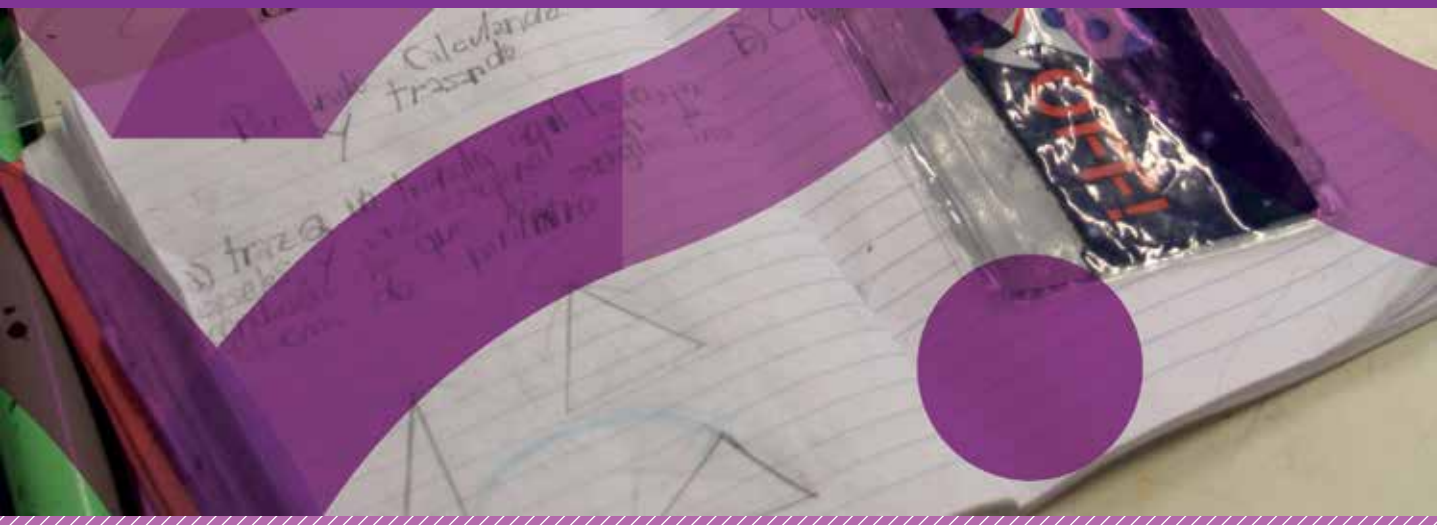
ARISTAS

REPORTE 6

¿CÓMO SON LAS EVALUACIONES NACIONALES DE OTROS PAÍSES?



INEEd
Instituto Nacional de Evaluación Educativa

Aristas
Evaluación Nacional de Logros Educativos

La elaboración de este documento estuvo a cargo de: Federico Burgell, Raisa López y Andrea Rajchman.

Corrección de estilo: Mercedes Pérez y Federico Bentancor
Diseño y diagramación: Diego Porcelli
Fotos de tapa: INEEd

Montevideo, 2022
ISSN: 2697-2786

Edificio Los Naranjos, Planta Alta, Parque Tecnológico del LATU
Av. Italia 6201, Montevideo, Uruguay
(+598) 2604 4649 – 2604 8590
ineed@ineed.edu.uy
www.ineed.edu.uy

Cómo citar: INEEd. (2022). *Reporte de Aristas 6*. ¿Cómo son las evaluaciones nacionales de otros países? Recuperado de https://www.ineed.edu.uy/images/Aristas/Publicaciones/Reportes/Reporte-6-Como-son-las-evaluaciones-nacionales-de-otros-paises.pdf

En la elaboración de este material se ha buscado que el lenguaje no invisibilice ni discrimine a las mujeres y, a la vez, que el uso reiterado de /o, /a, los, las, etcétera, no dificulte la lectura.

INTRODUCCIÓN

Las evaluaciones estandarizadas nacionales de logros educativos son una de las formas de acceder a información sobre la dinámica de los procesos y resultados en los sistemas educativos de distintos países. Algunos de los principales aportes de la participación en ellas consisten en que permiten visualizar los resultados educativos de los estudiantes en forma conjunta, proporcionan información sobre el acceso real al conocimiento y a las capacidades alcanzadas por los estudiantes, ayudan a visibilizar aspectos centrales de la labor educativa, aportan información importante sobre diversos actores sociales del sistema educativo, y ponen en discusión qué aspectos del currículo intencional son exigibles a todos los estudiantes en cada ciclo educativo (Ravela et al., 2008).

Este reporte busca generar insumos para la mejora de la evaluación Aristas que aplica el Instituto y realizar aportes al sistema educativo uruguayo en relación con la operacionalización y medición del currículo intencional. Se espera que la exploración permita sistematizar información acerca de cómo se desarrolla en otros países la relación entre el currículo intencional y la tabla de dominios de las pruebas.

RELEVAMIENTO REALIZADO

Durante la década de 1990 los sistemas nacionales de evaluación estandarizada de la región se desarrollaron con especial ímpetu. Además, en 1994 se constituyó la red de sistemas de evaluación en torno al Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación (LLECE), en la que participan casi todos los países de la región. Varios participan también en el Programa para la Evaluación Internacional de Alumnos (PISA, por su sigla en inglés) y uno en el Estudio Internacional de Tendencias en Matemáticas y Ciencias (TIMSS, por su sigla en inglés) y en el Estudio Internacional de Progreso en Comprensión Lectora (PIRLS, por su sigla en inglés).

En este contexto, algunos de los cambios y mejoras más relevantes en los sistemas nacionales de evaluación de los países de América Latina son: mayor transparencia en la difusión y el uso de los resultados; tendencia a pasar de pruebas normativas a pruebas criteriales; mejoras en las capacidades técnico metodológicas para la construcción de pruebas y procesamiento de datos; mayor preocupación por los factores asociados a los desempeños; creciente participación en las evaluaciones internacionales, entre otros (Ravela et al., 2008). A partir de estos cambios, las evaluaciones nacionales de aprendizajes de los países de la región, con el objetivo de monitorear el logro de los aprendizajes establecidos en los documentos curriculares y favorecer mejoras en ellos, se han constituido en herramientas centrales de la política educativa (Manzi, García y Taut, 2019).

En el resto del mundo también se han realizado importantes avances en lo referido a evaluaciones estandarizadas nacionales e internacionales. En particular, hay una gran presencia de países que aplican pruebas de logro educativo a los alumnos, para conocer en qué medida se alcanzan los conocimientos y competencias que se espera que estos hayan adquirido al finalizar un ciclo o nivel educativo. Estas evaluaciones proporcionan información que permite evaluar el sistema educativo en su conjunto y a las políticas educativas vigentes. Su aporte principal radica en la información que proporcionan para la toma de decisiones en materia de política educativa (Ravela, 2006).

En este contexto, varios investigadores (Ferrer, 2006; Murillo y Román, 2010) refieren a la importancia de que los sistemas educativos definan el propósito de sus sistemas de evaluación, así como el uso que harán de los resultados que aportan.

Tomando como objetivo la exploración acerca de cómo se da la relación entre el currículo intencional y la definición conceptual y operacional de las evaluaciones nacionales de otros países, se analizó información acerca de las pruebas nacionales de siete:

1. Sistema de Medición de la Calidad de la Educación (Simce) – Chile
2. Aprender – Argentina
3. Evaluación Censal de Estudiantes (ECE) – Perú
4. Saber – Colombia
5. Sistema de Evaluación de la Educación Básica (Saeb, por su sigla en portugués) – Brasil
6. National Assessment of Educational Progress (NAEP) – Estados Unidos
7. National Assessment Program – Literacy and Numeracy (NAPLAN) - Australia

Para hacer la selección, se procuró que hubiera cinco países de la región, considerando que la mayoría de los institutos nacionales de evaluación de América Latina se han desarrollado principalmente en los últimos años, en forma similar al desarrollo del Instituto Nacional de Evaluación Educativa (INEEd). Además, se incluyeron en el análisis las evaluaciones nacionales de Estados Unidos y de Australia, ya que son dos países en los que el sistema de evaluación tiene vasta trayectoria[1].

La información recogida sobre estas pruebas fue la siguiente:

1. Información general de la prueba:
    a. áreas en la que se aplica;
    b. grados;
    c. frecuencia con que se hacen las evaluaciones y justificación de la frecuencia, y
    d. ¿la prueba es censal o muestral?
2. Diseño y usos de la prueba:
    a. ¿el diseño es un único cuadernillo?, ¿matricial con cuadernillos?, ¿adaptativa?;
    b. ¿para qué se usan los resultados?, ¿la prueba tiene consecuencias?
3. Constructo de la prueba:
    a. ¿cómo se operacionaliza el constructo?;
    b. ¿tiene dominios definidos por los documentos curriculares?, y
    c. ¿usa el diseño de test basados en evidencias?
4. Cantidad y tipo de ítems:
    a. cantidad de ítems por prueba, y
    b. tipos de ítems (múltiple opción simple, complejos, emparejamiento, abiertos breves, abiertos de desarrollo).
5. Pilotaje de ítems:
    a. ¿se pilotean los ítems?, y
    b. ¿se hace un operativo piloto previo?
6. Niveles de desempeño:
    a. cantidad de niveles, y
    b. ¿algún nivel fija el estándar mínimo?
7. Proceso de equiparación:
    a. ¿cuántos ítems se usan de anclaje?, ¿qué características tienen?;
    b. ¿cómo se eligen los ítems de anclaje?

[1] Los países incluidos en el análisis fueron seleccionados para representar las decisiones conceptuales y metodológicas de algunas evaluaciones. No hubo intención de indicar que representan a todos los sistemas nacionales de evaluación ni que son los más importantes para el análisis o consideración.

Del relevamiento realizado, se encontró que en los siete casos analizados se evalúan las áreas de matemática, lectoescritura y ciencias sociales o competencias ciudadanas. Además, en algunas también se evalúan ciencias naturales, tecnología, artes y economía.

En cuanto a los grados en que se aplican las pruebas, la de Argentina se realiza en dos grados (al finalizar primaria y secundaria), las de Perú y Estados Unidos en tres (segundo y cuarto de primaria, y segundo de media; y en los grados cuarto, octavo y doceavo, respectivamente), la de Brasil en cuatro grados (segundo, quinto y noveno de primaria, y tercero de media), y las de Chile, Colombia y Australia en cinco grados o más.

La mayoría de las evaluaciones analizadas se aplican anual o bianualmente. Además, algunos de los institutos llevan a cabo algunas de sus evaluaciones también cada tres o cuatro años.

Se realizan tanto muestras como censos y, en algunos casos, ambos. En Australia, por ejemplo, las evaluaciones de lectoescritura y matemática se aplican anualmente en formato censal, y las de las otras áreas se aplican trianualmente, con carácter muestral. En Chile, algunas de las evaluaciones sumativas pasaron de censales a muestrales. En los países que realizan censos, el objetivo principal es brindar información por centro educativo. Únicamente en Australia se entrega información a nivel de niño (aunque con pruebas que no son adaptativas).

En cinco de las evaluaciones relevadas el formato de las pruebas es matricial: Chile, Argentina, Brasil, Colombia y Estados Unidos. En Perú hay pruebas de los dos tipos (matricial y de cuadernillo único), y en Australia son de cuadernillo único. Además, cabe mencionar que ninguna de las siete evaluaciones es adaptativa, para dar cuenta del monitoreo de los logros a nivel nacional.

En lo que refiere a la cantidad de ítems por evaluación, en Argentina son 24 por cuadernillo y 72 en total, en Perú son 100 por prueba (pero no se explicita cuántos por cuadernillo), en Colombia entre 36 y 54 por cuadernillo, en Brasil 22, y en Australia entre 36 y 64, dependiendo el grado y área. Para las evaluaciones de Chile y Estados Unidos no se encontró información respecto a la cantidad de ítems por prueba, aunque en el segundo caso hay información acerca de la duración de tiempo que se otorga para cada prueba. En cuanto al tipo de ítems que se usan, en todas las evaluaciones se incluyen ítems de opción múltiple. Además, en algunas de ellas también hay de respuesta construida: Chile, Perú, Brasil, Estados Unidos y Australia. En el caso de Colombia todos los ítems son de múltiple opción. En Argentina no está disponible esa información. En Estados Unidos y Australia algunos de los ítems de opción múltiple incluyen tareas interactivas.

Para las siete evaluaciones relevadas se pilotean los ítems que se usan en las pruebas. En el caso de Brasil no se aclara si se realiza o no un operativo específico con esta finalidad. En el resto de ellas se realiza un operativo piloto previo a la aplicación definitiva.

En cuanto a la operacionalización del constructo, los dominios de las siete evaluaciones están definidos por los documentos curriculares nacionales. En algunos casos un comité especial desarrolla los marcos de las evaluaciones como forma de interpretar los documentos

curriculares (Argentina, Perú y Estados Unidos), y en los otros los dominios reflejan los documentos curriculares sin necesidad de hacer interpretaciones. En las pruebas de Argentina, Perú y Estados Unidos (países en los que los marcos de las evaluaciones son interpretaciones del currículo), los marcos definen el constructo a evaluar a través de la explicitación de los contenidos específicos y los procesos cognitivos implicados. En los casos de Chile, Colombia, Brasil y Australia, en los que los marcos de las pruebas están explícitamente definidos por los documentos curriculares, los dominios (*blueprint*) están presentados a través de una tabla en la que, además, se explicita la proporción de cada dominio en la prueba, que se corresponde directamente con las proporciones presentes en los documentos curriculares.

De las siete pruebas relevadas, dos aplican la metodología de test basados en evidencias (Colombia y Brasil), dos no lo hacen (Chile y Estados Unidos) y las tres pruebas restantes (Argentina, Australia y Perú) no explicitan aplicar dicha metodología o cuál es la metodología empleada.

Respecto a la cantidad de niveles de desempeño de cada una de las pruebas, dos de ellas tienen tres (Chile y Estados Unidos) y en ambas el nivel 2 corresponde al estándar mínimo. Tres fijan cuatro niveles de desempeño (Argentina, Perú y Colombia) y en ninguna de ellas se explicita que alguno de ellos corresponda al estándar mínimo, aunque en Argentina, por la nomenclatura de los niveles, se podría asumir que el 2 es el mínimo aceptable (básico). En la evaluación de Australia se presentan progresiones de aprendizaje discriminadas en diez niveles, de los cuales seis corresponden a cada grado evaluado. Es decir, se definieron diez niveles de logro por área, que cubren desde los grados más bajos hasta los más altos del sistema educativo. En cada área y grado, de esos diez niveles los alumnos se distribuyen en seis, dentro de los cuales el segundo de cada grado corresponde al estándar mínimo. Por otra parte, en las evaluaciones de Brasil también se presentan progresiones de aprendizaje compuestas por entre ocho y diez niveles, dependiendo del área y grado, y también se fija un nivel que corresponde al estándar mínimo.

Por último, en lo que refiere a los ítems de anclaje entre distintas ediciones de las pruebas, solamente se encontraron referencias explícitas en los documentos de las evaluaciones de Chile y Colombia. En los documentos técnicos de las pruebas de Chile se explicita que se usa aproximadamente un 20% de ítems de anclaje, y que estos se eligen de forma tal que sean representativos de la matriz de la prueba, que cubran todo el rango de habilidad y que tengan buenos parámetros estadísticos. Además, solo se utilizan como anclaje ítems de opción múltiple. Por su parte, en las pruebas de Colombia se usan para el anclaje entre el 25% y 40% de los ítems, y en la equiparación se vuelven a calibrar los ítems de anclaje, para valorar si se emplean los nuevos parámetros o los anteriores.

Para las pruebas de Estados Unidos no se encontró el porcentaje ni las características de los ítems de anclaje, aunque sí se hallaron referencias respecto a los modelos de equiparación aplicados y el tipo de calibración empleado. En las pruebas de Australia tampoco se encontró referencias a los ítems de anclaje y su selección; sí hay referencias a la equiparación horizontal y vertical.

Síntesis

El relevamiento realizado muestra que las evaluaciones analizadas tienen distintas características. Se pudo observar que en los países en los que las definiciones curriculares son más claras, las pruebas están centradas en estas definiciones, por lo que el constructo no se basa en una interpretación de los documentos curriculares, sino que los considera directamente.

Tanto en estos casos como cuando los marcos consisten en una interpretación de los documentos curriculares, la operacionalización del marco se hace a través de tablas o matrices en las que se explicitan los contenidos a evaluar, así como las habilidades o procesos cognitivos a través de los cuales se abordan los contenidos que definen los dominios de cada evaluación. Además, en estos países los marcos presentan la proporción de cada dominio en la prueba en relación con la proporción de dichos dominios en los documentos curriculares. En algunos países también se operacionalizan los marcos utilizando test basados en evidencias.

En al menos seis de las siete pruebas analizadas se pilotean los ítems en un operativo previo a la evaluación definitiva. La cantidad de niveles de desempeño varía entre tres y diez niveles entre las distintas evaluaciones. Cuatro de ellas definen un nivel como el correspondiente al estándar mínimo expresado en el currículo.

SIMCE - CHILE

1.Información general de la prueba	Nacional	Chile.
	Áreas y grados	Actualmente, las pruebas Simce se aplican en los niveles segundo, cuarto, sexto y octavo de enseñanza básica y en segundo año de educación media. Abordan contenidos curriculares de las asignaturas de lenguaje y comunicación, matemática, ciencias naturales e historia, geografía y ciencias sociales. Sin embargo, no todos los años se aplican todas las pruebas en todos los niveles. La aplicación de cada año está determinada por un Plan de Evaluaciones.
	Frecuencia de aplicación	Anual.
	Censal o muestral	Pruebas censales. A partir del 2018 se incorporan otras evaluaciones (formativa y progresiva) y la evaluación en segundo de primaria pasa a ser muestral porque no se consideró adecuado que tuviera consecuencias a partir de ese grado.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	Matricial.
	¿Para qué se usan los resultados?	Las evaluaciones de carácter censal ofrecen información por establecimiento.
	¿La prueba tiene consecuencias?	Sí, para los establecimientos.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	Los marcos de las evaluaciones se basan en los objetivos de aprendizaje del currículo oficial. Cada prueba tiene tres dimensiones. En la prueba cubren cada dimensión con un porcentaje de ítems similar al de abordaje de dichas dimensiones en el currículo.
	¿Tiene dominios definidos por los documentos curriculares?	A partir de este marco de referencia se generan las especificaciones técnicas de las pruebas, esto es, los documentos que establecen las características particulares de diseño de los instrumentos, y que buscan garantizar cobertura y comparabilidad en la medición de los aprendizajes curriculares. Las especificaciones técnicas las elabora el Departamento de Construcción de Pruebas (DCP) de la Agencia de Calidad de la Educación y las revisa la Unidad de Currículum y Evaluación (UCE) del Ministerio de Educación. Las especificaciones técnicas de las pruebas distinguen y definen ejes temáticos (o de contenidos) y ejes de habilidades (o de dominios cognitivos) y además establecen las matrices de evaluación teóricas, que señalan los porcentajes de preguntas con los que serán representados los ejes en la prueba. Estos porcentajes se definen a partir de un análisis de énfasis curriculares que realizan en conjunto profesionales de la Agencia y de la UCE.
	¿Usa el diseño de test basados en evidencias?	Las preguntas de las pruebas Simce se construyen a partir de objetivos de evaluación que corresponden a una operacionalización de los ejes descritos en el apartado anterior.
4.Cantidad y tipo de ítems	Cantidad de ítems por prueba.	No se explicita.
	Tipos de ítems	La mayoría de los ítems son de múltiple opción y algunos de respuesta abierta.

5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí. El proceso de elaboración de las preguntas de las pruebas Simce 2015 se llevó a cabo mediante la modalidad de contratación directa de elaboradores externos por área y nivel escolar evaluado. Una vez elaborados los conjuntos de preguntas para cada una de las pruebas, los ítems fueron distribuidos en cuadernillos para su aplicación experimental conducente a su validación.
	¿Se hace un operativo piloto previo?	Se realiza piloto un año antes de las evaluaciones definitivas. Las pruebas experimentales se aplican el año anterior a la aplicación censal definitiva, a una muestra representativa de la población evaluable. El análisis de estas pruebas permite obtener información cuantitativa y cualitativa de los ítems para determinar si poseen la suficiente calidad como para conformar una prueba Simce definitiva. En el caso de las preguntas abiertas, la información permite validar las pautas y rúbricas, y elaborar los manuales de corrección. En el caso de las preguntas de opción múltiple, el proceso permite obtener información sobre la calidad de los enunciados y las opciones. El resultado de los análisis cuantitativos y cualitativos da origen a un conjunto de preguntas que, sumadas a otras provenientes de aplicaciones previas, conformarán las pruebas definitivas. Todas estas preguntas deben cumplir con los requisitos de representatividad y cobertura curricular, cobertura del rango de habilidades de los estudiantes y alineación a los estándares de aprendizaje, en los niveles en que existan estándares.
6. Niveles de desempeño	Cantidad de niveles	Tres.
	¿Algún nivel fija el estándar mínimo?	Sí, los niveles de aprendizaje adecuado y elemental están asociados a una lista de requisitos mínimos que detalla los aprendizajes que, al menos, debe demostrar un estudiante en la prueba Simce para alcanzar cada uno de ellos. El nivel de aprendizaje insuficiente, por su parte, no cuenta con requisitos mínimos, sino que agrupa a aquellos estudiantes que no demuestran de manera consistente el logro de los requisitos mínimos para alcanzar el nivel elemental.
7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	Un 20% de ítems es de anclaje.
	¿Qué características tienen?	Se usan de anclaje solo ítems de múltiple opción.
	¿Cómo se eligen los ítems de anclaje?	La selección se realiza en función de una serie de criterios que aseguren su calidad: que cuenten con parámetros TRI adecuados, que no presenten problemas de construcción, que no estén liberados (ni asociados a un estímulo liberado), y que en su conjunto abarquen el rango de dificultad de la prueba y sean representativos de la matriz teórica de la prueba. La primera acción del armado consiste en seleccionar el conjunto de ítems de equiparación entre años (*equating* entre años) para cada prueba. Estos conjuntos conforman subpruebas de las mediciones de origen, esto es, una representación del total de ítems de cada evaluación, en términos de constructo y características psicométricas. Estos ítems de *equating* corresponden a cerca del 20% de las preguntas de la prueba de la medición anterior.
8. Observaciones o comentarios	Los ítems de anclaje deben estar en el mismo orden relativo en las pruebas en las que se usan. Además, no deben variar su posición en más de cinco lugares. Los demás ítems se reparten en los cuadernillos, de forma que se respeten los siguientes criterios: • cada cuadernillo debe cubrir la tabla de dominios; • cada cuadernillo debe tener similar carga cognitiva; • los primeros ítems de cada forma deben ser sencillos y motivadores, y los de mayor dificultad o más complejos deben estar en posiciones intermedias; • los ítems de respuesta abierta no deben ubicarse en las posiciones finales; • las claves (letras de opciones correctas) deben estar distribuidas de manera homogénea al interior de cada prueba y se debe asegurar que no existan patrones, y • si hubiera ítems dependientes o encadenados, se deben ubicar en formas distintas, para garantizar la independencia local de las preguntas. Una vez finalizado el posicionamiento de las preguntas en cada forma, se evalúa que las formas exhiban una función de información y una curva característica similares entre sí y, en caso de detectar falencias en este ámbito, se realizan las modificaciones pertinentes.	
9. Enlaces a reportes técnicos e informes de resultados	http://archivos.agenciaeducacion.cl/Informe_Tecnico_SIMCE_2015_Final.pdf http://archivos.agenciaeducacion.cl/Sistema_Nacional_de_Evaluacion_17abr.pdf	

PRUEBAS APRENDER – ARGENTINA

1. Información general de la prueba	Nacional	Argentina.
	Áreas y grados	Las pruebas Aprender reemplazan al Operativo Nacional de Evaluación (ONE) desde 2016. Se presenta como una evaluación nacional de carácter estandarizado que mide los logros de aprendizaje de los estudiantes que están por finalizar los niveles primario y secundario en áreas básicas de conocimiento como son matemática, lengua, ciencias naturales y ciencias sociales. El operativo Aprender se caracteriza por diseñar pruebas que toman como marco de referencia las definiciones curriculares nacionales, provinciales y de la Ciudad Autónoma de Buenos Aires.
	Frecuencia de aplicación	Hasta la fecha, Aprender se aplicó de manera consecutiva entre los años 2016 y 2019.
	Censal o muestral	En un primer momento, las pruebas fueron anuales y censales, tanto en el nivel primario como en el secundario, pero desde 2018 se ejecutan privilegiando un nivel de enseñanza por año.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	Matricial.
	¿Para qué se usan los resultados?	Reportes por escuela, jurisdicción y nacionales.
	¿La prueba tiene consecuencias?	No.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	La evaluación Aprender, siguiendo la línea de trabajo iniciada con el ONE, es una prueba referida a criterios. Estas pruebas buscan conocer los contenidos y capacidades que los estudiantes dominan, a través de un conjunto de ítems relevantes y representativos de la disciplina evaluada. Las pruebas de criterio privilegian la comparación de los logros de los estudiantes con respecto a los desempeños esperados, fijados en los núcleos de aprendizajes prioritarios (NAP) y en los diseños curriculares jurisdiccionales.
	¿Tiene dominios definidos por los documentos curriculares?	Los NAP son una fuente clave en el proceso de evaluación, ya que determinan un marco de aprendizajes mínimos que todos los estudiantes deberían adquirir y, por lo tanto, sobre los cuales se construyen los instrumentos de evaluación, definidos junto a los diseños curriculares de cada provincia y los consensos jurisdiccionales, en un trabajo articulado de la Red de Evaluación Federal para la Calidad y Equidad Educativa (REFCEE). Para cada prueba, se definen los procesos cognitivos a relevar y los contenidos correspondientes.
	¿Usa el diseño de test basados en evidencias?	No se explicita.
4. Cantidad y tipo de ítems	Cantidad de ítems por prueba	La prueba Aprender 2019, para cada área/disciplina, quedó conformada por un total de 72 ítems, que se distribuyeron en seis modelos de pruebas con dos bloques de 12 ítems cada uno. Los bloques fueron encadenados de modo tal que ocupasen, alternativamente, la primera y la segunda posición en la prueba, conformando un total de 24 ítems para cada disciplina de quinto y sexto año de nivel secundario.
	Tipos de ítems	No se explicita.

5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	Sí.
6. Niveles de desempeño	Cantidad de niveles	Se distinguen cuatro niveles: 1) por debajo del nivel básico, 2) básico, 3) satisfactorio y 4) avanzado.
	¿Algún nivel fija el estándar mínimo?	No se especifica. Por la nomenclatura, podría ser el nivel 2.
7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	No se especifica.
	¿Qué características tienen?	
	¿Cómo se eligen los ítems de anclaje?	
8. Observaciones o comentarios	Aprender 2019 inició su proceso con el diseño y aplicación de la prueba piloto, que fue implementada en abril del mismo año y sirvió de base para la construcción de las versiones definitivas de los instrumentos administrados durante setiembre. En función de los resultados de la prueba piloto, se realizó un proceso de análisis que tuvo como objetivo mejorar la validez de los instrumentos finales, mediante la eliminación de aquellos ítems que mostraron un funcionamiento anómalo.	
9. Enlaces a reportes técnicos e informes de resultados	https://www.argentina.gob.ar/sites/default/files/evaluacion_educacion_secundaria_argentina_2019.pdf https://www.argentina.gob.ar/educacion/evaluacion-informacion-educativa/aprender	

PRUEBAS ECE – PERÚ

1. Información general de la prueba	Nacional	Perú.
	Áreas y grados	La ECE informa en qué medida los alumnos están logrando los aprendizajes que se espera de ellos según el currículo oficial. Se evalúa comunicación (lectura y escritura) y matemática en segundo y cuarto de primaria, y segundo grado de secundaria; también se evalúa historia, geografía y economía en segundo grado de secundaria. Comunicación y matemática se evalúan anualmente en todos los grados, mientras que historia, geografía y economía se evalúan en forma bianual. La periodicidad y las características del diseño permiten trazar tendencias a lo largo del tiempo para saber si los resultados están mejorando o si las brechas entre las distintas poblaciones de estudiantes se están reduciendo.
	Frecuencia de aplicación	No se explicita.
	Censal o muestral.	Censal.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	En la ECE se utilizan, dependiendo del grado y las características de la competencia a evaluar, dos tipos de diseño de pruebas: una en la que se aplica un mismo conjunto de ítems a toda la población (de "forma única") y otra en la que existen varias "formas" o versiones de la prueba, que son equivalentes entre sí porque miden la misma competencia.
	¿Para qué se usan los resultados?	La ECE ha sido concebida para ser censal porque se pretende devolver resultados a los distintos actores de las instancias de gestión educativa: directores y docentes de todas las escuelas, especialistas y autoridades de las Unidades de Gestión Educativa Local (UGEL), las Direcciones Regionales de Educación (DRE) y el Ministerio de Educación (Minedu), con el fin de movilizarlos hacia la mejora de los aprendizajes, tanto en las áreas como en las competencias evaluadas. Se realiza en todas las escuelas públicas y privadas del país que tengan más de cinco estudiantes en el grado a evaluar.
	¿La prueba tiene consecuencias?	No se explicita.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	Para cada prueba, se define el constructo a evaluar y el modelo de evaluación. En lectura, por ejemplo, se relevan capacidades, textos y contextos. En matemática, se relevan capacidades, contenidos y contextos.
	¿Tiene dominios definidos por los documentos curriculares?	En el caso de la ECE, los constructos y la lógica de su progresión están bien asentados en la literatura existente, tanto en los documentos curriculares (Diseño Curricular Nacional, Mapas de Progreso del Aprendizaje) como en otros documentos de uso pedagógico, como las Rutas del Aprendizaje.
	¿Usa el diseño de test basados en evidencias?	No se explicita.

4. Cantidad y tipo de ítems	Cantidad de ítems por prueba	Dentro del área de comunicación la prueba de lectura consta de 100 ítems en cada uno de los niveles. Se realiza, además, una prueba de escritura en segundo grado de secundaria, en la que los estudiantes deben redactar cuatro textos (dos narrativos, uno descriptivo y uno argumentativo). En el área de matemática las pruebas constan de 100 ítems en cada uno de los niveles. En el área historia, geografía y economía la prueba también tiene 100 ítems.
	Tipos de ítems	Se distinguen dos tipos de ítems: de opción múltiple y de respuesta construida. En algunos casos los ítems cuentan con crédito parcial. En ciertas ocasiones, los ítems cuentan con un estímulo común (en el caso de lectura esto siempre es así).
5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	La construcción de los ítems es puesta a prueba tanto en procesos de juicio experto como en aplicaciones de campo. La aplicación de campo permite obtener evidencias de validez vinculadas a la estructura interna de las mediciones realizadas, esperando que estas sean unidimensionales, y posibilita detectar posibles sesgos en la construcción de los ítems, evidenciados por su funcionamiento diferencial.
6. Niveles de desempeño	Cantidad de niveles	Se consideran cuatro niveles de logro: satisfactoria, en proceso, en inicio y previo al inicio.
	¿Algún nivel fija el estándar mínimo?	No se especifica.
7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	No se explicita.
	¿Qué características tienen?	No se utilizan procedimientos para estimar la confiabilidad directamente (por ejemplo, test-retest), ya que se dificulta su aplicación en evaluaciones estandarizadas masivas. Sí se utilizan métodos indirectos que calculan la consistencia interna sobre la base del promedio de las correlaciones entre los ítems o como una derivación del cociente entre la sumatoria de varianzas de los ítems sobre la varianza total (en teoría clásica) o de la separación de las personas (en el caso de los modelos Rasch).
	¿Cómo se eligen los ítems de anclaje?	No se explicita.
8. Observaciones o comentarios		
9. Enlaces a reportes técnicos e informes de resultados	http://umc.minedu.gob.pe/wp-content/uploads/2017/12/Marco-de-Fundamentaci%C3%B3n-ECE.pdf	

PRUEBAS SABER – COLOMBIA

1. Información general de la prueba	Nacional	Colombia.
	Áreas y grados	Lenguaje, matemática y competencias ciudadanas, en tercero y quinto de primaria, y en noveno de educación media. A partir del 2021 se incorporan también pruebas en séptimo de primaria.
	Frecuencia de aplicación	Las nuevas pruebas inician la serie en 2021, y la idea es evaluar las trayectorias educativas de los estudiantes desde tercero y quinto, cada dos años, con escalamiento vertical. Desde 2023 se incluirá también a los estudiantes de séptimo con la primera evaluación definitiva (en 2021 se realiza el piloto).
	Censal o muestral	Muestral y censal.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	Matricial. Usa el método de bloques incompletos balanceados.
	¿Para qué se usan los resultados?	Desde el 2000 se realizaban pruebas muestrales, representativas de algunos departamentos, pero desde el 2012 se comenzaron a hacer pruebas censales aplicadas por los docentes, además de una muestra controlada. Además, en 2017 se incorporó el resultado por estudiante.
	¿La prueba tiene consecuencias?	No.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	Para cada prueba, se presentan los dominios, definidos por las competencias y componentes propios del área. Además, se presenta una tabla con la proporción de ítems de cada uno de los dominios en cada prueba. La prueba de lenguaje evalúa dos competencias: la comunicativa-lectora y la comunicativa-escritora. Para la evaluación de las competencias se consideran tres componentes transversales: el sintáctico, el semántico y el pragmático. En la prueba de matemática las competencias son: el razonamiento y la argumentación; la comunicación, la representación y la modelación, y el planteamiento y resolución de problemas. Los tres componentes son: el numérico variacional, el geométrico-métrico y el aleatorio.
	¿Tiene dominios definidos por los documentos curriculares?	Su diseño está alineado con los estándares básicos de competencias establecidos por el Ministerio de Educación Nacional, que son los referentes comunes a partir de los cuales es posible establecer qué tanto los estudiantes y el sistema educativo en su conjunto están cumpliendo unas expectativas de calidad en términos de lo que saben y lo que saben hacer. A partir del 2018 se decidió implementar nuevas pruebas, teniendo en cuenta los estándares básicos de competencias por ciclo (no por grado).
	¿Usa el diseño de test basados en evidencias?	Modelo basado en evidencias.
4. Cantidad y tipo de ítems	Cantidad de ítems por prueba	En la prueba de lenguaje de tercero se responden 36 ítems, en quinto 36 y en noveno 54. En la prueba de matemáticas de tercero se responden 40 ítems, 48 en quinto y 54 en noveno. En la prueba de competencias ciudadanas se responden 54 ítems en quinto y 54 en noveno (no se releva en tercero).
	Tipos de ítems	Todos los ítems son de opción múltiple.

5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	El año previo al operativo se realiza un operativo piloto. A estos ítems se les aplican análisis univariados TC y análisis multivariados TRI 2p, se analiza el sesgo. Se hacen evaluaciones piloto por la validez y confiabilidad de la prueba, para la consistencia interna. Con los resultados obtenidos se hacen análisis factoriales y confirmatorios.
6. Niveles de desempeño	Cantidad de niveles	Cuatro niveles de desempeño en las pruebas de matemática y lenguaje en tercero, quinto y noveno.
	¿Algún nivel fija el estándar mínimo?	No se especifica.
7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	Entre un 25% y 40% de ítems de anclaje.
	¿Qué características tienen?	Para equiparar con ediciones anteriores de la prueba se analizan los parámetros obtenidos en el anclaje y se los compara para decidir si usar el parámetro nuevo o el anterior, si hubiera diferencias. Se estudian posibles justificaciones para el cambio en parámetros de ítems de anclaje.
	¿Cómo se eligen los ítems de anclaje?	No se explicita.
8. Observaciones o comentarios	Desde el 2000 se venía realizando un tipo de pruebas, pero a partir del 2018 se realizaron definiciones curriculares y se adaptaron las pruebas.	
9. Enlaces a reportes técnicos e informes de resultados	https://www.icfes.gov.co/web/guest/objetivo-saber-3579 https://www.icfes.gov.co/documents/20143/176813/Guia+pruebas+saber+3+5+9+lineamientos+para+las+aplicaciones+muestral+y+censal+2015+2+v2.pdf/5108835e-1282-75f4-c386-9bd904f379fd	

PRUEBAS SAEB – BRASIL

1. Información general de la prueba	Nacional	Brasil. El Saeb está integrado por tres evaluaciones nacionales: la evaluación nacional de educación básica (Aneb), la evaluación nacional de rendimiento escolar (Prueba Brasil) y la evaluación nacional de alfabetización (ANA).
	Áreas y grados	Lengua y matemática en segundo, quinto y noveno de primaria, y tercero de secundaria. Humanidades y ciencias naturales en noveno de primaria (lengua y matemática en segundo de primaria, y humanidades y ciencias naturales en noveno de primaria se incorporaron en 2019).
	Frecuencia de aplicación	Cada dos años.
	Censal o muestral	Censal en lo público para lengua y matemática de quinto y noveno de primaria, y tercero de secundaria. Muestral en lo público para lengua y matemática de segundo de primaria, y humanidades y ciencias naturales en noveno de primaria. Muestral en todo lo privado.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	Matricial con cuadernillos. Se utiliza una metodología denominada bloques incompletos balanceados (BIB). Para cada área de conocimiento, se arman 7 bloques, con 11 ítems cada uno, totalizando 77 ítems. Cada cuaderno de prueba se arma agrupando 2 bloques de lengua y 2 de matemática. La combinación de los bloques resulta en 21 cuadernos de prueba diferentes. Cada estudiante responde solamente un cuadernillo de prueba con 22 ítems de lengua y 22 de matemática (datos de quinto año, de la Prueba Brasil del 2013).
	¿Para qué se usan los resultados?	Los objetivos de Saeb son: (i) evaluar la calidad, la equidad y la eficiencia de la educación; (ii) producir indicadores educacionales para Brasil, sus regiones y unidades federales y, cuando es posible, para los municipios y las instituciones escolares; (iii) subsidiar la elaboración, el monitoreo y la valoración de políticas públicas basadas en evidencias, con vistas al desarrollo social y económico de Brasil, y (vi) desarrollar competencia técnica y científica en el área de la evaluación educacional.
	¿La prueba tiene consecuencias?	No se explicita.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	Tienen "matrices de referencia" alineadas con la Base Nacional Común Curricular (BNCC).
	¿Tiene dominios definidos por los documentos curriculares?	No se explicita.
	¿Usa el diseño de test basados en evidencias?	Sí.

4. Cantidad y tipo de ítems	Cantidad de ítems por prueba.	22 de lengua y 22 de matemática por cuadernillo (datos de quinto año, de la Prueba Brasil del 2013).
	Tipos de ítems	Hay ítems de múltiple opción y de respuesta construida. Los de respuesta construida solo están presentes en la pruebas más recientes (lengua y matemática de segundo de primaria y humanidades y ciencias de noveno de primaria), que tienen matrices de referencia construidas recientemente; el resto de las pruebas, que se realizan a partir de matrices de referencia construidas con anterioridad, mantienen exclusivamente los ítems de múltiple opción, para posibilitar la comparación con la serie histórica de resultados de las pruebas.
5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	No se explicita.
6. Niveles de desempeño	Cantidad de niveles	Variable según área de conocimiento y grado, la que tiene más niveles va del 0 al 10 (matemática de quinto) y la que tiene menos niveles va del 1 al 8 (lengua de noveno) (datos de la prueba 2013).
	¿Algún nivel fija el estándar mínimo?	Sí.
7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	No se explicita.
	¿Qué características tienen?	
	¿Cómo se eligen los ítems de anclaje?	
8. Observaciones o comentarios		
9. Enlaces a reportes técnicos e informes de resultados	https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/saeb_documentos_referencia_versao_preliminar.pdf https://download.inep.gov.br/educacao_basica/prova_brasil_saeb/resultados/2013/caderno2013_v2016.pdf	

NAEP – ESTADOS UNIDOS

1. Información general de la prueba	Nacional	Estados Unidos.
	Áreas y grados	En matemáticas, lectura, ciencia, escritura, tecnología e ingeniería, artes, educación cívica, geografía, economía e historia de los Estados Unidos. Se aplica en los grados cuarto, octavo y duodécimo.
	Frecuencia de aplicación	En los grados cuarto y octavo es cada dos años, en grado duodécimo, cada cuatro.
	Censal o muestral	Muestral.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa	Matricial.
	¿Para qué se usan los resultados?	Los resultados se divulgan ampliamente. Portavoces políticos y otras personas han utilizado los resultados de la NAEP para respaldar mensajes positivos y negativos sobre la calidad del sistema educativo de los Estados Unidos. La NAEP monitorea las tendencias en el desempeño de los subgrupos. Se presta especial atención a la tasa de progreso de los grupos minoritarios, especialmente a los aumentos en el puntaje de comprensión lectora desde 1971. Un objetivo de la NAEP es medir los cambios en el desempeño a lo largo del tiempo. Los resultados de diferentes administraciones deben basarse en una escala común para que se puedan realizar comparaciones válidas.
	¿La prueba tiene consecuencias?	No.

3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	Cada elemento de las pruebas de la NAEP se desarrolla para medir uno de los objetivos, que se organizan en áreas principales de contenido.
	¿Tiene dominios definidos por los documentos curriculares?	Cada evaluación de la NAEP se basa en un marco organizativo. El marco es el anteproyecto que guía el desarrollo de la evaluación y el contenido a evaluar. La Junta Directiva de Evaluación Nacional desarrolla los marcos de la NAEP para las evaluaciones en cada materia. Los marcos definen el contenido específico de la materia y las habilidades de pensamiento que necesitan los estudiantes para lidiar con los problemas complejos que encuentran dentro y fuera del aula. Los marcos de la NAEP se diseñan a través de un proceso de desarrollo que garantiza que cumplan con los requisitos educativos actuales. La Junta de Gobierno es responsable de desarrollar tanto el marco como las especificaciones de prueba que sirven como modelo para la evaluación. Los marcos se diseñan a través de un proceso de desarrollo que garantiza que cumplan con los requisitos educativos actuales. Los pasos en el desarrollo de cada una de las pruebas de la NAEP son: • la Junta Directiva de Evaluación Nacional desarrolla marcos de contenido y especificaciones de elementos en cada área temática; • el comité de desarrollo de instrumentos en cada área temática capacita al personal de la NAEP sobre cómo se pueden medir los objetivos descritos en el marco y acerca de las prioridades para la evaluación (dentro del contexto del marco de evaluación) y los tipos de ítems a desarrollar; • los especialistas con experiencia en la materia y experiencia en la creación de ítems de acuerdo con las especificaciones desarrollan y revisan las preguntas de evaluación; • el personal de desarrollo de exámenes de la NAEP y los especialistas de exámenes externos revisan los ítems y las guías de codificación que los acompañan; • se revisan los cuestionarios de contexto, se hacen revisiones editoriales y de imparcialidad; • se preparan materiales de la prueba piloto; • se lleva a cabo una prueba piloto en muchos de los estados y jurisdicciones programados para participar en la evaluación definitiva del año siguiente; • sobre la base de los análisis de la prueba piloto, los ítems se seleccionan para su inclusión en la evaluación definitiva; • cada comité de desarrollo de instrumentos del área temática aprueba la selección de ítems para incluir en la evaluación definitiva del año siguiente, y • se diseñan los bloques y cuadernillos de ítems para la evaluación definitiva.
	¿Usa el diseño de test basados en evidencias?	No.
4. Cantidad y tipo de ítems	Cantidad de ítems por prueba	No se explicita.
	Tipos de ítems	Múltiple opción, respuesta construida breve, respuesta construida extensa y tareas interactivas basadas en escenarios.
5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	Sí, el año previo a la evaluación definitiva.
6. Niveles de desempeño	Cantidad de niveles	Tres: básico, competente y avanzado.
	¿Algún nivel fija el estándar mínimo?	Los estudiantes que se desempeñan en o por encima del nivel ***competente*** de la NAEP demuestran un desempeño académico sólido y competencia sobre materias desafiantes. Cabe señalar que el nivel de desempeño ***competente de la NAEP*** no representa el dominio del nivel de grado según lo determinado por otros estándares de evaluación (por ejemplo, evaluaciones estatales o distritales).

7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	No se explicita.
	¿Qué características tienen?	
	¿Cómo se eligen los ítems de anclaje?	
8. Observaciones o comentarios		
9. Enlaces a reportes técnicos e informes de resultados	https://nces-ed-gov.translate.goog/nationsreportcard/assessments/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=nui,sc https://nces.ed.gov/nationsreportcard/tdw/analysis/ https://nces.ed.gov/nationsreportcard/tdw/instruments/	

NAPLAN - AUSTRALIA

1. Información general de la prueba	Nacional	Australia.
	Áreas y grados	Los estudiantes de los grados tercero, quinto, séptimo y noveno son evaluados en las habilidades de lectoescritura y aritmética. Además, en los grados sexto y décimo, y cada tres años, se hacen evaluaciones de carácter muestral sobre ciencias, educación, cívica, ciudadanía, tecnología (alternando áreas).
	Frecuencia de aplicación	Anual.
	Censal o muestral	Censal.
2. Diseño y usos de la prueba	Cuadernillo único, matricial con cuadernillos o adaptativa.	Matricial.
	¿Para qué se usan los resultados?	Los resultados de las pruebas NAPLAN brindan información sobre el desempeño de los estudiantes en las áreas de lectoescritura y aritmética y respaldan las mejoras en la enseñanza y el aprendizaje. Los datos de los resultados de las pruebas NAPLAN brindan a las escuelas y los sistemas la capacidad de medir los logros de sus estudiantes en comparación con los estándares mínimos nacionales y el desempeño de los estudiantes en otros estados y territorios. Todos los estudiantes que participaron en las pruebas NAPLAN reciben un informe individual de sus resultados.
	¿La prueba tiene consecuencias?	No.
3. Constructo de la prueba	¿Cómo se operacionaliza el constructo?	No se explicita.
	¿Tiene dominios definidos por los documentos curriculares?	Las pruebas de NAPLAN reflejan el currículo australiano.
	¿Usa el diseño de test basados en evidencias?	No se explicita.
4. Cantidad y tipo de ítems	Cantidad de ítems por prueba	Entre 178 y 256 ítems por prueba, entre 36 y 64 por cuadernillo.
	Tipos de ítems	De múltiple opción, de respuesta construida e interactivos.
5. Pilotaje de ítems	¿Se pilotean los ítems?	Sí.
	¿Se hace un operativo piloto previo?	Sí.
6. Niveles de desempeño	Cantidad de niveles	Hay diez niveles en total, pero para cada grado corresponden seis de los niveles.
	¿Algún nivel fija el estándar mínimo?	Sí. De los seis niveles, el mínimo es el segundo nivel.

7. Proceso de equiparación	¿Cuántos ítems se usan de anclaje?	No se explicita. Se menciona que se hace equiparación vertical y horizontal.
	¿Qué características tienen?	
	¿Cómo se eligen los ítems de anclaje?	
8. Observaciones o comentarios		
9. Enlaces a reportes técnicos e informes de resultados	https://www.acara.edu.au/assessment https://www.nap.edu.au/ https://nap.edu.au/docs/default-source/resources/naplan-2019_technical-report_final.pdf	

BIBLIOGRAFÍA

Ferrer, G. (2006). *Estándares en Educación. Implicancias para su aplicación en América Latina* (1.a ed.). PREAL.

Manzi, J., García, M. R. y Taut, S. (Eds.). (2019). *Validez de evaluaciones educacionales en Chile y Latinoamérica*. Santiago de Chile: Ediciones Universidad Católica de Chile.

Ravela, P. (2006). *Para comprender las evaluaciones educativas* (1.a ed.). Recuperado de https://www.mineducacion.gov.co/cvn/1665/articles-125590%7B%5C_%7Darchivo%7B%5C_%7Dpdf.pdf

Ravela, P., Arregui, P., Valverde, G., Wolfe, R., Ferrer, G., Martínez, F., ... Wolff, L. (2008). Las evaluaciones educativas que América Latina necesita. En R*evista Iberoamericana De Evaluación Educativa*. Santiago de Chile.