Cuadernillo técnico
de evaluación educativa

1

Nociones básicas en medición y evaluación en el contexto educativo

Centro UC
Medición - MIDE

Textos de divulgación

INEE
Instituto Nacional para la Evaluación de la Educación
México

Cuadernillo técnico
de evaluación educativa

1

Nociones básicas en medición y evaluación en el contexto educativo

Centro UC
Medición - MIDE

Textos de divulgación

INEE
Instituto Nacional para la Evaluación de la Educación
México

Nociones básicas en medición y evaluación en el contexto educativo

Av. Vicuña Mackenna 4860
Macul, Santiago, Chile, cp 7820436

Barranca del Muerto 341, col. San José Insurgentes,
Alcaldía Benito Juárez, Ciudad de México, cp 03900

Autores
María Paulina Flotts de los Hoyos, MIDE UC
María Beatriz Rodríguez Frias, MIDE UC

Editora
María Rosa García González, MIDE UC

Corrección de estilo
Arturo Cosme Valadez, INEE
Lissette Sepúlveda Cepeda, MIDE UC

Coordinación General
Adriana Guadalupe Aragón Díaz, INEE
Marcela Cuevas Ossandón, MIDE UC
Marcela Ramírez Jordán, INEE

Diseño
www.iunta.cl

Índice

Presentación

El Instituto Nacional para la Evaluación de la Educación de México, INEE, y el Centro de Medición MIDE UC, de la Pontificia Universidad Católica de Chile, han gestado una colaboración para el desarrollo y fortalecimiento de capacidades en evaluación educativa, en profesionales del Instituto y de los equipos responsables de los Programas Estatales de Evaluación y Mejora Educativa (PROEME) y del Proyecto Nacional de Evaluación y Mejora Educativa de Escuelas Multigrado (PRONAEME), en el marco del Sistema Nacional de Evaluación Educativa (SNEE), en México.

El documento que a continuación presentamos constituye un material de consulta que forma parte de una serie de nueve cuadernillos, cuyo propósito es orientar la comprensión de los conceptos centrales de la medición y la evaluación educativas y su impacto en el diseño de instrumentos; considerando que el proceso evaluativo es una suma de decisiones que deben cuidar la coherencia de cada uno de los elementos y fases que lo componen.

Este material se ha organizado en una serie de cuadernillos con base en las siguientes temáticas:

1. Nociones básicas en medición y evaluación en el contexto educativo.
2. Confiabilidad, validez e imparcialidad en evaluación educativa.
3. Definición del marco de referencia de la evaluación.
4. Desarrollo de instrumentos de evaluación: pruebas.
5. Desarrollo de instrumentos de evaluación: cuestionarios.
6. Desarrollo de instrumentos de evaluación: pautas de observación.
7. Desarrollo de instrumentos de evaluación: tareas de desempeño y rúbricas.
8. Análisis y uso de resultados.
9. Uso de resultados y retroalimentación.

Esperamos que este material resulte de utilidad para los profesionales que se desempeñan en el contexto de la medición y evaluación educacional. En los cuadernillos encontrarán nociones y conceptos fundamentales, además de recomendaciones prácticas, y sugerencias bibliográficas para quienes deseen profundizar en cada una de las temáticas trabajadas.

Nociones básicas en medición y evaluación en el contexto educativo

Resumen

En este cuadernillo se abordan conceptos básicos asociados con la medición y evaluación educativa. Se describen conceptos centrales, como la diferencia entre medir y evaluar, y los distintos tipos de evaluación que se distinguen en la literatura y en la práctica.

Además, se hace una presentación del proceso de construcción de instrumentos y dispositivos para levantar información, mostrando la variedad de herramientas que suelen ocuparse en el contexto educativo para medir y evaluar una amplia diversidad de constructos. Se pone énfasis en que no hay mejores o peores instrumentos en sí, sino que son más o menos pertinentes dada la naturaleza de aquello que se quiere conocer y del tipo de resultados que se espera obtener.

El tema central de este cuadernillo es la coherencia que debe existir entre los distintos componentes y fases del proceso evaluativo. Diseñar e implementar una evaluación conlleva una serie de decisiones siempre adoptadas en función del uso intencionado que se quiere dar a los resultados: qué se evalúa, cómo se hace y analiza, y qué tipo de información se genera. El conjunto debe ser consistente con el objetivo establecido.

Introducción

Este cuadernillo es el primero de una serie de nueve documentos que abordan conceptos y aspectos vinculados con la medición y evaluación en el contexto educativo. Esta cualidad de ser "el primer cuadernillo" justifica que su enfoque sea el de un barrido sobre varios temas, en los que se implican distintos niveles de abstracción y análisis de dimensiones concretas. De algún modo, el texto busca entregar al lector un esquema sinóptico -fundamentado y con un nivel de detalle que facilite la comprensión- del proceso evaluativo, identificando sus principales hitos y los requerimientos que deben satisfacerse para obtener resultados confiables y que puedan ser usados para tomar decisiones.

El texto se organiza en seis secciones. La primera presenta las distinciones básicas que es preciso manejar al abordar el tema de la evaluación en el contexto educativo. La segunda aborda en detalle las distintas clasificaciones de la evaluación y, a partir de ellas, se distinguen los tipos de evaluación educativa; se revelan las implicaciones de cada uno de acuerdo con la información que se espera levantar y los instrumentos o dispositivos para hacerlo, el tipo de conclusiones que se puede extraer a partir de ellas, y el uso que se hará de los resultados obtenidos. La tercera sección aborda el tema de la coherencia que debe tener un proceso o sistema de evaluación a lo largo de sus fases o componentes, y cómo ello se relaciona con el proceso de validación de un instrumento. Tal consistencia se puede plantear a partir de las respuestas que el diseñador debe dar a las preguntas: *¿qué evaluar?, ¿para qué evaluar?, ¿cómo evaluar?, ¿cuándo evaluar?* y *¿A quiénes evaluar?* La respuesta a estas interrogaciones se vincula con los hitos de un proceso de evaluación. En la cuarta sección se presentan los instrumentos o dispositivos orientados a recolectar evidencia para la evaluación; se pone énfasis en que su elección depende siempre de la naturaleza de aquello que interesa conocer y, asimismo, de las restricciones o limitaciones que imponen el contexto y las situaciones específicas en las cuales será llevado a cabo el proceso evaluativo.

En la quinta sección se dejan atrás estas nociones y distinciones conceptuales, y se da paso a la descripción del proceso de construcción de instrumentos; se detallan las características de cada etapa y se expone la importancia que tiene cada una para garantizar la calidad y solidez de la información que se busca levantar.

La sexta y última sección aborda el tema de qué se mide y evalúa en el contexto educativo; se presentan los usos que se dan a la medición y evaluación, y se alude a los principales tipos de ellos presentes en los países de la región, tanto a nivel de aula como a gran escala (logros de aprendizaje de los estudiantes, factores asociados con dichos logros, evaluación del desempeño docente y evaluación de programas).

A partir de esta revisión se espera contribuir a consolidar un lenguaje básico en este dominio, que permita hacer distinciones conceptuales relevantes y favorecer el trabajo de quienes participan en los procesos de diseño, aplicación, análisis y uso de la información en un marco de rigor técnico... También deseamos motivar la lectura de los demás cuadernillos de esta colección.

I. Evaluación educativa: ¿cuáles son las distinciones básicas a tomar en cuenta?

La evaluación es un elemento constitutivo y, por tanto, insoslayable del contexto educativo. La experiencia revela de manera inequívoca que todos los actores del sistema se ven expuestos en forma reiterada y sistemática a procesos de evaluación, sea en el papel de evaluadores o en el de evaluados. Esa misma experiencia muestra, la mayor parte de las veces, que la evaluación se percibe como un hito aislado y desvinculado del proceso de enseñanza aprendizaje o, incluso, en un escenario peor, como una experiencia lejana, impuesta, sin sentido.

Pero evaluar, evaluar bien, es una herramienta que puede ser de gran utilidad y provecho para orientar la toma de decisiones en el contexto educativo. Es sensato pensar que estas pueden ser mejor conducidas cuando se basan en información confiable, cuando hay evidencia que las respalde. Y la evaluación es precisamente eso: evidencia... Evidencia de lo que un estudiante sabe y es capaz de hacer; de los resultados, o de su falta, luego de la implementación de un programa; del desempeño de un maestro en el salón; de los aprendizajes de la población escolar de un distrito, ciudad o país; de la influencia de factores socioculturales como facilitadores o inhibidores del aprendizaje, entre otros. Evaluar bien es dar la oportunidad de que esta evidencia se haga visible y pueda ser sistematizada.

Siguiendo a Ravela (2006: 29):

> La evaluación -bien realizada- puede ser una herramienta de cambio de enorme potencial. Si los sistemas educativos mejoraran los distintos tipos de evaluación que ocurren a diario, ello tendría un enorme impacto en el sistema educativo: los alumnos recibirían mejor apoyo en sus procesos de aprendizaje, las evaluaciones de certificación serían más justas y garantizarían que los individuos estén preparados para lo que se supone fueron formados, el sistema seleccionaría a los individuos más competentes para desempeñar funciones de conducción y responsabilidad institucional, los profesores y las escuelas aprenderían más de su experiencia, las familias conocerían mejor qué es lo que sus hijos están intentando aprender y qué dificultades tienen, la sociedad

en general tendría mayor conocimiento y compromiso con la educación, las políticas educativas podrían estar sustentadas en una base de información más sistemática.

Hacer una buena evaluación supone satisfacer una serie de requisitos para que sea un proceso serio, confiable y robusto. Es importante distinguir algunos conceptos básicos para ser precisos en el uso del lenguaje y no inducir equívocos o confusiones a los actores del proceso educativo y a las audiencias que reciben los resultados de una evaluación.

Primera distinción: medir y evaluar

Aunque se suelen usar como sinónimos intercambiables, en términos técnicos medir y evaluar no son lo mismo. Siguiendo a Scriven (2013), la evaluación es el acto o proceso cognitivo por el cual se establece una afirmación respecto de la calidad, valor o importancia de cierta entidad. La medición, por su parte, es un proceso por el cual se asignan números a atributos, observables y no observables, de acuerdo con parámetros y reglas claramente definidas (Fenton y Pfleeger, 1997). En palabras simples, medir es asignar números y evaluar es hacer un juicio integral acerca de las cualidades del objeto de interés. Ambas acciones pueden considerarse complementarias; de hecho, el resultado de una medición puede ser un insumo para la evaluación; por otra parte, un juicio evaluativo permite dar sentido y significado al dato de una medición, otorgándole así un marco que promueva la acción y la toma de decisiones.

¿Por qué es importante esta distinción? Porque ayuda a determinar los alcances de los datos y los juicios, y a establecer sus limitaciones. El dato resultante de una medición difícilmente basta, por sí mismo, para concluir si se avanza o no en las metas propuestas, si se está en un buen o mal escenario o si se alcanzan los niveles de aprendizaje o desempeño esperados. Todos estos son juicios que se formulan tomando en consideración otras variables y elementos, desde el caso más simple -por ejemplo, el resultado de una medición anterior, para ver si hay avance o progreso- hasta otros más complejos, como la medición de variables asociadas, que son revisadas por un panel de expertos. Por otra parte, un juicio evaluativo solo adquiere solidez cuando hay evidencia que lo respalda, y los datos que aporta una buena medición son precisamente eso: evidencia relevante.

En suma, cuando interesa conocer los avances o logros educativos de un estudiante, curso, escuela, distrito o país, se puede tener una mejor fotografía si se utilizan los dispositivos de medición apropiados para levantar sistemáticamente información, la que luego, combinada con otros datos y referentes, permite fundamentar un juicio que integra y pondera.

Segunda distinción: niveles de la evaluación en el contexto educativo

Tradicionalmente se distinguen dos niveles cuando se habla de evaluación educativa: de aula y a gran escala. La diferencia no solo tiene que ver con el contexto micro o macro en que se lleva a cabo la evaluación, también tiene implicaciones de otro tipo, tal como se ilustra en la tabla 1 (Committee..., 2003):

TABLA 1

COMPARACIÓN ENTRE LAS EVALUACIONES DE AULA Y A GRAN ESCALA

	Evaluación de aula	Evaluación a gran escala
Foco	La retroalimentación individual del proceso de enseñanza y aprendizaje para profesores y alumnos.	La toma de decisiones a gran escala, rendición de cuentas y retroalimentación de las políticas.
Requerimientos fundamentales	Se diseña de acuerdo con un contexto y propósito determinados y no exige un requerimiento técnico métrico.	Para asegurar la comparabilidad de resultados entre grupos y años debe: • ser única para todo el sistema; • aplicarse en condiciones uniformes; • responder a requisitos técnicos estrictos.
Entrega de resultados	Resultados y retroalimentación inmediatos.	El procesamiento implica resultados diferidos.
Nivel de detalle de la información que entrega	Al desarrollarse continuamente, puede dar información minuciosa.	Suele ser una evaluación de síntesis, por lo que entrega información poco específica (en términos de los objetivos de aprendizaje esperados).
Ámbito de aprendizajes	Evaluar una amplia gama de aprendizajes, con foco tanto en productos como en procesos.	Tiene limitaciones para evaluar algunas competencias relevantes -por ejemplo, expresión oral- dada las dificultades logísticas que implica.

Fuente: elaboración propia.

La evaluación de aula es diseñada y administrada por el profesor. Es él o ella quien construye los reactivos, sitúa el ejercicio en un momento particular del año escolar, asigna puntos o calificaciones y decide cómo informar y usar los resultados. Suele ser una evaluación contextualizada, es decir, que toma en consideración la experiencia de enseñanza aprendizaje concreta del grupo de estudiantes y, en este sentido, tiene un fuerte componente local. Se aplica con regularidad dentro del período escolar (trimestre

o semestre) y genera resultados, habitualmente expresado en una calificación, de manera contingente al momento de su aplicación. Puede tomar la forma de muy diversos instrumentos y dispositivos, lo cual favorece que el maestro pueda evaluar aprendizajes de distinto tipo y naturaleza (conocimientos declarativos, aplicación de procedimientos, desempeños, ejecuciones, etcétera).

La evaluación a gran escala es diseñada, administrada y analizada a escala de administración general de los sistemas educativos, y suele tener como foco la obtención de información confiable y válida para la política educativa de un distrito, región o país. Supone altos niveles de estandarización, tanto en el diseño de los instrumentos como en su aplicación, con el fin de garantizar la comparabilidad de los resultados; debe satisfacer una serie de requerimientos técnicos y métricos, lo cual tiene alto impacto en el rigor de los procesos de construcción, pilotaje (o prueba de campo) y análisis estadístico basado en teorías de medición. Se suele aplicar en forma periódica, pero espaciada anualmente, por lo que no permite llevar un monitoreo cercano y cotidiano del proceso de aprendizaje. Sus resultados se entregan varios meses después de la aplicación. Por su carácter masivo y las exigencias de estandarización, la evaluación a gran escala suele ser más restrictiva que la de aula respecto a los conocimientos, habilidades y competencias medidas.

El paso entre las evaluaciones de aula y a gran escala suele no ser fácil ni fluido. Habitualmente, los actores del sistema educativo los consideran asuntos independientes y desconectados. En sistemas que tienen evaluaciones estandarizadas a gran escala, es común que las escuelas las perciban amenazantes y contrapuestas a las motivaciones de la comunidad escolar. Sin embargo, es posible dar un giro a esta percepción negativa si se toma en cuenta que

> [...] las evaluaciones estandarizadas y las evaluaciones en el aula son complementarias y no antagónicas. Cada una permite 'ver' o 'hacer' algunas cosas, pero no otras. La evaluación externa sirve para poner el foco de atención en aquello que todos los alumnos deberían aprender, pero, por supuesto, no puede ni pretende dar cuenta de todos los aprendizajes. La evaluación en el aula, cuando se hace bien, puede ser mucho más rica en su apreciación de los procesos de aprendizaje de alumnos específicos, pero no puede nunca ofrecer un panorama de lo que ocurre a nivel del conjunto del sistema educativo (Ravela, 2006, p. 73).

Un interesante desafío para los maestros y las comunidades escolares es aprovechar la información que entrega una evaluación externa y estandarizada y cotejarla con los resultados de sus propias evaluaciones, ponderar su coherencia o incoherencia, buscar

hipótesis explicativas acerca de ello y, a partir de esta reflexión, aprovecharla como un insumo para retroalimentar y mejorar la propia práctica pedagógica, sin perder de vista que

> [...] solo un buen maestro puede llevar a cabo la evaluación más importante de cada alumno. Una evaluación que incluya todos los aspectos del currículo y los niveles cognitivos más complejos, que tenga en cuenta las circunstancias de cada niño, y se haga con la frecuencia necesaria para ofrecer retroalimentación oportuna para que el alumno pueda mejorar. Este tipo de evaluaciones son las que deben hacerse en cada aula regularmente, con acercamientos más finos que los que pueden emplearse a gran escala (Martínez, 2009, p. 10).

A estos dos tipos de la evaluación educativa se puede agregar un tercero: la evaluación de programas (Patton, 2002), es decir, el estudio respecto de cuán bien se está implementando o se ha implementado un programa y por qué. Puede hacerse para determinar su efectividad (una vez habilitado) o bien, para mejorarlo (durante su instrumentación); en el primer caso, el foco está puesto en evaluar el grado de cumplimiento de las metas del programa o intervención, esperando como resultado que sea posible establecer juicios y generalizaciones sobre las condiciones bajo las cuales esos esfuerzos son efectivos; en el segundo caso, el énfasis se concentra en evaluar fortalezas y debilidades que permitan mejorar la implementación del programa.

II. ¿Cuáles son los distintos tipos de evaluación?

En la literatura especializada y en la práctica profesional docente se distinguen varios tipos de evaluación. Para exponerlos, se hace en este documento un ordenamiento en función de tres criterios de clasificación de las evaluaciones: según su finalidad, según sus usos y consecuencias, y según su referente.

Tipos de evaluación según su finalidad: evaluación formativa y evaluación con fines de certificación (sumativa)

A partir de la distinción hecha por Michael Scriven en los años sesenta, se suele hablar de *evaluación sumativa* para referir a aquella que ocurre al final de un periodo educativo y que tiene como propósito fundamental calificar al estudiante; y de *evaluación fomativa,* aludiendo a la que ocurre durante el proceso de enseñanza con la finalidad de adaptarlo a las necesidades de aprendizaje de los estudiantes para mejorar su desempeño (Scriven, 1967).

La *evaluación sumativa,* según se apuntó, está intrínsecamente vinculada con la calificación, toda vez que cumple la función de informar y certificar públicamente acerca

del grado en que cada estudiante ha logrado los aprendizajes esperados de un período educativo dado. Realiza, por tanto, "una función social de acreditación de conocimientos y capacidades, propia de los sistemas de educación formal" (Ravela, Picaroni y Loureiro, 2018, p. 207). Precisamente por esto se suele aplicar al final de un período de aprendizaje -de ahí su carácter sumativo- y poner el foco en los resultados, no en el proceso.

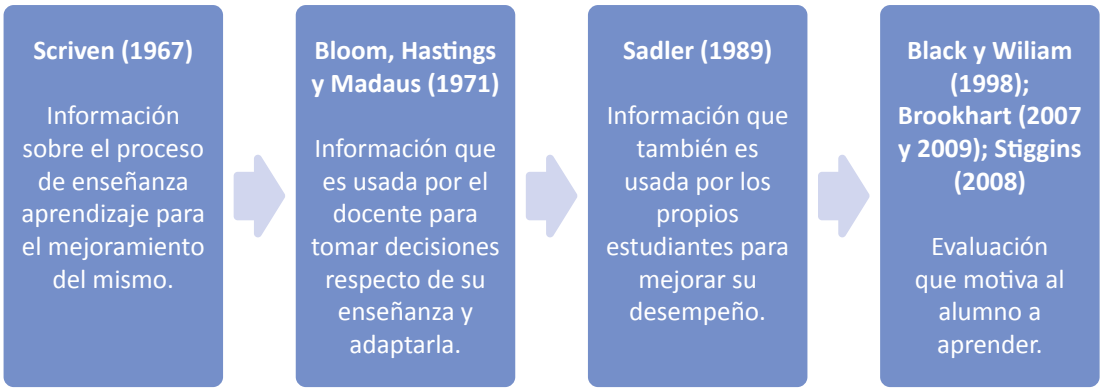
La *evaluación formativa,* por su parte, tiene como finalidad retroalimentar a estudiantes y maestros en las distintas etapas del proceso de aprendizaje (Bloom, 1969, *citado en* Wiliam, 2011) y apoyar al docente para tomar decisiones respecto de su enseñanza a partir de lo observado (Bloom, 1971, *citado en* Martínez, 2012). "El carácter formativo de la evaluación, por tanto, no es algo intrínseco a la actividad propuesta, sino que refiere al uso de la información recogida" (Ravela, Picaroni y Loureiro, 2018, p. 146). Lo que hace formativa a una evaluación no es el tipo de instrumentos que se emplean, ni el momento en que se aplica, ni tampoco si es o no calificada: por ejemplo, un examen se puede utilizar para determinar los aprendizajes de una unidad y decidir cómo seguir avanzando a la siguiente.

Según estos autores, el concepto de evaluación formativa es homologable al de *evaluación para el aprendizaje* (y se distingue del de evaluación sumativa o con fines de certificación): "en los últimos años algunos autores han reelaborado la distinción entre evaluación formativa y sumativa, utilizando las expresiones de evaluación *del* aprendizaje y evaluación *para* el aprendizaje, indicando que en el primer caso se está intentando valorar lo que los estudiantes han logrado, en tanto que en el segundo, la evaluación es parte intrínseca de los procesos de enseñar y aprender" (Ravela, Picaroni y Loureiro, 2018, p. 146).

Martínez Rizo (2012), a partir de lo que ha postulado Susan Brookhart (2007 y 2009), expone la expansión del concepto de evaluación formativa (figura 1).

FIGURA 1

ESQUEMA SINÓPTICO DE LA EXPANSIÓN DEL CONCEPTO DE EVALUACIÓN FORMATIVA



Fuente: elaboración propia.

La evaluación para el aprendizaje o evaluación formativa ha sido conceptualizada precisamente como un puente entre la enseñanza y el aprendizaje (Wiliam, 2011), toda vez que su función principal es hacer visible la brecha que hay entre lo que el profesor enseña y lo que los estudiantes efectivamente aprenden. "Una capacidad clave de un buen docente es la de percibir y valorar la distancia entre lo que se propuso lograr -sus intenciones educativas- y lo que realmente alcanzó cada estudiante -aprendizaje-" (Ravela, Picaroni y Loureiro, 2018, p. 154). En función de esta brecha, el maestro puede reformular sus acciones pedagógicas y retroalimentar a los estudiantes.

Según Wiliam (2011), la evaluación formativa (o para el aprendizaje) implica la participación de tres actores: el docente, el estudiante y el grupo de pares. Además, el autor distingue tres procesos clave de este tipo de evaluación:

1. Clarificar y compartir las intenciones educativas de manera que sea comprensible para los estudiantes (dónde deberían llegar).

2. Generar evidencia acerca de qué están aprendiendo (dónde están en relación con las intenciones educativas).

3. Llevar a cabo devoluciones y dar orientación que le permita a los estudiantes ajustar su desempeño y seguir aprendiendo (cómo avanzar hacia la meta deseada).

Bajo el enfoque de la evaluación para el aprendizaje se asume, entonces, que cada estudiante es capaz de mejorar su desempeño en un proceso que lo involucra activamente. Requiere que los maestros compartan con sus alumnos lo que se espera de ellos en términos de metas de aprendizaje y les entregue una retroalimentación formativa, es decir, que muestre qué deben hacer para mejorar su desempeño y avanzar hacia la meta deseada a partir de la evidencia recogida en las actividades de evaluación.

Por último, cabe señalar que el carácter formativo o sumativo de una evaluación no solo se puede apreciar cuando se habla de logros de aprendizaje de los estudiantes en el aula. Aplica también a otros niveles de la evaluación y a otros actores del proceso educativo. Por ejemplo, como se señaló en el apartado anterior, la evaluación de un programa puede ser realizada durante su implementación para detectar fortalezas y debilidades y, a partir de ellas, entregar recomendaciones con el fin de mejorarlo (evaluación formativa); también puede ser hecha al finalizar el programa para determinar si los resultados obtenidos están en línea con las metas esperadas (evaluación sumativa).

Tipos de evaluación según sus usos y consecuencias: evaluación de bajas y de altas consecuencias

Las *evaluaciones de bajas consecuencias* son aquellas en las que no hay una consecuencia directa específica para los partícipes; su foco está puesto en "contribuir a mejorar la comprensión de la situación educativa y propiciar acciones y decisiones que permitan cambiar y mejorar" (Ravela, 2006, p. 24). Una *evaluación de altas consecuencias,* por otra parte, tiene implicaciones directas e importantes para una persona, programa o institución, es decir, para el evaluado (por ejemplo, el acceso a la educación superior, mejoras salariales, decisiones de desvinculación laboral, cierre o término anticipado de un programa, entre otras). Al igual que la modalidad anterior, puede evaluar a distintos niveles del sistema educativo y estar destinada a sus diversos actores.

Esta clasificación se vincula muy estrechamente con la presentada en el apartado anterior, pues usualmente las evaluaciones formativas tienen bajas consecuencias para las personas o sistemas evaluados, mientras que las sumativas suelen estar asociadas con altas consecuencias.

Los sistemas educativos implementan evaluaciones de altas y bajas consecuencias en distintos niveles y teniendo como partícipes a distintos a actores. Ejemplos de ello pueden verse en la tabla 2.

TABLA 2

EJEMPLOS DE EVALUACIONES DE BAJAS Y ALTAS CONSECUENCIAS EN DISTINTOS NIVELES Y ACTORES

Evaluaciones de bajas consecuencias	Evaluaciones de altas consecuencias
• Evaluaciones que hace el profesor para conocer los avances de sus alumnos en el aprendizaje de una unidad determinada. • Evaluaciones de desempeño docente para detectar necesidades de orientación profesional que permitan al maestro mejorar su enseñanza. • Evaluaciones a los centros educativos con el propósito de fomentar el diálogo de sus actores para superar las dificultades existentes. • Evaluaciones del currículum o de proyectos específicos con el fin de detectar necesidades de cambio y mejora. • Evaluaciones nacionales e internacionales de logros educativos orientadas a la identificación de fortalezas y debilidades, con vistas a la formulación de políticas educativas.	• Evaluaciones para certificar o acreditar que un individuo o institución posee ciertas características o atributos (y que implica una constancia formal, un reconocimiento social). • Evaluaciones para ordenar postulantes a un cargo, plaza o cupo (por ejemplo, postulantes a la universidad en un sistema nacional de admisión; maestros que postulan a una plaza en la planta docente de un distrito; ordenar centros educativos para determinar quiénes serán objeto de una intervención del nivel central, etc.). • Evaluaciones del desempeño docente o directivo para establecer quienes continuarán en su cargo y quienes deben dejar la dotación profesional. • Evaluaciones de proyectos, innovaciones o programas educativos con la finalidad específica de resolver acerca de su continuidad o término.

Fuente: Ravela (2006).

Cabe señalar que las evaluaciones de altas consecuencias también son referidas como aquellas que tienen por finalidad responsabilizar o promover la rendición de cuentas (*accountability*). Se trata de una evaluación que incluye mecanismos y procesos destinados a asegurar el cumplimiento de los objetivos y las obligaciones de los actores educativos, en especial -aunque no exclusivamente- de las escuelas, los directivos y los docentes (Hatch, 2013).

Tipos de evaluación según su referente: evaluación referida a normas y evaluación referida a criterios

Esta forma de clasificar las evaluaciones está en función del referente con el cual se contrastan los resultados para poder establecer un juicio. Se distinguen bajo esta mirada dos tipos de evaluaciones: las vinculadas a normas y las asociadas a criterios.

La *evaluación referida a normas* permite ordenar a los evaluados -personas, instituciones o sistemas- con el fin de compararlos a partir de su resultado o desempeño en la evaluación; el foco no está puesto en conocer cuánto sabe o es capaz de hacer cada evaluado, sino en conocer qué posición ocupa en el conjunto. A partir de este tipo de evaluaciones no es posible saber qué representa, en términos de dominio de habilidades y contenidos, tener un puntaje *x*, solo se sabe que *x* es mejor que *x-1* y peor que *x+1.*

La *evaluación referida a criterios*, en cambio, compara el desempeño o resultado del evaluado con una definición clara y explícita de lo que se espera que conozca y sea capaz de hacer en un determinado dominio o ámbito; no contrasta a los evaluados entre sí, sino con un estándar esperado (Ravela, 2006).

Esta distinción tiene impacto no únicamente en el tipo de resultados que cada una entrega -y por tanto en sus alcances y limitaciones-, sino también en las decisiones que deben tomarse para construir los instrumentos con los cuales se levantará la información. Al elaborar pruebas con un enfoque normativo "se eliminan tanto las preguntas muy fáciles como las muy difíciles. Esto se hace debido a que el propósito de la medición es ordenar o comparar a los individuos, y las preguntas que mejor sirven a esta finalidad son las de dificultad intermedia" (Ravela, 2006, p. 48): las muy fáciles las contestan todos, las muy difíciles nadie o casi nadie, con lo cual no es posible discriminar entre los evaluados. En las pruebas construidas bajo un enfoque criterial, por otra parte, "las preguntas se seleccionan de modo tal que permitan describir toda la gama de niveles de desempeño posibles, desde los más simples hasta los más complejos" (Ravela, 2006, p. 48).

La evaluación referida a normas es similar a las evaluaciones construidas bajo *enfoque psicométrico*, mientras que la referida a criterios se corresponde con un *enfoque edumétrico*. Este último se generó como reacción a la teoría psicométrica durante la década de 1960, básicamente como una crítica al uso de grupos normativos en pruebas educacionales convencionales y como respuesta a la necesidad de obtener información más explícita sobre los aprendizajes efectivos de los alumnos. Las principales diferencias entre ambos enfoques, de acuerdo con Popham y Husek (1969), pueden verse en la tabla 3.

TABLA 3

COMPARACIÓN ENTRE EL ENFOQUE PSICOMÉTRICO Y EDUMÉTRICO

	Enfoque psicométrico (evaluación referida a normas)	**Enfoque edumétrico (evaluación referida a criterios)**
Variabilidad	Central	Irrelevante
Construcción de ítems	Foco en recoger la mayor variabilidad posible.	Foco en la representación del criterio.
Análisis de ítems	Centralidad de la capacidad discriminativa.	Capacidad discriminativa del ítem es deseable, pero son aceptables ítems que no cuenten con esta propiedad si refieren a un atributo importante del criterio.
Reportes de resultados	La puntuación se entiende en referencia al grupo normativo y la interpretación de resultados se hace en relación con un grupo de referencia.	Se debe entregar información relevante para decisiones educacionales, incluyendo también la proporción de estudiantes que logran cierto criterio establecido.

Fuente: elaboración propia con base en Popham y Husek (1969).

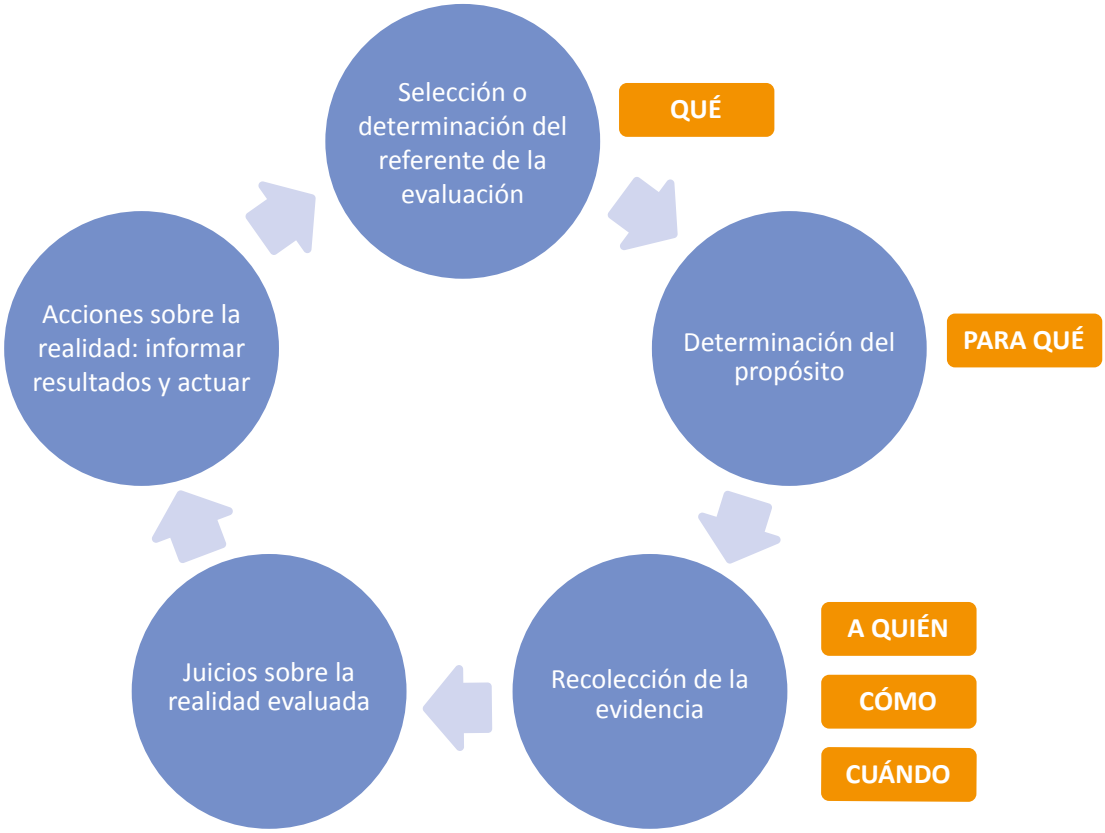
III. El ciclo evaluativo

La calidad de una evaluación se juega en cada uno de sus aspectos y componentes; en esto adquiere un papel fundamental el grado de coherencia que existe entre ellos. Diseñar una evaluación puede ser visto como un proceso continuo de toma de decisiones, que deben estar en sintonía y ser consistentes entre sí, para asegurar que la evaluación cumpla sus propósitos y entregue la información esperada, sólida y robusta.

Una forma de asegurar esta coherencia consiste en vincular cada hito o fase del diseño del instrumento con las preguntas que deben ser abordadas en él. Dicho de otro modo: la consistencia se puede entender a partir de las respuestas que el diseñador da a las preguntas ¿Qué evaluar?, ¿para qué?, ¿cómo?, ¿a quiénes? y ¿Cuándo?, como se ilustra en la figura 2.

FIGURA 2

EL CICLO EVALUATIVO



Fuente: elaboración propia.

La selección o determinación del referente responde a la pregunta *qué se evalúa*. Es el punto de partida del diseño de una evaluación, toda vez que determina las decisiones que se irán luego tomando en relación con el tipo de dispositivo más pertinente para recoger la evidencia, qué análisis de la información deberán hacerse, cuál es el tipo de resultados que se espera obtener y cuál el uso intencionado de los mismos. En algunas evaluaciones este referente es totalmente explícito; por ejemplo, el currículum en un sistema de evaluación de logros de aprendizaje de los estudiantes a escala nacional; o bien, los perfiles, parámetros e indicadores en el sistema de evaluación del desempeño profesional. En otros casos, el referente es menos evidente y se requiere realizar cierto trabajo para explicitarlo, por ejemplo, al evaluar las habilidades socioemocionales de los estudiantes (no hay un "marco curricular de las emociones").

Cuando hay un referente claro y explícito, esta fase del ciclo implica seleccionar qué será objeto de la evaluación. Lo habitual es que los referentes sean más amplios de lo

que una evaluación permite abordar; se requiere entonces priorizar en función de algún o algunos criterios (los cuales se definen a partir del propósito de la evaluación y de los usos de sus resultados). Cuando se enfrentan referentes más difusos o menos explícitos, al trabajo de selección se añade el de definición del referente o marco de la evaluación. Para ello, se puede recurrir a documentación oficial, a normativas, a modelos teóricos o conceptuales sólidos, a paneles de expertos que en forma colegiada hagan la definición esperada o a cualquier otra fuente de información que sea robusta... No se debe perder de vista que la definición de *qué evaluar* es la base de todo el proceso.[1]

Estrechamente relacionada con la definición del marco de referencia (el *qué*), se sitúa la del *para qué,* es decir, la del propósito de la evaluación. Es fundamental tener claro si se diseñará una evaluación con fines formativos o sumativos, y si se trata de una evaluación de bajas o altas consecuencias. Los distintos propósitos determinan, entre otras cosas, el tipo de reactivo que se construye, la cobertura del referente que se espera y el grado de dificultad previsto.

La claridad del *qué* y el *para qué* ayuda a decidir cuál será la evidencia a recolectar (*a quién*, *cuándo* y *cómo* se evalúa). La definición de cómo evaluar implica la selección del tipo de instrumento que se va a emplear, según se aborda en la sección siguiente.[2]

Esta fase del ciclo evaluativo suele ser la más visible, ya que en ella se construyen o adaptan los instrumentos y dispositivos con los cuales se levanta la información: es lo que el evaluado *ve* y a lo que se enfrenta, lo que los usuarios y receptores de la información quieren conocer, lo que se comenta cuando se quiere criticar o legitimar un sistema de evaluación. Existen diversos mecanismos para garantizar la calidad de los dispositivos e instrumentos, cuya importancia es nodal. Sin embargo, su visibilidad y centralidad no deben opacar, en quienes diseñan la evaluación, la necesidad de precisar de manera seria y responsable la delimitación del marco de referencia, el propósito de la evaluación, el modo en que se analizarán sus resultados y la manera en que se usarán.

El ciclo de la evaluación contempla que, una vez recogida la evidencia, se lleven adelante procedimientos para establecer juicios sobre la realidad evaluada. Estos son, en definitiva, valoraciones respecto de los datos levantados en función de un referente. Para que sean robustos, deben tener como base rigurosos procedimientos de análisis.[3]

---

[1] El tema será desarrollado con detalle en el tercer cuadernillo de esta serie.

[2] Y con más detalle, en los cuadernillos cuatro a siete.

[3] El cuadernillo ocho de la colección profundiza en esta fase del ciclo evaluativo.

Cuando ya se tienen juicios sólidos y fundados en la evidencia, el ciclo evaluativo considera la realización de ciertas acciones ejercidas sobre la realidad evaluada, siendo la más básica informar los resultados y entregar retroalimentación a los evaluados.[4]

Una forma complementaria de comprender el ciclo de vida de una evaluación es describir de forma detallada las fases o etapas que clásicamente sigue. Ello será abordado más adelante, en un apartado especial.

Por último, es importante comprender que "la vocación principal de toda evaluación es modificar la realidad, pero la evaluación por sí misma no produce cambios si no hay actores que usen los resultados y tomen decisiones a partir de las valoraciones resultantes de la misma" (Ravela, 2006, p. 42). Tener presente este fin último de la evaluación revela la importancia crucial que reviste guardar la coherencia entre las decisiones que se adoptan durante el devenir del ciclo evaluativo.

IV. ¿Cuáles son los distintos tipos de instrumentos de evaluación?

Si bien otros cuadernillos de esta colección abordan en detalle los distintos instrumentos y dispositivos de evaluación, es razonable que al abordar las "nociones básicas en medición y evaluación", se incluya un apartado donde aparezca una breve descripción de la diversidad de posibilidades para recolectar la evidencia que permitirá luego hacer los juicios en función del referente y los propósitos de la evaluación. Para ordenar esta diversidad, proponemos la siguiente clasificación:

1. Instrumentos para evaluar conocimientos: pruebas o exámenes.
2. Instrumentos para medir opiniones, actitudes y creencias: encuestas, entrevistas y cuestionarios.
3. Instrumentos para evaluar desempeños: tareas de ejecución, observación y rúbricas.

Instrumentos para evaluar conocimientos

Las *pruebas o exámenes* son probablemente los instrumentos de evaluación más ampliamente utilizados en el contexto educativo. Cuando se trata se valorar conocimientos, tienen claras ventajas sobre otros tipos de herramientas: permiten cubrir mayor extensión

[4] El noveno cuadernillo aborda con detalle este importante asunto.

de contenidos; ofrecen la opción de combinar distintos formatos de pregunta para diversificar los constructos; y abren la posibilidad de formular preguntas con diversos grados de complejidad. En el caso de las mediciones a gran escala estandarizadas, el uso de pruebas tiene además otras ventajas métricas: por una parte, la teoría de medición está más desarrollada respecto de este instrumento que de otros alternativos, lo cual facilita modelar los resultados, tanto en términos de puntuación como para establecer puntos de corte; además, en la comunidad académica existen consensos acerca de las exigencias métricas de las preguntas y los instrumentos en su conjunto, lo que respalda las decisiones en torno a su construcción y ensamblaje. A estas ventajas se suman las facilidades logísticas y operativas para su aplicación a un elevado número de evaluados.[5]

Instrumentos para medir opiniones, actitudes y creencias

Otro grupo de instrumentos o dispositivos de evaluación es aquel que permite recoger información relativa a opiniones, actitudes y creencias. En concreto, se trata de encuestas, cuestionarios y entrevistas[6].

La *encuesta* es una técnica de indagación cuantitativa que, valiéndose de un *cuestionario* -integrado principalmente por preguntas cerradas y semicerradas- y de un procedimiento sistemático y estandarizado de aplicación, permite obtener información sobre variables de interés relacionadas con los comportamientos, actitudes, valores, creencias y opiniones de una población. Una encuesta permite llegar a conclusiones generales a partir del agregado de datos individuales. Para ello, se selecciona una muestra proporcional y representativa de la población, cuya definición se hace a partir de criterios estadísticos (procedimiento de muestreo).

Mientras las encuestas buscan generalizar los resultados basándose en una muestra representativa (respuestas que se procesan estadísticamente), las *entrevistas* generan información cualitativa, rica y profunda, pero no generalizable. Pueden ser *estructuradas, semiestructuradas* o *no estructuradas*; todas ellas comparten que la evidencia se recolecta a partir de una conversación conducida por el entrevistador, en función de los propósitos y objetivos que guían la evaluación.

---

[5] En el cuadernillo cuatro se profundiza en el proceso de construcción de las pruebas, en los distintos tipos de reactivos que es posible incorporar en ellas según la naturaleza del constructo que se quiere medir y, junto con ello, se presenta el análisis de los alcances y limitaciones de este tipo de instrumento de recolección de evidencia.

[6] Las características, alcances y limitaciones de cada una de estas modalidades son abordados en profundidad en el cuadernillo cinco de la colección.

Instrumentos para evaluar desempeños

Muchas veces el interés de la evaluación no es establecer los conocimientos que una persona tiene respecto de un determinado dominio, sino su capacidad de realizar un procedimiento o ejecutar una tarea. En estos casos, *saber* no es igual a *dominar conocimientos*, sino a *ser capaz de hacer* algo de acuerdo con ciertos parámetros o estándares esperados.[7]

Merece la pena señalar que se trata de dispositivos que evidentemente lucen muy distintos a una prueba, en tanto miden a través de *tareas de desempeño* objetivos de evaluación que suponen la capacidad para llevar a cabo una actividad; es decir, deben dar cuenta no solo del dominio de conocimientos y habilidades cognitivas en el evaluado, sino también de su capacidad de ejecución de una tarea o la demostración de cierto desempeño.

Otro elemento distintivo de este tipo de instrumentos es que, junto con el diseño de las indicaciones o instrucciones para solicitar al evaluado la tarea o especificar el tipo de desempeño que se espera de él, supone siempre el uso de una herramienta complementaria, aquella que permite registrar lo observado y emitir un juicio respecto de ello: una *pauta de observación, lista de cotejo o rúbrica.* El instrumento de evaluación es, en consecuencia, un *paquete (pack)* que incluye dos elementos: las indicaciones o instrucciones al evaluado y la pauta o rúbrica con que será juzgada la ejecución o el desempeño.

Aspectos a considerar al escoger un instrumento de evaluación

Según se ha expuesto en los apartados precedentes, el abanico de posibilidades para recoger información que tiene el diseñador de una evaluación es muy amplio. La diversidad va desde un examen de álgebra con reactivos de respuesta cerrada a una entrevista semiestructurada para conocer la percepción de un apoderado sobre un programa aplicado en la escuela de su hijo, por poner un par de ejemplos. La pregunta que naturalmente surge frente a esta variedad es: ¿hay algún instrumento que sea mejor que otro, que sea más sólido o robusto en términos técnicos?

La respuesta es que ningún instrumento, por sí mismo, es mejor que otro. Lo que determina su grado de adecuación es la coherencia con la naturaleza de lo que se quiere medir o evaluar. Cuando se trata de evaluar conocimientos, un examen es por cierto más pertinente -y, por lo mismo, una opción técnicamente más robusta- que una encuesta.

[7] Los cuadernillos seis y siete abordan en profundidad este tipo de instrumentos de evaluación: pautas de observación, tareas de desempeño y rúbricas.

Cuando el objetivo es ponderar el modo en que un maestro conduce una clase, es más adecuada la observación directa de su desempeño que un examen.

Esta regla, que pone la naturaleza de lo que se va a medir como epicentro de la decisión acerca de qué instrumento de evaluación utilizar, no es el único criterio. Además, el diseñador de la evaluación debe atender consideraciones prácticas, logísticas e incluso presupuestarias. "Lo perfecto es enemigo de lo bueno", señala el adagio popular; y en este caso, si no hay condiciones para, por ejemplo, recoger evidencia con entrevistas en profundidad respecto de la valoración que hacen los actores de un programa educativo (porque no se cuenta con personal capacitado para llevarlas a cabo, transcribirlas y luego codificarlas con el fin de generar categorías explicativas), entonces una segunda buena opción es generar y aplicar una encuesta.

Al decidir qué tipo de instrumento o dispositivo de levantamiento de información es el más pertinente para alcanzar cierto objetivo de evaluación, es también útil tener en consideración el concepto de *evaluación auténtica*. Este fue acuñado a fines de los años ochenta por Grant Wiggins, del siguiente modo: "pruebas auténticas son desafíos representativos de las tareas propias dentro de una disciplina determinada. Son diseñadas para enfatizar un grado de complejidad realista (pero, a la vez, justo y razonable); enfatizan la profundidad más que la amplitud. Para hacer esto, necesariamente deben involucrar tareas o problemas poco estructurados y que tengan cierto grado de ambigüedad" (Wiggins, 1989). La evaluación auténtica propone enseñar a través de situaciones similares a las que se generan y ocupan el conocimiento en la vida real, y tienen las siguientes características:

- Son realistas y plausibles, emulan lo más posible situaciones que se dan en la vida real.
- Son complejas e intelectualmente desafiantes, implican poner en juego habilidades de valoración, creación y análisis, y activar conocimientos complejos (en el sentido de no ser datos aislados o atomizados).
- Tienen una finalidad o propósito claro, exigen buscar soluciones a situaciones novedosas.
- Implican la elaboración de un producto final dirigido a audiencias o destinatarios reales, más allá del docente.
- Ponen al estudiante en la situación de desempeñar ciertos roles similares a los que desempeñan personas en la vida real.

- Las situaciones que se plantean como contexto para la evaluación tienen restricciones e incertidumbres, de modo que los estudiantes deben ser creativos para activar sus conocimientos y habilidades al buscar una solución.

- Requieren que los estudiantes movilicen un variado repertorio de estrategias, ensayen distintas soluciones, evalúen su eficacia y pertinencia, y hagan los ajustes necesarios. Por tratarse de un proceso con *idas y venidas*, el profesor debe acompañar y retroalimentar de forma permanente a los estudiantes.

- Están pensadas para que los estudiantes trabajen en ellas por un periodo extenso de tiempo, no es una evaluación que se lleve a cabo en un momento puntual y acotado.

- Generalmente se desarrollan a partir de un trabajo colaborativo entre los estudiantes, tal como ocurre en la vida real.

- Tiene como aspectos esenciales la autoevaluación y la evaluación entre pares (Ravela, Picaroni y Loureiro, 2018).

El enfoque de la evaluación auténtica aporta una perspectiva que puede ser útil al escoger cómo se diseñarán los instrumentos para levantar información, en tanto sea posible y pertinente que se sitúe al evaluado en un contexto que emule o replique las tareas a desempeñar. Aun cuando suele ser una aproximación a la que se hace referencia en el marco de la evaluación de aprendizajes en el aula, también es de utilidad cuando se evalúan otros ámbitos y otros actores, por ejemplo, cuando en un portafolio de evaluación docente se incorporan como evidencia las tareas que un maestro desarrolla durante su trabajo cotidiano (planificar una lección o unidad pedagógica, preparar material didáctico o evaluar los aprendizajes de sus estudiantes).

En suma, optar por un dispositivo de evaluación u otro es una decisión que siempre debe tener fundamentos y estar alineada con el propósito de la evaluación y con los usos que se quiera dar a los resultados que de ella se esperan. Cabe precisar que muchas veces se pueden y deben utilizar distintos tipos de instrumentos en una misma evaluación; por ejemplo, cuando se quiere evaluar un constructo complejo desde múltiples perspectivas (o informado a partir de distintas fuentes), o bien, cuando existen altas consecuencias y es necesario contar con información completa y exhaustiva sobre el fenómeno que es objeto de interés.

V. El proceso de construcción de instrumentos de medición y evaluación

Construir un instrumento de evaluación no es una tarea simple, ni tampoco algo que se improvisa. Aun cuando existen ciertas particularidades que dependen del tipo de dispositivo, de si se trata de una evaluación de aula o a gran escala, y de los usos que se pretende dar a la información resultante, es posible distinguir ciertas etapas comunes del proceso de construcción:

1. Definición del marco de referencia.
2. Construcción de preguntas o ítems (diseño de la tarea evaluativa).
3. Revisión de expertos.
4. Ensamblaje de instrumentos para pilotaje.
5. Pilotaje y análisis psicométrico.
6. Ensamblaje definitivo.
7. Aplicación definitiva.
8. Análisis de datos aportados por la aplicación definitiva, estimación de puntuaciones.
9. Reporte de resultados
10. Acompañamiento en la interpretación y uso de resultados.

La *definición del marco de referencia* es, tal como se presentó, el punto de partida de la construcción de cualquier instrumento de evaluación -de aula o a gran escala; formativa o sumativa; de bajas o altas consecuencias; de sujetos, programas o sistemas-, toda vez que delimita con precisión qué se evalúa, cuáles son los límites y las especificaciones que luego deberán ser recogidas en los instrumentos y sobre las que se espera entregar resultados. La *construcción de preguntas, ítems o reactivos* puede ser vista como el corazón del proceso: es el segmento donde el diseñador concreta el marco de referencia de la evaluación en una herramienta mediante la que se recoge la evidencia en que se fundará el juicio evaluativo buscado. La fase siguiente, *revisión de expertos,* contribuye a respaldar la validez de contenido de la evaluación, toda vez que implica montar un proceso de análisis de las preguntas o reactivos por parte de especialistas en la disciplina o dominio evaluado, y en temas de medición educativa. Tales revisiones externas ayudan a detectar errores o imprecisiones, y garantizan la calidad de los reactivos que conforman la evaluación.

Las fases implicadas en el proceso de pilotaje o estudio de campo son propias de las mediciones estandarizadas y suelen estar ausentes en los procesos de evaluación de aula. El *ensamblaje de los instrumentos para pilotaje* consiste en organizar los ítems construidos emulando el instrumento definitivo -en términos de su correspondencia con las especificaciones de la evaluación-, para ser aplicados luego en un *estudio piloto,* es decir, a un grupo de sujetos de características equivalentes a quienes serán evaluados realmente, con el propósito de recolectar datos que permitan luego hacer un *análisis psicométrico* de los reactivos e instrumentos. Se prueban siempre más ítems de los que se necesitan para la aplicación definitiva, porque es esperable que en la fase piloto haya algunos que no satisfagan los criterios técnicos esperados.

En el *análisis psicométrico* se estudian típicamente las siguientes propiedades de los reactivos: a) grado de dificultad; b) capacidad discriminativa (grado en que el ítem permite distinguir quién sabe y quién no, quién tiene un alto nivel en el atributo medido y quién uno bajo); c) patrón de omisión; d) en caso de reactivos de opción múltiple, distribución de respuestas en los distintos distractores; y e) análisis de sesgo o funcionamiento diferencial del ítem. Para hacer estos análisis existen dos teorías de medición: la *Teoría Clásica de Test* y la *Teoría de Respuesta al Ítem*;[8] se trata de modelos estadísticos complementarios y su empleo entrega información muy valiosa (Manzi y San Martín, 2003).

Una vez seleccionados los reactivos a partir de los resultados del análisis del piloto, se *ensambla una nueva versión del instrumento* y se hace la *aplicación definitiva* (en el caso de la evaluación de aula, es la única aplicación del proceso), en la que resulta fundamental asegurar condiciones de equidad e imparcialidad para todos los evaluados. Recogida la información generada por la prueba, se hace *el análisis de los datos definitivos y la estimación de puntajes.*[9]

Según vimos al abordar el ciclo de la evaluación, siguen las fases de generación y entrega de *reportes de resultados y acompañamiento en su interpretación y uso*.[10] Conviene enfatizar que una evaluación cumple con su sentido si los datos que genera son útiles para la toma de decisiones bien fundamentadas, basadas en evidencia robusta y técnicamente sólida. El proceso de construcción de instrumentos debe seguir un riguroso trabajo en cada una de sus fases, con el fin de garantizar la confiabilidad, validez y ecuanimidad de la medición.[11]

---

[8] Ambas serán presentadas detalladamente en el cuadernillo dos.

[9] El cuadernillo ocho da cuenta de este proceso.

[10] El cuadernillo nueve trata con profundidad estos temas.

[11] En el cuadernillo dos se abordará cada uno de estos conceptos y su vinculación con el diseño de los instrumentos y sistemas de evaluación.

VI. ¿Qué se evalúa en el contexto educativo?

Según fue señalado, las evaluaciones están presentes en la práctica educativa, con diversos actores como participantes y de manera permanente. La mayor parte de las evaluaciones refieren a los logros de aprendizaje de los estudiantes, sea a nivel de aula o en mediciones a gran escala nacionales e internacionales. En este segundo caso, los programas de evaluación educativa suelen incluir estudios de factores asociados, es decir, indagan variables de los contextos escolares, familiares y sociales que se relacionan positiva o negativamente con los logros de los alumnos. Igualmente, a nivel macro, ha adquirido presencia de manera creciente en los países de la región la evaluación del desempeño de los docentes. Se trata de un asunto que merece ser profundizado al abordar la pregunta *qué se evalúa* en el contexto educativo. Finalmente, desde una perspectiva algo distinta, son frecuentes también las evaluaciones a programas específicos, por lo que se abordarán en la parte final de este apartado.

Evaluación de logros de aprendizaje

En su trabajo regular, los maestros diseñan e instrumentan evaluaciones para monitorear en qué medida sus estudiantes adquieren los conocimientos y desarrollan las habilidades que prescriben los aprendizajes esperados en currículos y planes de estudio. "En las aulas, los maestros y profesores evalúan a sus alumnos, algunas veces con el propósito de conocer qué han aprendido y cuáles son sus dificultades, de modo de ayudarlos en su proceso de aprendizaje o con el propósito de otorgarles una calificación" (Ravela, 2006, p, 17). Como muy bien señalan Ravela, Picaroni y Loureiro, mientras el discurso pedagógico parece otorgar más valor a la evaluación de los logros de aprendizaje con fines formativos que a la evaluación con fines de certificación (calificación), "en las aulas predomina el uso constante de las notas y calificaciones por encima de la evaluación formativa" (2018, p. 139). En cualquier caso, sea para un uso formativo o de calificación, la evaluación en el salón de clases se concentra casi exclusivamente en la medición del grado de apropiación que los estudiantes tienen de los objetivos de aprendizaje en las distintas asignaturas y dominios.

A nivel de evaluación estandarizada a gran escala ocurre lo mismo. Las evaluaciones nacionales —como el Plan Nacional para la Evaluación de los Aprendizajes (PLANEA) en México, Saber en Colombia o el Sistema de Medición de Calidad de la Educación (SIMCE) en Chile— son diseñadas para medir, a nivel del sistema educativo en su conjunto, los logros de los estudiantes en diferentes áreas de aprendizaje del currículo (regularmente Lenguaje, Matemáticas, Ciencias Sociales y Ciencias Naturales). Los estudios internacionales también ponen el foco en ciertos logros de aprendizaje determinados a partir de lo que paneles de expertos señalan como niveles de competencia esperados para cierta población -el caso del Programa Internacional para la Evaluación de Estudiantes (PISA, por sus siglas

en inglés)- o aquellos aprendizajes comunes a los currículos de los países participantes -el caso del Estudio Regional Comparativo y Explicativo (ERCE), que evalúa la calidad de la educación en América Latina y El Caribe liderado por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO, por sus siglas en inglés)-.

El foco en los logros de aprendizaje es coherente con el uso mayoritario de las pruebas como instrumentos de recolección de la evidencia, tal como ya se expuso. De un modo quizás *reduccionista*, los logros de aprendizaje suelen vincularse con saberes, conocimientos y habilidades en el marco de las asignaturas escolares, sin que se ponga mucho énfasis, por ejemplo, en la medición de habilidades más transversales y de un ámbito menos disciplinario, como la capacidad de trabajar en equipo, la regulación de la expresión emocional o las habilidades para la convivencia democrática y la vida en comunidad. Estos otros atributos están comenzando a adquirir cierta presencia en el debate sobre medición educativa, pero aún de manera incipiente y, por cierto, con un nivel de desarrollo y sistematización mucho menor que la evaluación de logros de aprendizaje tradicional[12].

Estudios de factores asociados con el aprendizaje

Muchos programas de evaluación educativa, nacionales e internacionales, incluyen además de las pruebas sobre logros de aprendizaje, otros instrumentos -habitualmente cuestionarios para estudiantes, profesores, directivos y padres-, con el fin de recabar información acerca de variables de los contextos escolares, familiares y sociales, y el propósito de identificar cuáles son los aspectos (o variables) que inciden en los resultados de las pruebas de logro. El objetivo último de este tipo de estudios de factores asociados consiste en informar a quienes toman decisiones a nivel de sistema, para orientar el diseño de las políticas educativas.

Obsérvese que al hablar de *factores asociados* el foco se sitúa precisamente en estudiar los grados de asociación (no de causalidad) entre estas variables y los aprendizajes de los estudiantes. Es una distinción importante, toda vez que un estudio que intentase establecer causalidad entre variables debería tener otro diseño -experimental, por ejemplo- y un marco teórico base para determinar la dirección y fuerza de los eventuales resultados.

[12] Una experiencia interesante sucede en Chile, donde desde hace algunos años el Simce contempla la medición de indicadores de desarrollo personal y social, además de la de logros de aprendizaje. A través de cuestionarios, se evalúan hábitos de vida saludable, participación y formación ciudadana, clima de convivencia escolar, autoestima y motivación escolar. Véase el sitio web del Simce, disponible en <http://bit.ly/2oJOBY8>.

Los estudios de factores asociados suelen mostrar un grupo de variables vinculadas de manera significativa con el aprendizaje de los estudiantes, las cuales pueden conceptualizarse como factores sociales: "sabemos, antes que nada, que los principales 'factores' que inciden sobre los aprendizajes son aquellos de carácter sociocultural: el nivel educativo de los padres de los alumnos, el equipamiento cultural del hogar, su situación económica" (Ravela, 2006, p. 230). A este respecto es poco lo que puede hacer la escuela, pero la evidencia es muy consistente en mostrar el alto grado de asociación entre dichos elementos y los logros escolares. Un segundo tipo de factores asociados al aprendizaje, sobre los que la escuela sí tiene injerencia, son aquellos "que pueden ser objeto de política educativa: el liderazgo educativo, el clima del centro escolar, la existencia de expectativas altas en relación al desempeño de los alumnos, la dotación de libros y textos en medios desfavorecidos, la experiencia y estabilidad de los equipos docentes, etcétera; todo aquello en que se puede intervenir a través de la toma de decisiones dentro del sistema educativo" (Ravela, 2006, p. 230).

Identificar los factores asociados al aprendizaje es, sin duda, un aporte de los programas de evaluación a gran escala para orientar la política educativa de los países y pone de manifiesto el valor de la evaluación en la toma de decisiones.

Evaluación del desempeño docente

Un ámbito que ha cobrado cada vez más presencia en la medición educativa es la evaluación del desempeño profesional docente: "Una de las actuales preocupaciones de los sistemas educativos de América y Europa es la de desarrollar sistemas de carrera docente y de evaluación del desempeño docente que contribuyan al desarrollo profesional de los maestros y profesores y, con ello, a la mejora de la calidad de la enseñanza" (UNESCO, 2007, p. 23).

Este tipo de evaluación presenta complejidades adicionales a las que tiene la evaluación asociada con los logros de aprendizaje de los estudiantes. Según Danielson y McGreal (2000), una evaluación docente efectiva debe considerar de manera especial: a) una definición coherente de lo que se espera de una buena enseñanza (el qué y la definición de un estándar de calidad); b) técnicas y procedimientos para medir todos los aspectos o componentes de la enseñanza (el cómo); y c) evaluadores capacitados y entrenados que puedan hacer juicios consistentes acerca del desempeño de los docentes a partir de la evidencia que ellos entregan en su quehacer.

La evidencia que por lo regular se escoge para emitir un juicio evaluativo acerca del desempeño docente, es variada y de distinta naturaleza. Se trata de fuentes de información tan diversas como la observación del trabajo del profesor en el aula, su autoevaluación, la

reflexión que hace respecto de sus estrategias de enseñanza, los materiales de su planeación didáctica y otros artefactos creados por el maestro (por ejemplo, guías de trabajo para los estudiantes). En menor medida, algunos sistemas de evaluación docente consideran otros tipos de evidencia, como las comunicaciones del profesor con los padres y la comunidad, la participación en actividades de perfeccionamiento y algunas muestras del trabajo de los alumnos (Danielson y McGreal, 2000; Martínez, 2016).

A escala mundial, los sistemas de evaluación docente, al igual que los de logros de aprendizaje, varían en términos de sus consecuencias: algunos buscan identificar necesidades de capacitación de los maestros, otros detectar bajos desempeños vistas a tomar decisiones de desvinculación, y un tercer grupo pretende identificar a los mejores profesores para poder entregarles incentivos económicos o simbólicos. Estos sistemas comparten la necesidad de explicitar lo que se entiende por *una enseñanza de calidad* y el objetivo de recoger evidencia variada y de distintas fuentes para consolidar un juicio evaluativo que le haga justicia a la complejidad del constructo de interés.

Evaluación de programas educativos

En el ámbito educativo es también frecuente la evaluación de programas.

> Intervenciones de alto impacto, costo efectivas y sostenibles requieren, en principio, un claro entendimiento del conjunto de los elementos, las relaciones y dinámicas que existen dentro de una determinada realidad. En la medida que exista una comprensión sólida, se podrán formular, validar, implementar y evaluar de manera más consistente las diversas acciones asociadas a la gestión de un proyecto y/o programa en sus diversos momentos (identificación de necesidades, definición de intervenciones, selección de alternativas, asignación de recursos, implementación de la estrategia y evaluación y aprendizaje) (Ortiz y Rivero, 2007, p. 3).

La evaluación de programas educativos se entiende como uno de los componentes esenciales de su proceso completo de diseño e implementación. En este contexto, existen distintos modelos o enfoques que ofrecen metodologías y herramientas para diseñar los programas y evaluarlos; los más conocidos son *marco lógico* y la *teoría de cambio* (Ortiz y Rivero, 2007).

La *metodología de marco* lógico es una forma de planificación por objetivos que se utiliza de manera esencial -pero no exclusiva- en la gestión de proyectos de cooperación para el desarrollo. A partir de que se identifican un problema y sus alternativas de solución,

propone cuatro análisis: de los involucrados, de los problemas, de objetivos y de estrategias. Suele expresarse en una matriz de marco lógico, herramienta que resume lo que el proyecto pretende hacer y cómo, cuáles son sus supuestos clave y cómo los insumos y productos del proyecto serán supervisados y evaluados (Ortegón, Pacheco y Prieto, 2005).

La *teoría de cambio*, por su parte, explica cómo acciones consistentes, de manera lógica, predecible y probadamente derivarán en el cambio deseado. Permite abordar el análisis de una situación que requiere modificarse con el fin de alcanzar un cambio positivo. Parte de una visión de éxito que identifica resultados primarios, secundarios, terciarios, etcétera -todos ellos precondiciones "unos de otros"- que articuladamente permiten alcanzar el cambio deseado. Los componentes de una teoría de cambio son: visión de éxito, precondiciones, intervenciones, supuestos e indicadores, es decir, métricas que permiten conocer si se alcanza o no el éxito en la implementación de las precondiciones (Ortiz y Rivero, 2007).

Sea cual sea el enfoque empleado, conviene tener presente que la evaluación de un programa es un componente esencial de su proceso de diseño e implementación, y que, al igual que en el caso de cualquier otro sistema de evaluación, deben tomarse resguardos para asegurar la coherencia entre sus distintos elementos y garantizar la pertinencia de las acciones y dispositivos generados para dar cuenta del objeto que interesa medir y conocer.

Consideraciones finales: ideas fuerza

En este cuadernillo se han revisado conceptos centrales de la medición y evaluación educativas. Al momento de darle un cierre, es oportuno señalar algunas ideas transversales que han guiado la exposición. La primera es que no existe una única y mejor forma de medir ni de evaluar. La calidad del proceso se vincula de manera importante con la coherencia entre sus componentes: el objetivo y propósito de la evaluación, la naturaleza del objeto que se quiere conocer, el tipo de instrumento o dispositivo que se emplea para levantar información, el tipo de resultados que se espera obtener y el uso que se les quiere dar. Es tal consistencia la que contribuye de manera privilegiada a la solidez técnica de la evaluación.

En segundo lugar, cabe destacar que diseñar una evaluación es un proceso en el cual continuamente se toman decisiones. No hay una *receta* o un conjunto de fórmulas que se puedan seguir al pie de la letra para garantizar el éxito; por el contrario, quien diseña debe ir, en cada fase del proceso, evaluando las posibilidades que se le presentan, analizando los pros y contras de cada una, y escogiendo la más pertinente y viable de habilitar. En este sistema de decisiones se atienden y sopesan los elementos mencionados en el cuadernillo, y que se desarrollarán con detalle en los siguientes: el propósito perseguido, las consecuencias involucradas, los objetos de medida, las herramientas técnicas disponibles, los usuarios de la información, etcétera.

Finalmente, el interés y cuidado en diseñar evaluaciones de calidad no es un mero ejercicio intelectual: encuentra su fundamento en la necesidad de obtener información robusta, sólida y bien fundamentada. Si se van a tomar decisiones a partir de los datos de una medición o evaluación, deben ser de calidad; de lo contrario se puede crear una falsa ilusión de certezas, lo que posiblemente repercuta de manera no deseada en las decisiones y posteriores intervenciones. Evaluar, evaluar bien, es un imperativo para conducir el mejoramiento educativo.

En una época en la cual los sistemas educativos y los países invierten costosos recursos en medición y evaluación, es importante tener presente que evaluar no es un fin en sí mismo: es siempre un medio al servicio del cumplimiento de un propósito; un propósito que demanda información válida y confiable para orientar decisiones y diseñar soluciones.

Referencias

BLACK, P. y Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7-74.

BLOOM, B., Hastings, J. y Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. Nueva York: McGraw-Hill.

BROOKHART, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative Classroom Assessment: Theory into Practice* (pp. 43-62). Nueva York: Teachers College Press.

BROOKHART, S. M. (2009). Editorial. *Educational Measurement: Issues and Practice, 28*(1), 1-2.

COMMITTEE ON ASSESSMENT IN SUPPORT OF INSTRUCTION AND LEARNING, COMMITTEE ON SCIENCE EDUCATION K-12 y National Research Council. (2003). *Assessment in support of instruction and learning: Bridging the gap between large-scale and classroom assessment*. Washington, DC: The National Academies Press. https://doi.org/10.17226/10802

DANIELSON, C. y McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Princeton: Educational Testing Service.

FENTON, N.E. y Pfleeger, S.L. (1997). *Software metrics. A rigorous and practical approach*. Londres: CRC Press.

HATCH, T. (2013). Beneath the surface of accountability: Answerability, responsibility and capacity-building in recent educational reforms in Norway. *Journal of Educational Change, 14*(1), 1-15.

MANZI, J. y San Martín, E. (2003). La necesaria complementariedad entre la teoría clásica de la medición (TCM) y teoría de respuesta al item (IRT). *Estudios Públicos, 90*, 145-183.

MARTÍNEZ RIZO, F. (2009). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa, 11*(2), 1-18.

MARTÍNEZ RIZO, F. (2012). La evaluación formativa del aprendizaje en el aula en la bibliografía en inglés y francés. Revisión de literatura. *Revista Mexicana de Investigación Educativa, 17*(54), 849-875.

MARTÍNEZ RIZO, F. (2016). *La evaluación de docentes en educación básica. Una revisión de la experiencia internacional*. México: INEE.

ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA EDUCACIÓN, LA CIENCIA Y LA CULTURA [UNESCO]. (2007). *Evaluación del desempeño y carrera profesional docente. Un estudio comparado de 50 países de América y Europa*. Santiago: Andros Impresores.

ORTEGÓN, E., Pacheco, J. y Prieto, A. (2005). *Metodología del marco lógico para la planificación, el seguimiento y la evaluación de proyectos y programas*. Santiago: ILPES.

ORTIZ, A. y Rivero, G. (2007). *Desmitificando la teoría de cambio*. PACT. Recuperado de http://www.rootchange.org/about_us/resources/publications/DemistificandolaTeoriadeCambio.pdf

PATTON, M. (2002). *Qualitative research and evaluation methods* (3th ed). California: Sage.

POPHAM, J. y Husek, T. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9.

RAVELA, P. (2006). *Para comprender las evaluaciones educativas. Fichas didácticas*. Montevideo, Uruguay: PREAL.

RAVELA, P., Picaroni, B. y Loureiro, G. (2018). *¿Cómo mejorar la evaluación en el aula? Reflexiones y propuestas de trabajo para docentes.* Montevideo: Grupo Almagro Editores.

SADLER, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119-144.

SCRIVEN, M. (1967). *The methodology of evaluation*. En R. Tyler, R. Gagné y M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (pp. 39-83). Chicago, U.S.: AERA Monograph Series on Curriculum Evaluation.

SCRIVEN, M. (2013). The foundation and future of evaluation. En S. L. Donaldson (Ed)., *The future of evaluation in society. A tribute to Michel Scrive* (pp. 11-44). Estados Unidos: Information Age Publishing Inc.

STIGGINS, R. (2008). *Assessment manifiesto: A call for the development of balanced assessment systems*. Portland: ETS-ATI.

WILIAM, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.

Instituto Nacional para la Evaluación de la Educación
México

Centro UC
Medición - MIDE