamente significativas que ayudan a los docentes y a las escuelas en general a movilizar el aprendizaje de todos sus estudiantes.

Recomendación 6: Realizar un estudio de equating entre los estándares de desempeño del SIMCE y los estándares de pruebas internacionales.

El nivel de exigencia de los estándares de desempeño del SIMCE no es el mismo que el asociado a los estándares de desempeño de las evaluaciones internacionales. Esto, dadas las diferencias en los marcos de evaluación, criterios y procedimientos para definir los estándares. Para verificar empíricamente estas variaciones (y los resultados asociados a las mismas), podría ser de interés para el MINEDUC realizar un estudio de *equating* entre los estándares de desempeño del SIMCE y los estándares de pruebas internacionales. Esto se podría hacer cuando el SIMCE y una evaluación internacional sean administrados el mismo año, en el mismo grado y en la misma asignatura. Por ejemplo, se podría hacer con los datos de PIRLS 2016 (recogidos en 2015 en Chile) y la prueba SIMCE 2015 de lectura de 4º grado.

Siguiendo la metodología propuesta por Philips (2012) el equating se podría hacer utilizando los datos del SIMCE, pero solo para el subconjunto de estudiantes que participaron en la muestra internacional (o una muestra equivalente). Así, los datos para el estudio de equating provendrían de dos muestras iguales o equivalentes, lo que permitiría poner los resultados en una misma escala de puntajes. Este procedimiento tiene la ventaja de basarse en resultados reales de los estudiantes. Es distinto al procedimiento actualmente utilizado en Chile, en donde se hace un juicio profesional

sobre la equivalencia de los estándares de desempeño utilizados en pruebas nacionales e internacionales (UCE, 2014, p. 68)."

En conjunto con lo anterior, se podría hacer un estudio comparativo entre las evaluaciones (especificaciones de las pruebas y las pruebas mismas) efectivamente utilizadas en Chile y en la evaluación internacional. Esto permitiría conocer las similitudes y diferencias entre ambas evaluaciones, permitiendo así mejor comprender los resultados del equating. Este estudio podría utilizar alguna de las metodologías de análisis de alineamiento descritas en este informe, o alguna de las metodologías utilizadas en los estudios comparativos del NAEP.

8. BIBLIOGRAFÍA

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2000). A Comparison of the Angoff and Bookmark Standard Setting Methods (p. 13).

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/

CEPPE (2013). Learning Standards, Teaching Standards and Standards for School Principals: A Comparative Study. OECD Education Working Papers No. 99

Cizek (2012a). Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

Cizek (2012b). An Introduction to Contemporary Standard Setting: Concepts, Characteristics, and Contexts. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

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.

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.

Elacqua, G., Martínez, M., Santos, H., Urbina, D., Treviño, E. & Place, K. (2013). *Los efectos de las presiones de accountability sobre las políticas y prácticas en escuelas de bajo desempeño: el caso de Chile*. Santiago de Chile: PREAL.

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

Ferdous, A.A. y Plake, B.S. 2005 *Understanding the Factors that Influence Decisions of Panelists in a Standard-Setting Study*. Applied Measurement in Education. Vol 18, No. 3 pp. 257-267

Flórez, M. (2013). *Análisis crítico de la validez del Sistema de Medición de la Calidad de la Educación (SIMCE*). Reino Unido: Universidad de Oxford.

García-Huidobro, J. E. (2014). *Políticas Educativas de "Rendición De Cuentas" Y Políticas Sociales De "Corresponsabilidad". ¿Algo En Común?*. Educación y Políticas Sociales.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting Cut Scores: Post-Standard-Setting Panel Considerations for Decision Makers. Educational Measurement: Issues and Practice, 29(1), 38–44.

Haertel, E. H. (2002). Standard Setting as a Participatory Process: Implications for Validation of Standards-Based Accountability Programs. Educational Measurement: Issues and Practice, 21(1), 16–22.

Hambleton, R. (1999). Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. Lab Report 377. http://www.nciea.org/publications/SetStandards_Hambleton99.pdf

Hamilton, L., Stecher, B., & Yuan, K. (2008). *Standards-based reform in the United States: History, research, and future directions*. Paper commissioned by the Center on Education Policy, Washington, D.C. for its project on Rethinking the Federal Role in Education.

Lewis, D., CTB/MCGRAW-HILL, Mitzel, H., and PACIFIC METRICS. 2012. The Bookmark Standard Setting Procedure. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

Manzi, J., Bogolasky, F., Gutiérrez, G., Grau, V. & Volante, P. (2014). *Análisis sobre valoraciones, comprensión y uso del SIMCE por parte de directores escolares de establecimientos subvencionados*. Informe preliminar FONIDE (no publicado).

Meckes, L. and Carrasco, R. (2010) 'Two decades of SIMCE: an overview of the National Assessment System in Chile', Assessment in Education: Principles, Policy & Practice, 17: 2, 233 — 248.

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

Ministerio de Educación (2014). *Fundamentos Estándares de Aprendizaje 4º y 8º básico.* Documento de trabajo de la Unidad de Currículum y Evaluación del Ministerio de Educación.

Ministerio de Educación (2015). H*acia un sistema completo y equilibrado de evaluación de los aprendizajes en Chile: Informe Equipo de Tarea para la Revisión del SIMCE.* Santiago: Mineduc.

Ministerio de Educación de Chile (2003). *Evaluación de aprendizajes para una educación de calidad: Comisión para el desarrollo y uso del sistema de medición de la calidad de la educación*. Santiago de Chile: Mineduc.

Mullis. Using scale anchoring to interpret the TIMSS and PIRLS Achievement Scales. http://timssandpirls.bc.edu/methods/pdf/TP11_Interpret_Achievement.pdf

OCDE (2004). *Revisión de políticas nacionales de educación: Chile*. Paris: OCDE

OECD (2004). *Chile: Reviews of National Policies for Education.* Paris: OECD, Centre for Co-operation with non members.

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

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

Peterson, C. H., Schulz, E. M., & Engelhard, G. J. (2011). Reliability and Validity of Bookmark-Based Methods for Standard Setting: Comparisons to Angoff-Based Methods in the National Assessment of Educational Progress. Educational Measurement: Issues & Practice, 30(2), 3–14.

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.

Philips, G. (2012). The Benchmark Method of Standard Setting. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

Plake y Cizek (2012). Variations on a theme: The modified Angoff, Extended Angoff, and Yes/No Standard Setting Models. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

Plake, B. S., Impara, J. C., & Irwin, P. M. (2000). Consistency of Angoff-based Predictions of Item Performance: Evidence of Technical Quality of Results from the Angoff Standard Setting Method. Journal of Educational Measurement, 37 (4), 347–55.

Ravitch, D. (2010) The death and life of the great American school system: how testing and choice are undermining education; New York, Basic Books.

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

UCE (2014). Fundamentos Estándares de Aprendizaje Matemática, Lenguaje y Comunicación: Lectura – II Medio. Ministerio de Educación. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

Zieky, M. (2012). So much as changed: An Historical Overview of Setting Cut Scores. En Cizek (Ed.): Setting Performance Standards: Foundations, Methods, and Innovations (2nd Ed.). Routledge: New York.

9. ANEXOS

ANEXO 1: CONTACTO DE INSTITUCIONES O PROFESIONALES QUE PUEDAN REALIZAR ASESORÍAS EN LA MATERIA.

Escocia – Reino Unido
Marion MacRury
Scottish Survey of Literacy and Numeracy Team
Scottish Government
marion.macrury@gov.scot

España
Ruth Martín Escanilla
Jefe de área de evaluación
Instituto Nacional de Evaluación Educativa
Ministerio de Educación, Cultura y Deporte
ruth.martin@mecd.gob.es

Inglaterra – Reino Unido
Liz Twist
Interim Deputy Director for Test Development
Standards & TestingAgency (STA)
liz.twist@education.gsi.gov.uk

Holanda
Frans Kleintjes
Coordinador de Cito Internacional
Frans.kleintjes@cito.nl
http://www.cito.com/contact

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

NAEP – Estados Unidos
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

México
Felipe Martinez 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

Perú
Liliana Miranda Molina
Jefa de la Oficina de Medición de la Calidad de los Aprendizajes
lmiranda@minedu.gob.pe

Australia
Dr. Stanley Rabinowitz
Administrador general de Departamento de Evaluación y Reportes - ACARA
stanley.rabinowitz@acara.edu.au

Nueva Zelanda
Alison Gilmore
Co-Directora Unidad de investigación en evaluación educativa (EARU -- Educational Assessment Research Unit) de la Universidad de Otago, Nueva Zelanda
alison.gilmore@otago.ac.nz

ANEXO 2: 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 o General	Titulo	Autor	Descripción	Nombre de Archivo
General	Using scale anchoring to interpret the TIMSS and PIRLS 2011 Achievement scales	Ina Mullis	Este documento analiza el proceso de anclaje de la escala para los niveles de logro de las pruebas TIMSS y PIRLS 2011	G_Mullis_Scale TIMSS PIRLS_sa
General	What matters most for student assessment systems: a framework paper.	Marguirite Clarke. World Blank	Este documento entrega una visión general sobre las materias construyen un sistema de evaluación más efectivo para los estudiantes.	G_Clarke_matters st assessment_2012
General	National Testing of Pupils in Europe: Objectives, Organisation and Use of Results	Eurydice	Este informe el contexto y la organización de las pruebas nacionales en 30 países europeos y el uso de sus resultados a nivel de esttudiante, escuelas y sistema.	Hol_Eurydice_Nat_te sting_2009
Australia	Student report 2014. National Assessment program - Numeracy and literacy	ACARA	Este informe entrega los resultados del test nacional en matemáticas y lenguaje en el año 2014	AU_ACARA_2014 Report_2014
Australia	NAPLAN 2016 Alignment with Australian Curriculum	ACARA	Esta carta describe sucintamente el alineamiento entre los test y el curriculum	AU_ACARA_Letter Align_2016
Australia	National Assessment program. Literacy and Numeracy. 2014 Technical report	NAPLAN - ACARA	Este es el reporte técnico del test nacional del 2014 para matemáticas y lenguaje.	AU_NAPLAN_Tech Report 2014_2014
Australia	Developing the enabling context for School - based assessment in Queensland, Australia.	Reg Allen - World Bank	Este es un reporte para describir el sistema de evaluación de Queensland en Australia	AU_Allen_Assess System_2012

Australia	Measurement framework for Schooling in Australia	ACARA	Este informe entrega los resultados y el marco de evaluación en el sistema educativo de Australia	AU_ACARA_Measure framework_2015
Australia	2015 Naplan test reporting.	QCAA	Este es un informe que da cuenta de la prueba rendida por el grado 5 en la prueba estandarizada.	AU_QCAA_2015 Naplan test_2015
Australia	2015 NAPLAN test reporting.	QCAA	Este es un informe que muestra un ejemplo de informe al estudiante de grado 7 de sus resultados en la prueba estandarizada.	AU_QCAA_2015 Y7 report_2015
Australia	2016 NAPLAN student report	ACARA	Ejemplo de informe a los estudiantes de sus resultados en el test.	AU_ACARA_2016 NAPLAN Report_2015
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_Tecn ical 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_EQA O 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_Fra me 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 - Ontario	Understanding levels of achivement 2012. Junior Division	EQAO	Este documento entrega información para vincular la evaluación de aula con las evaluaciones que desarrolla	Can_Ont_EQAO_Und Levels JD_2012

			EQAO. Junior Division.	
Canada - Ontario	Curriculum conections in lenguage: reading and writing	EQAO	Este documento trabaja la conexión entre la prueba EQAO de lectura y escritura y matemáticas en educación primaria y junior, con el curriculum de la región de Ontario.	Can_Ont_EQAO_Curr conections_2011
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_EQA O Assess_2013
Chile	Los argumentos en favor de los mapas de progreso en Chile	Margaret Forster	Detalla los principales ejes del trabajo de asesoría de ACER (Australia) al Ministerio de Educación de Chile entre 2002 y 2007.	CL_Forster_Ases_Ace r-MINEDUC_2007
Chile	Fundamentos estándares de aprendizaje matemática, lenguaje y comunicación: Lectura II medio	UCE	Describe antecedentes para la elaboración de estandares de aprendizaje	CL_UCE_Estandares Aprendizaje_2014
Chile	Learning Standards, teaching Standards, and Standards for the school principals. A comparative studies.	OECD	Este informe realiza una comparación sobre la implementación de estándares en diferentes paises, tomando el caso de Chile.	CL_OECD_Standards_ 2013
Chile	Fundamentos. Estándares de aprendizaje matemática, lenguaje y comunicación: lectura 2 medio.	Unidad de Curriculum y Evaluación. Ministerio de Educación	Este documento contiene los antecedentes, fundamentos, definiciones y la descripción del proceso de elaboración de Estándares de Aprendizaje para II medio en las asignaturas de Matemática y Lenguaje y Comunicación: Lectura	CL_UCE_Estandares Lectura_2014
Chile	Anexos. Estándares de Aprendizaje. II Medio	Unidad de Curriculum y Evaluación. Ministerio de Educación	Este documento trabajo en torno a la definición del metodo Bookmark para le definición de los puntajes de corte.	CL_UCE_Est Aprend_2014

EEUU	A history of NAEP Achievement levels: Issues, Implementation, and impact 1989-2009.	Mary Lyn Bourque	Este documento analiza la historia del NAEP entre los años 1989 - 2009.	EEUU_Bourque_His NAEP_2009
EEUU	Evaluation of the National Assessment of Educational Progress	Chad Buckendahl	Este documento analiza el diseño de la evaluación	EEUU_Buckendahl_Ev NAEP_2009
EEUU	Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science	National Assessment Governing Board	Este documento describe los niveles de logro del NAEP ciencia 2009	EEUU_NAGB_NAEP Science_2009
EEUU	Design document for 12th grade NAEP Preparedness Research Judgmental Standard Setting Studies	National Assessment Governing Board	Este documento describe los procedimientos para el establecimiento de estándares basado en juicio de expertos para el NAEP 2012.	EEUU_NAGB_Judgme ntal standard_2010
EEUU	Reliability and Validity of Bookmark-based methods for standard setting: comparisons to Angoff-Based Methods in the National Assessment of Educational Progress	Christina Hamme Peterson, E. Matthew y George Engelhard.	Este artículo académico analiza el método bookmark en la prueba NAEP	EEUU_Hamme_book mark-based method_2011
EEUU	Developing achievement levels on the 2014 National Assessment of Educational Progress in Grade 8 Technology and Engineering literacy.	National Assessment Governing Board	Este documento describe el proceso y los resultados de un estudio diseñado e implementado por Pearson para desarrollar las recomendaciones de los niveles de logro para el NAEP 2014 en Technology y alfabetismo ingenieril.	EEUU_NAGB_Dev Achivement_2016

EEUU	Developing Achievement levels on the National Assessment of Educational Progress in Writing grades 8 and 12 in 2011	National Assessment Governing Board	Este documento entrega una descripción detallada del proceso de definición de los niveles de logro para el NAEP en el grado 8 y 12 en escritura.	EEUU_NAGB_Dev Achievement levels_2012
EEUU	Developing Achievement levels on the 2005 National Assessment of Educational Progress in Grade twelve Mathematics.	National Assessment Governing Board	Este documento informa sobre los materiales y el proceso de establecimiento de los niveles de logro de la prueba de matemáticas del NAEP 2004.	EEUU_NAGB_Dev Achievement_2005
EEUU	Science framework for the 2009 National Assessment of Educational Progress	National Assessment Governing Board	Este documento describe el marco de la prueba de ciencias en el NAEP 2009	EEUU_NAGB_Science frame_2008
EEUU	Assessment and Item specification for the NAEP 2009 Mathematics Assessment.	National Assessment Governing Board	Este documento informa sobre las especificaciones de la prueba y los itemes para el NAP 2009 de matemáticas	EEUU_NAGB_Assess and Item_2007
EEUU	Reading Assessment and Item Specification for the 2009 national assessment of educational progress.	National Assessment Governing Board	Este documento entrega las especificaciones de la prueba de lectura del NAEP 2009	EEUU_Read Assess NAEP2009_2009
EEUU	Science Assessment and Item Specifications for the 2009 National Assessment of Educational Progress	West Ed and the council of chief state school officers	Este documento entrega las especificaciones de la prueba e itemes para el NAEP 2009 de ciencia.	EEUU_WEd_Science 2009NAEP_2007
EEUU	Mathematics Framework for the 2013 National Assessment of Educational Progress	National Assessment Governing Board	Este documento entrega el marco de la prueba de matemáticas del NAEP 2013	EEUU_NAGB_Math fram NAEP2013_2012

EEUU	Reading Framework for the 2013 National Assessment of Educational Progress	National Assessment Governing Board	Este documento informa sobre el marco de la prueba de lectura del NAEP 2013.	EEUU_NAGB_Read Frame NAEP2013_2012
EEUU	Reading Framework for the 2015 National Assessment of Educational Progress	National Assessment Governing Board	Este documento entrega el marco para la prueba de lectura del NAEP 2015	EEUU_NAGB_Read Fram NAEP2015_2015
EEUU	Evaluation of the achievement levels for mathematics and reading on the national assessment of educational progress	Christopher Edley y Judith Koening	Este documento evalua los niveles de logro para matemática y lectura en el NAEP	EEUU_Edley_Ev of achievement_2016
EEUU	A history of NAEP Assessment Framework	Carol Jago	Este documento entrega una visión histórica del NAEP	EEUU_Jago_History NAEP_2009
EEUU	Why define achievement levels and set standards?		Esta presentación analiza los niveles de logro y los estándares de la prueba NAEP.	EEUU_Why achievement_sa
EEUU	NAEP 2012. Trends in academic progress. Reading 1971-2012. mathematics 1973-2012.	National Center for Educational Statistics	Este documento analiza las tendencias históricas del NAEP tanto para lectura como para matemáticas.	EEUU_NCES_NAEP trends_2013
EEUU	The mapmark standard setting method	Matthew Schulz y Howard Mitzel	Este documento describe el método de mapmark para definir los niveles de logro del NAEP 2005 del grado 12 de matemáticas.	EEUU_Schulz_Mapmark_sa
EEUU	Setting performance Standards on Educational Assessments and criteria for evaluating process.	Ronald Hambleton	Este documento trabaja los pasos seguidos para definir los estándares de desempeño de test de aprendizaje.	EEUU_Hambleton_Sett performance_sa
EEUU	NAEP Achievement Levels	Michael Ward, Susan Loomis y Christina Peterson	Esta presentación describe en términos generales el NAEP y en particular el desarrollo de los estándares.	EEUU_Ward_NAEP Achievement_sa

EEUU- New York	New York State Testing Program 2013: English Language Arts Mathematics Grades 3–8. Technical Report	Pearson	Este informe detallada información técnica de las pruebas de Lengua Inglesa Común (ELA) y Matemáticas 2013 del grado 3-8.	EEUU_NY_Pearson_Tech_Rep_ELA_2013
EEUU- New York	PARENTS' GUIDE TO ASSESSMENTS IN NEW YORK	National PTA	Este documento entrega información para padres sobre las nuevas evaluaciones en el estado de Nueva York para ELA y matemáticas.	EEUU_NY_NatPTA_Parents-Guide_ELA_2016
EEUU- New York	JANE DOE SAMPLE MIDDLE SCHOOL. ENGLISH LANGUAGE ARTS. SCHOOL 2015-2016 GRADE 6 TEST RESULTS	New York State Education Department	Este es un informe de resultados de la escuela que muestra resultados por estándares de desempeño y distintas comparaciones con agregados (distrito, estado).	EEUU_NY_EdDep_Sch_Report_2016
EEUU- New York	New York State Testing Program Common Core. Mathematics Test Performance Level Descriptions Grade 3	New York State Education Department	Este documento presenta los cuatro niveles de desempeño para el grado 3 de matemática.	EEUU_NY_EdDep_Stan_Math_ELA-3_2014
EEUU- New York	Programa de Evaluación del estado de Nueva York Contenidos Básicos Comunes de 3.º a 8.º Evaluación de Artes del idioma inglés. Comprender los informes de puntaje de Artes del idioma inglés de los Contenidos Básicos Comunes de 3.º a 8.º	New York State Education Department	Este documento ayuda a comprender e interpretar los informes de puntaje ELA del año 2015.	EEUU_NY_EdDep_2015
EEUU - Virginia	Frequently asked questions about SOL testing	Virgina Department of Education	Este documento entrega respuestas a las preguntas más frecuentes sobre SOL evaluaciones.	EEUU_VA_DE_FAQ_2015
EEUU - Virginia	SOL Grade 3. Reading performance levels descriptors		Este documento describe los niveles de desempeño de los estudiantes en el examen de lectura en grado 3.	EEUU_VA_levels descriptors_2013

EEUU - Virginia	Virgina Standards of learning assessments. Technical Report 2014-2015 Administrator Cycle	Virginia Department of Education	Este programa entrega información sobre los antecedentes y características técnica del programa de evaluación del estado de Virginia.	EEUU_VA_DE_St Learning_2015
EEUU - Virginia	English standards of learning for Virginia Public School	Board of Education Commonwealth of Virginia	Este documento describe los estándares de aprendizaje para inglés adoptados por el estado de Virginia.	EEUU_VA_BEV_Engl standards_2010
EEUU - Virgina	VA State Profile. Virginia Standards of Learning (SOL) End-of-course exams.	Center on Education Policy	Documento que define las características y parámetros de la prueba al final de grado 12.	EEUU_VA_CEP_State profile_2010
EEUU - Virgina	Virginia. Profile of State high school exit exam policies.	Center on Education Policy	Documento que define las características y parámetros de la prueba al final de grado 12.	EEUU_VA_CEP_Profil e State_2011
EEUU - Virginia	Virginia Standards of Learning (SOL) tests. Cut scores as adopted by the Virginia board of Education	Virginia board of education	Este documento informa sobre los puntos de corte de las pruebas SOL.	EEUU_VA_VBE_Cut scores_2016
Escocia	Assessment for Curriculum for Excellence. Strategic Vision Key Principles	Scottish Government	Este documento establece la visión estratégica del Gobierno de Escocia para la evaluación dentro de su marco de referencia, Curriculum for Excellence.	Esc_Sc-Gov_Ass_CfE_SA
Escocia	Curriculum for excellence. Building the curriculum 5a framework for assessment	Scottish Government	Guía profesionales sobre los estándares y expectativas, las habilidades y resultados esperados por el 'Curriculum of Excellencel' .	Esc_Sc-Gov_Asses_Frame_20 11
Escocia	Curriculum for excellence. Building the curriculum 5a framework for assessment	Scottish Government	Guía profesionales sobre los estándares y expectativas, las habilidades y resultados esperados por el 'Curriculum of Excellencel' .	Esc_Sc-Gov_Asses_Frame_20 11
Escocia	Scottish Survey of Literacy and Numeracy 2015 (Numeracy)	NS (National Statistics Scotland)	Informe de resultados nacionales de la prueba SSLN 2015 de numeracy.	Esc_NS_Inf-Res_SSLN_Math_201 6

Escocia	Scottish Survey of Literacy and Numeracy (SSLN) 2015. Survey Design Document	SA	Entrega información técnica sobre la aplicacion y analisis de la prueba SSLN 2015	Esc_Surv_Desig_SSLN_2015
España	EVALUACIÓN GENERAL DE DIAGNÓSTICO 2009. MARCO DE LA EVALUACIÓN	Instituto de Evaluación (IE)	Presenta el marco de las evaluaciones muestrales de diagnóstico del sistema educativo y el marco de referencia de evaluaciones de diagnóstico de las escuelas.	Esp_IE_Marc_Ev_EGD 2009_2009
España	Evaluación general de diagnóstico 2009. Educación Primaria. Cuarto curso. INFORME DE RESULTADOS	Instituto de Evaluación (IE)	Informe de resultados nacionales de la Evaluación General Diagnóstica de 4 primaria del año 2009.	Esp_IE_Inf_Res_4pri_ EGD2009_2010
España	Evaluación general de diagnóstico 2010. Educación secundaria obligatoria. Segundo curso. INFORME DE RESULTADOS	Instituto de Evaluación (IE)	Informe de resultados nacionales de la Evaluación General Diagnóstica de 2 secundaria del año 2010.	Esp_IE_Inf_Res_2sec_ EGD2010_2011
Holanda	OECD Thematic Review on Migrant Education Country Background Report for the Netherlands	OECD	Proporciona información y datos evaluativos sobre la educación de los migrantes en los Holanda.	Hol_OECD_Back_Rep ort_2009
Holanda	Een nadere beschouwing. Over de drempels met taal en rekenen	SA	Discusión sobre los esperado en Lenguaje y matemática.	Hol_SA_Een_nadere_ 2009
Holanda	HOW TO MEASURE AND EXPLAIN ACHIEVEMENT CHANGE IN LARGE-SCALEASSESSMENTS : A REJOINDER	Hickendorff y otros	Se discute temas de validez y desafíos metodológicos en evaluaciones de gran escala para decidir qué y cómo medir, y la manera de fomentar la estabilidad. Se analiza el caso de Holanda.	Hol_Hickendorff_Achi ev-Change_2009
Holanda	World Data on Education. VII Ed 2010/2011	Unesco	Presenta una revisión detallada de la educación primaria y secundaria holandesa.	Hol_Unesco_Netherla nds_2012
Holanda	OECD Reviews of Evaluationand Assessment in	Nusche y otros	Informa sobre sistema educativo nacional, marco de evaluación de estudiantes, docentes, escuelas y	Hol_Nusche_OECD-Netherlands_2014

	Education: Netherlands		sistema, fortalezas, desafíos y recomendaciones.	
Holanda	Referentiekader taal en rekenen. De referentieniveaus (Marco de referencia Lenguaje y Matemática. Niveles de referencia)	Ministerie van OCW (Ministerio de Educación Holanda)	Presenta y describe los niveles de referencia para lenguaje y matemática y sus ejes curriculares.	Hol_MiEd_Niv_Refer_2009
Holanda	OECD Review of evaluation and assessment in Education	OECD	Revisión del sistema de evaluación y monotoreo del sistema educativo holandes.	HL_OECD_Review of Eval_2014
Holanda	OECD Review of evaluation and assessment. Framework for improving school outcomes	Jaap Sheerense, Melanie Ehren, Peter Sleegers y Renske de Leeuw	Documento elaborado por la Universidad de Twente como base para el documento elaborado por la OECD	HL_Sheerense_Ed Evaluation_2012
Holanda	The End of Primary School Test	Marleen van der Lubbe	Entrega información sobre Cito test, resultados y uso.	Hol_van-der-Lubbe_Cito_test_NA
Holanda	COUNTRY BACKGROUND REPORT FOR THE NETHERLANDS	Jaap Scheerens y otros	Entrega información detalla sobre el sistema educacional holandés y de la evaluación en todos sus niveles.	Hol_Scheerens_Back_report_2012
Holanda	UPDATE TO THE COUNTRY BACKGROUND REPORT FOR THE NETHERLANDS	Jaap Scheerens y otros	Actualización de información entregada en 2012 sobre el sistema educacional holandés y de la evaluación en todos sus niveles.	Hol_Scheerens_Back_report_2013
Holanda	Netherlands: Quality Assurance in Early Childhood and School Education	Eurydice	Informa sobre el sistema de aseguramiento de la calidad de educacion pre-escolar y regular.	Hol_Eurydice_Qual_Assu_Netherlands_2015
Inglaterra	School inspection handbookHandbook for inspecting schools in England under section 5 of the Education Act 2005	Ofsted	Describe información general sobre la inspeccion de escuelas, los indicadores que se consideran y el uso de los resultados.	Ing_Ofsted_Inspec_Handb_2016
Inglaterra	Setting the grade standards of new	Ofqual	Presenta información sobre nueva metodología de establecimiento de	Ing_Ofqual_Stan_sett_GCSE_2016

	GCSEs in England – part 2		estándares para las pruebas GCSE. Material para uso de consulta pública.	
Inglaterra	Standards and Testing Agency. Annual Report and Accounts. For the year ended 31 March 2016	STA (Standard and Testing Agency)	Presenta información para rendir cuentas al parlamento sobre acción de STA.	Ing_STA_Ann_Report_2016
Inglaterra	Standards and Testing Agency. Business plan 1 April 2015-31 March 2016	STA (Standard and Testing Agency)	Presenta información anual sobre actividades de STA asociadas a las evaluaciones en todos los niveles que le corresponden.	Ing_STA_Buss_plan_2016
Mexico	Manual técnico. Establecimiento de nivles de competencia.	Instituto Nacional para la Evaluación de la Educación. Dirección de pruebas y medición	Documento oficial para la interpretación de los niveles de logro de los examenes para la calidad y el logro educativo.	MX_INEE_Niv Comp_2006
Mexico	Marco de referencia del examen de la calidad y el logro educativos (Excale). Español, tercer grado de educación primaria.	Instituto Nacional para la Evaluación de la Educación.	Este documento difunde el Marco de Referencia que da sustento al Examen de la calidad y el logro educativo (Excale) de la asignatura español de tercer grado de educación primaria.	MX_INEE_Marc Ref Espanol_2010
Mexico	Manual de Procedimientos. Dirección de pruebas y medición	Instituto Nacional para la Evaluación de la Educación. Dirección de pruebas y medición.	Este documento establece los procedimientos que deben observar las areas participantes en la construcción de la Excale.	MX_INEE_Proced Dir pruebas_2005
México	Manual Técnico. Diseño de Examenes de la calidad y el logro educativos.	Instituto Nacional para la Evaluación de la Educación.	Documento que establece el marco de referencia de cada excale, fundamentación teórica y definición de dimensiones.	MX_INEE_Ex de Calidad_2009
Nueva Zelanda	The Development of the Student Assessment System in New Zealand	Lester Flockton	Este documento detalla los cambios y el estado actual de las evaluaciones nacionales (monitoreo de estándares), exámenes (calificaciones) y evaluación de aula (basada en la escuela).	NZ_Flockton_Stu_Ass_Syst_2012

Nueva Zelanda	NMSSA 2016 Information for schools, parents, whānau and caregivers	EARU (Educational Assessment Research Unit)	Este documento presenta información general sobre la NMSSA, tales como propositos, características de la prueba, muestra y resultados.	NZ_EARU_Presentacion_NMSSA_2016
Nueva Zelanda	Achievement in English: reading. Summary of results from the 2014 National Monitoring Study of Student Achievement for teachers and principals	EARU (Educational Assessment Research Unit)	Este documento presenta información general sobre la NMSSA, tales como propositos, características de la prueba, muestra y resultados.	NZ_EARU_Inf_escuela_NMSSA_2014
Nueva Zelanda	Wanangatia te Putanga Tauira. National Monitoring Study of Student Achievement. Technical Information 2014. Social Studies, English: Reading	EARU (Educational Assessment Research Unit)	Este documento entrega información técnica del estudio y sus resultados 2014. Hay información sobre alineamiento curricular de áreas evaluadas NMSSA 2014	NZ_EARU_Inf_Tecn_2014
Nueva Zelanda	Wanangatia te Putanga Tauira. National Monitoring Study of Student Achievement. English: Reading 2014 Overview	EARU (Educational Assessment Research Unit)	Este documento entrega información general de los resultados del estudio de lectura en inglés	NZ_EARU_Inf result_4y8prim_NMSSA_2015
Perú	Preguntas frecuentes sobre la Evaluación Censal de Estudiantes ECE	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento es el disponible online para responder preguntas frecuentes al público general sobre distintos aspectos de la prueba.	Per_UMC_Preg_frec_2016
Perú	REPORTE TÉCNICO DE LA EVALUACIÓN CENSAL DE ESTUDIANTES (ECE 2015) SEGUNDO Y CUARTO (EIB) DE PRIMARIA Y SEGUNDO DE SECUNDARIA	MEP (ministerio de Educación de Perú)	Este documento detalla información sobre procesos de construcción de las pruebas, la muestra, la aplicación y las estrategias de análisis psicométrico	Per_MEP_Inf_Tec_ECE_2015

Perú	EVALUACIÓN DE LOS APRENDIZAJES DE LOS ESTUDIANTES EN LA EDUCACIÓN BÁSICA REGULAR	MEP (ministerio de Educación de Perú)	Este documento describe la normativa de evaluación para educación inicial, primaria y secundaria peruana.	Per_MEP_Eva_Ed-Regular_2005
Perú	Informe para el docente. Qué logran nuestros estudiantes en Lectura 2 grado Primaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a profesores y muestra resultados nacionales y de escuela en prueba ECE 2do primaria Lectura	Per_UMC_Inf_Prof_ECE-Lec_2prim_2015
Perú	Informe para el docente. Qué logran nuestros estudiantes en Matemática 2 grado Primaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a profesores y muestra resultados nacionales y de escuela en prueba ECE 2do primaria Matemática	Per_UMC_Inf_Prof_ECE-Mat_2prim_2015
Perú	REPORTE TÉCNICO DE LA EVALUACIÓN CENSAL DE ESTUDIANTES (ECE 2014) SEGUNDO Y CUARTO (EIB) DE PRIMARIA Y SEGUNDO DE PRIMARIA	MEP (ministerio de Educación de Perú)	Este documento detalla información sobre procesos de construcción de las pruebas, la muestra, la aplicación y las estrategias de análisis psicométrico	Per_MEP_Inf_Tec_ECE_2014
Perú	REPORTE TÉCNICO DE LA EVALUACIÓN MUESTRAL 2013 DE ESTUDIANTES DE 6 GRADO PRIMARIA	MEP (ministerio de Educación de Perú)	Este documento detalla información sobre procesos de construcción de las pruebas, la muestra, la aplicación y las estrategias de análisis psicométrico	Per_MEP_Inf_Tec_ECE_6prim_2014
Perú	Informe para el docente. Qué logran nuestros estudiantes en Matemática? 2 grado Secundaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a profesores y muestra resultados nacionales y de escuela en prueba ECE 2do secundaria Matemática.	Per_UMC_Inf_Prof_ECE/Mat_2sec_2015
Perú	Informe para el docente. Qué logran nuestros estudiantes en Lectura? 2 grado Secundaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a profesores y muestra resultados nacionales y de escuela en prueba ECE 2do secundaria Lectura.	Per_UMC_Inf_Prof_ECE-Lect_2sec_2015
Perú	Informe para el docente. Qué logran nuestros estudiantes en Escritura 2 grado Secundaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a profesores y muestra resultados nacionales y de escuela en prueba ECE 2do secundaria Escritura	Per_UMC_Inf_Prof_ECE-Esc_2sec_2015

Perú	Sistematización del proceso de elaboración de los estándares de aprendizaje nacionales	Jessica Tapia Soriano	Este documento sistematiza el proceso de elaboración de los estándares nacionales, sus fundamentos, potencialidades y principales riesgos.	Per_Tapia_Proc_Elab_Estan_2015
Perú	Evaluación Censal de Estudiantes (ECE) Segundo grado de primaria. Cuarto grado de primaria de IE EIB. MARCO DE TRABAJO	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento presenta el marco de trabajo de la Evaluación Censal de Estudiantes (ECE) para 2do y 4to básico.	Per_UMC_Mar_Ev_ECE_2y4bas_2009
Perú	¿Qué logran nuestros estudiantes en la ECE? 2.º grado de Primaria	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento está dirigido a las escuelas y muestra resultados nacionales y de escuela en prueba ECE 2do primaria.	Per_UMC_Inf_Esc_ECE-2prim_2015
Perú	Informe para Docentes y Directores. Qué logran nuestros estudiantes en Lectura? 4 grado de Primaria EIB. Castellano como segunda lengua	UMC (Oficina de la Medición de la Calidad de la Educación)	Este documento presenta los niveles de logro y los resultados la escuela en ECE 2015 en Lectura en castellano como segunda lengua en 4.° grado EIB (Educación Intercultural Bilingüe).	Per_UMC_Inf_Esc_ECE-4primEIB_2015