Cuadernillo técnico
de evaluación educativa

4

Desarrollo de instrumentos de evaluación: pruebas

Centro UC
Medición - MIDE

Textos de divulgación

INEE
Instituto Nacional para la Evaluación de la Educación
México

Cuadernillo técnico
de evaluación educativa

4

Desarrollo de instrumentos de evaluación: pruebas

Centro UC
Medición - MIDE

Textos de
divulgación

INEE
Instituto Nacional para la
Evaluación de la Educación
México

Desarrollo de instrumentos de evaluación: pruebas

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

Autoras
Patricia Mahias Finger, MIDE UC
María del Pilar Polloni Erazo, 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 1

Resumen 2

Introducción 3

Etapas en la construcción de una prueba 4

Construcción de preguntas cerradas 6

Construcción de preguntas abiertas o de respuesta construida 15

Proceso de revisión de ítems por equipos y expertos 20

Procesos de pilotaje de pruebas: cuantitativo y cualitativo 25

Ensamblaje de pruebas definitivas 29

Establecimientos de puntos de corte (standard setting) 30

Consideraciones finales: ideas fuerza 32

Referencias 34

Anexo 1: Criterios para la elaboración de ítems de opción múltiple 36

Anexo 2: Resolución del ejercicio 1: Análisis de una pregunta abierta 37

Anexo 1: Resolución del ejercicio 2: Análisis crítico de ítems cerrados 39

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.

Desarrollo de instrumentos de evaluación: pruebas

Resumen

En este cuadernillo se describe el proceso de construcción de un instrumento de evaluación, específicamente de una prueba, distinguiendo las etapas que lo componen. Se exponen las principales características de cada una de las fases de construcción y los criterios que se deben considerar en la elaboración y revisión de ítems cerrados y abiertos. Se especifican, además, los principales actores involucrados en cada instancia de revisión. Adicionalmente, se presentan los procesos de pilotaje, que permiten verificar el funcionamiento y calidad de los ítems, haciendo referencia a estudios cualitativos y cuantitativos. Por último, se explican de modo general los procesos que permiten establecer puntos de corte, también conocidos por su nombre en inglés, *standard setting*, para aquellos instrumentos que requieren fijar el desempeño de los evaluados a través de categorías. De esta forma, se presentan los principales criterios técnicos a tener en cuenta en los pasos que se requiere seguir para construir pruebas, considerando que este tipo de instrumento es uno de los más ampliamente utilizados a nivel internacional para evaluar logros de aprendizaje de estudiantes en el sistema educacional.

Introducción

El presente documento tiene como propósito realizar una descripción de las principales etapas que contiene el desarrollo de una prueba, específicamente la elaboración de ítems cerrados y abiertos, considerando que su correcta construcción converge directamente en la evidencia de validez de contenido de un instrumento. El desarrollo de una prueba implica el proceso de medir algún aspecto del conocimiento, habilidades, capacidades, intereses, actitudes u otras características de un individuo, mediante preguntas, tareas o la combinación de estas. Es posible señalar que una buena prueba está constituida por un conjunto de buenos ítems, que cubren de manera satisfactoria el objeto de interés y cuyos procedimientos de diseño y desarrollo deben respaldar la validez de las interpretaciones de los puntajes para sus usos previstos (American Educational Rese arch Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME], 2014).

Este cuadernillo abordará el proceso de desarrollo de una prueba a través de los siguientes temas. En la sección I se describirán las etapas que contiene la construcción de un instrumento de evaluación, abarcando desde la delimitación del contenido hasta la elaboración y ensamblaje de los ítems específicos que lo constituyen. En la sección II se detallarán las características y recomendaciones para la elaboración de diversos tipos de ítems cerrados, tales como: términos pareados, ítems de verdadero o falso, de completación e ítems de selección múltiple. En la sección III se abordará la construcción de preguntas abiertas, distinguiendo cuándo es pertinente utilizarlas y qué consideraciones se deben tener en cuenta para su elaboración. En la sección IV se expondrán las diversas instancias de revisión de un ítem y los criterios a considerar en cada una, entendiendo que son fundamentales para resguardar la evidencia de validez de contenido del instrumento en construcción. Una vez hecho esto, en la sección V se explicará de modo general en qué consisten los procesos de pilotaje de las pruebas, los cuales permiten la verificación del funcionamiento y calidad de los ítems desde un enfoque cuantitativo y/o cualitativo. En la sección VI se presentará cómo debe llevarse a cabo el proceso de ensamblaje de un instrumento, es decir, la descripción de cómo se arma una prueba para su aplicación definitiva. Por último, en la sección VII se finalizará este documento con una referencia general al procedimiento para establecer puntos de corte, o *standard setting*, en aquellos instrumentos que requieren distinguir el desempeño de los evaluados a través de categorías. De esta forma, se completan las orientaciones técnicas asociadas con la construcción de pruebas como instrumentos de evaluación.

I. Etapas en la construcción de una prueba

El proceso de elaboración de instrumentos de evaluación contempla una serie de etapas que se relacionan, por una parte, con la delimitación del instrumento en su conjunto y, por otra, con la construcción de los ítems específicos que lo constituirán. Cada una de estas fases debe darse de forma secuenciada y su adecuada ejecución es determinante para la siguiente. La figura 1 muestra las fases involucradas en el proceso de construcción de una prueba.

FIGURA 1

ETAPAS DE CONSTRUCCIÓN DE UNA PRUEBA

Propósito y destinatario del instrumento

El primer paso del ciclo de evaluación es establecer el propósito del instrumento y definir la población a la que está dirigido. Esto es especialmente relevante si se considera que las pruebas deben diseñarse de tal forma que respalden la validez de las interpretaciones de los puntajes de la prueba para los usos previstos (AERA, APA y NCME, 2014).

Especificaciones para la evaluación

Luego de haber definido el objetivo de evaluación, se debe delimitar con claridad cuál es el constructo que se pretende medir. Para ello, y dado que un constructo no siempre es accesible en forma directa para ser evaluado, se hace necesario operacionalizarlo, distinguiendo elementos más concretos y observables que lo definan y caractericen, de modo que sean susceptibles de ser sometidos a medición. En este proceso, se hace una definición exhaustiva de los contenidos y habilidades involucrados, cuidando que no falte ni sobre ningún elemento central, ya que esta especificación será el referente para la construcción del instrumento.

La definición del constructo puede organizarse de varias formas: como un listado de contenidos, a través de mapas conceptuales con sus contenidos e interrelaciones, o a partir de tablas de especificaciones (Moreno, Martínez y Muñiz, 2015). Se sugiere utilizar el formato "tabla de especificaciones" que, en términos concretos, se organiza como una matriz donde se especifican los temas y subtemas que serán evaluados; se determinan el o los objetivos de evaluación de cada uno y, a partir de ellos, se definen los indicadores que plantean las acciones observables y, por lo tanto, medibles, que deben ser capaces de realizar los evaluados. Finalmente, se detallan los pesos porcentuales de los temas y subtemas, con el fin de distinguir cuáles deben tener mayor o menor presencia en el instrumento final. El procedimiento de elaboración de tablas de especificaciones es abordado en profundidad en el cuadernillo 3 de esta serie.

Definición del tipo de ítems y extensión del instrumento

La decisión sobre el tipo de ítems que deben ser incluidos en un instrumento dependerá del objetivo de evaluación, de la cantidad de contenidos que se busque evaluar y de los recursos disponibles. Si se pretende evaluar una gran cantidad de contenidos, será más pertinente considerar ítems de selección múltiple antes que preguntas abiertas. Si, por el contrario, interesa evaluar habilidades o contenidos que requieren de una mayor profundización o elaboración de las respuestas como, por ejemplo, el proceso reflexivo que se lleva a cabo ante una determinada tarea, será más adecuado incluir preguntas abiertas. Un factor relevante a considerar son los recursos y el tiempo con que se cuenta para la construcción del instrumento, ya que son determinantes, principalmente para decidir sobre la amplitud del instrumento y sus posibilidades de cumplir con todas las etapas necesarias para su proceso de construcción.

Proceso de elaboración de ítems de pruebas

El proceso de elaboración de ítems de pruebas debe cumplir con ciertas fases que son fundamentales para lograr que sean de buena calidad (Haladyna, Downing y Rodriguez, 2002; Moreno, Martínez y Muñiz, 2004). Dichas fases, que se desarrollarán con detalle en las secciones siguientes, involucran la construcción individual de preguntas cerradas o abiertas, y procesos de revisión de equipos y expertos.

Estudio piloto

Una vez construidos los ítems, se ensambla una prueba y se aplica a una población similar a la que será evaluada posteriormente. De esta forma, se realiza un estudio piloto a partir del cual se analizan el funcionamiento de los ítems y del instrumento, con el

objeto de realizar los ajustes necesarios antes de su aplicación definitiva. En secciones siguientes, se describen las orientaciones que guían el análisis de datos cuantitativos y cualitativos en procesos de pilotaje.

Ensamblaje del instrumento definitivo

Luego del análisis y ajuste de los ítems a partir de la aplicación piloto, se procede al ensamblaje o armado de la prueba definitiva. Una vez que el instrumento es aplicado, se realiza el análisis de resultados del mismo, en que se considera la obtención de puntajes de los evaluados, y se realizan procedimientos estadísticos para analizar propiedades métricas del instrumento. El detalle de este procedimiento se desarrolla en profundidad en el cuadernillo 8 de la serie. Luego de la aplicación definitiva corresponde llevar a cabo la entrega del reporte de resultados y la retroalimentación, lo cual se expone puntualmente en el cuadernillo 9.

II. Construcción de preguntas cerradas

Con el propósito de recoger evidencia sobre el constructo a evaluar, se deben escoger una estrategia y un instrumento de evaluación que permitan observar si los desempeños asociados están o no logrados. De acuerdo con la naturaleza del aprendizaje, se selecciona y construye un instrumento o se diseña una situación evaluativa. En esta sección se describirán las características de diversos tipos de ítems utilizados en pruebas, específicamente de respuesta cerrada, que permiten mapear una amplia variedad de conceptos y habilidades en forma eficiente.

Construcción de ítems cerrados: términos pareados, de completación, de ordenamiento, y verdadero y falso

A continuación, se describirán brevemente cuatro tipos de ítems de respuesta cerrada: términos pareados, de completación, de ordenamiento, y verdadero y falso. Posteriormente, se delineará la construcción de ítems de selección múltiple, que son los más ampliamente utilizados en pruebas, por lo que se abordarán en mayor profundidad.

Términos pareados o pareamiento de términos y conceptos

Se refieren a aquellos ítems en los que se debe establecer una correspondencia entre dos conjuntos de términos o conceptos sencillos. Permiten medir fácilmente conocimientos puntuales y específicos.

Recomendaciones para su construcción:

- En cada columna, utilizar conceptos del mismo orden lógico.
- Colocar más elementos en la segunda columna que en la primera, para no otorgar puntos por descarte.

Ejemplo de ítem de términos pareados

En la columna A se encuentra una lista de capas de la atmósfera y en la columna B aparece una breve descripción de cada una de ellas. Ubique el número de la columna A en la descripción que corresponda de la columna B.

Columna A	Columna B
1. Tropósfera 2. Estratósfera 3. Termósfera	___ En esta capa hay presencia de ozono (ozonósfera). ___ En esta capa, los gruesos gases son los responsables de frenar los meteoritos. ___ En esta capa se producen los fenómenos meteorológicos. ___ En esta capa se producen las auroras boreales, vistas por lo regular en las regiones polares.

Ítems de completación

Son aquellos en los que se deben completar oraciones o frases en donde faltan ciertas palabras que expresan conceptos. Permiten medir fácilmente conocimientos puntuales y específicos.

Recomendaciones para su construcción:

- Los espacios deben tener aproximadamente la misma extensión.
- Usar un espacio en blanco por cada palabra a completar.
- Evitar frases ambiguas o demasiado extensas.
- Utilizar este tipo de ítems cuando el universo de posibilidades de respuesta sea acotado.

Ejemplo de ítem de completación

Complete los conceptos que faltan en cada línea de las siguientes oraciones:

A. Un nucleósido consta de _____ y _____; un nucleótido de _____ más _____.

B. La subclase _____ se caracteriza por la presencia de una bolsa o marsupio.

Ítems de ordenamiento u ordenamiento de términos o conceptos

Son aquellos en los que se deben ordenar términos o conceptos sencillos siguiendo un criterio lógico (por ejemplo, ocurrencia temporal). Estos ítems permiten medir fácilmente conocimientos puntuales y específicos.

Recomendaciones para su construcción:

- Usar este tipo de ítems cuando el orden solicitado sea indiscutido e inequívoco.
- El criterio para ordenar debe ser "objetivo", no estar sujeto a apreciaciones o puntos de vista. A menos que ese punto de vista se declare explícitamente.

Ejemplo de ítem de ordenamiento

Indique, usando números del 1 (primero) al 4 (último), el trayecto sucesivo que recorre un espermio desde su producción hasta su emisión en la eyaculación.

___ túbulos seminíferos
___ epidídimo
___ vaso deferente
___ uretra

Ítems de verdadero y falso

Son los ítems en que el evaluado debe pronunciarse sobre la veracidad o falsedad de una afirmación. Permiten medir de manera relativamente sencilla conocimientos específicos, pero un poco más complejos que los abordados en los ítems anteriores.

Recomendaciones para su construcción:

- Seleccionar ideas o conceptos lo menos discutibles posible y evitar afirmaciones que expresen opiniones, a no ser que ellas se relacionen explícitamente con una fuente y así sea enunciado en el ítem o prueba.
- Incluir solo una idea central por ítem y cuidar que el carácter verdadero o falso de la afirmación esté dado por esta y no por un detalle trivial en su contenido.
- Evitar el uso de expresiones que en sí mismas importen ambigüedad o vaguedad (algunas veces, posiblemente, con frecuencia, etc.). La interpretación de estas expresiones es tan difícil de objetivar que resta validez al juicio entregado como medida del aprendizaje.
- Evitar expresiones terminantes o absolutas (siempre, todos o nunca), salvo que se aplique para la afirmación en forma inequívoca.
- Redactar los ítems en forma tan clara y sencilla como sea posible.
- Ordenar las preguntas al azar e informarlo así a los evaluados.

Ejemplo de ítem verdadero o falso

Escriba una X en la columna que corresponde, según si el enunciado expuesto es verdadero o falso.

	Verdadero	Falso
La fiebre por dengue es una enfermedad viral leve transmitida por mosquitos y caracterizada por fiebre, erupción, dolor articular y muscular.		
Existen vacunas que previenen algunos tipos de dengue.		
La fiebre por dengue hemorrágico es más dañina que la fiebre por dengue.		
Una persona que sufrió un tipo de dengue nunca volverá a padecer dengue por ninguno de los virus.		

Ítems de opción o selección múltiple

Constituyen el formato más ampliamente utilizado en evaluaciones a gran escala. Evalúan la capacidad del estudiante para seleccionar por escrito la respuesta correcta de entre varias opciones que se le suministran. Son fáciles de puntuar, pero difíciles de construir. Una de sus ventajas es que admiten la evaluación de una amplia gama de contenidos y habilidades (Gierl, Bulut, Guo y Zhang, 2017; Moreno et al., 2015).

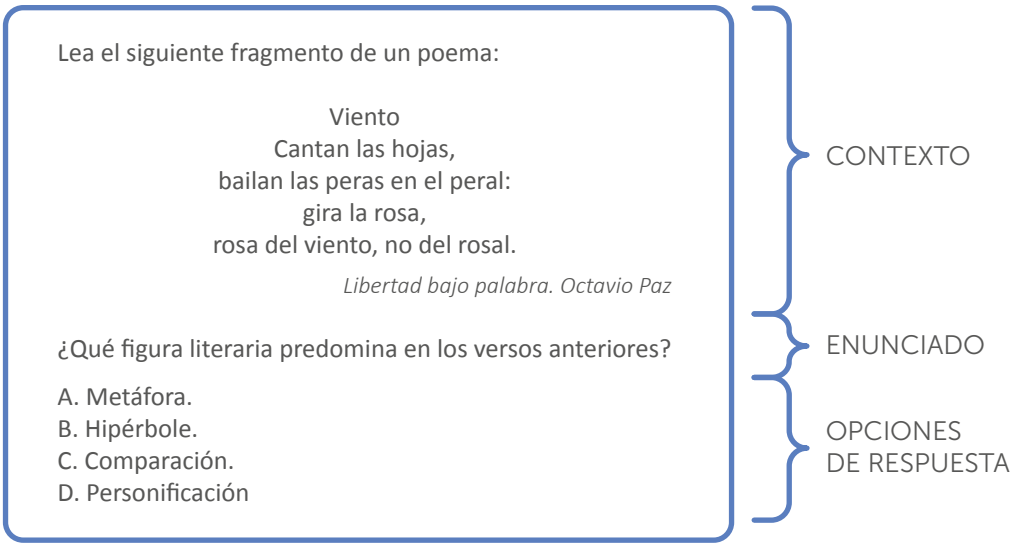
Consideraciones generales para la elaboración de ítems de opción múltiple:

- Deben ser elaborados en concordancia con la tabla de especificaciones y abordar aspectos relevantes de cada disciplina. Por ello, tienen que construirse con el objetivo de evaluar un indicador presente en la tabla de especificaciones. Es relevante que el ítem sea coherente con el indicador declarado en la tabla, a fin de asegurar que mida lo que se propuso evaluar.
- Es importante elaborarlos cuidando su claridad y brevedad, por ello es fundamental usar adecuadamente la gramática y puntuación; evitar complejidades innecesarias en el uso del lenguaje y disminuir al mínimo necesario la extensión del texto en cada elemento del ítem.
- Es preciso cuidar que no presenten sesgo, evitando hacer alusión a situaciones contextuales que sean más familiares y que, por lo tanto, favorezcan a ciertos grupos de la población. Además, se debe cuidar que no reproduzcan estereotipos raciales, culturales o de género (por ejemplo, mujeres cocinando y hombres en la oficina).

Los ítems de selección múltiple presentan en su estructura el contexto, el enunciado y las opciones de respuesta, tal como se muestra en la figura 2.

FIGURA 2

ESTRUCTURA DE UN ÍTEM DE OPCIÓN MÚLTIPLE



Para una correcta elaboración de ítems de selección múltiple, es fundamental tener en cuenta una serie de criterios respecto de su contexto, enunciado y opciones de respuesta (Haladyna y Downing, 1989a, 1989b). Estos se presentan a continuación:

Contexto

Corresponde a la descripción de una situación que sirve como base para que el evaluado tenga un punto de referencia concreto al momento de enfrentarse al enunciado. Puede ser una imagen, un texto, una situación, un gráfico, etcétera.

Recomendaciones acerca del contexto:

- Incluir contextos necesarios y verosímiles que permitan el uso o aplicación del conocimiento. Cuidar que la dificultad del ítem no se vea artificialmente aumentada por la inclusión del contexto.
- En relación con el uso de imágenes, utilizar cuando sean fundamentales para responder el ítem o para motivar a los evaluados a responder. Evitar que la imagen distraiga o entregue pistas que ayuden a responder.

Enunciado

Corresponde a la pregunta o tarea concreta que se le solicita al evaluado.

Recomendaciones acerca del enunciado:

- Expresar claramente la tarea que se demanda al evaluado. Es recomendable que esté formulado como una pregunta y no como una frase inconclusa.
- Debe estar planteado en positivo. Si es inevitable ocupar palabras como "no" o "excepto", se sugiere subrayarlas o destacarlas.

Si el contenido a evaluar requiere de una graduación en las respuestas, en el enunciado se puede preguntar por "la mejor respuesta". En estos casos se debe especificar claramente el criterio de graduación, es decir, con qué lógica una de las opciones es la correcta y las otras no.

Algunos ejemplos de preguntas de este tipo son:

- ¿Cuál de las siguientes acciones permite enfrentar de manera más directa el problema de...?
- ¿Cuál de las siguientes herramientas permite fabricar la pieza X con mayor precisión?

Existen también los denominados ítems de doble proceso, que en su estructura presentan un enunciado con una serie de afirmaciones, y las opciones de respuesta están constituidas por su combinación. Se suelen utilizar cuando es difícil encontrar cuatro opciones plausibles, no obstante, se recomienda evitar este tipo de preguntas, ya que son muy difíciles de construir, pues se debe cuidar que no se respondan por el uso de reglas lógicas de combinación de las afirmaciones y, así, evitar la "varianza irrelevante para el constructo".

Ejemplo de ítem de doble proceso

Respecto al principio escalar de Fayol se puede decir que:

I. Representa la distribución de la autoridad en una organización.

II. Representa las líneas formales de comunicación en una organización.

III. Describe tres niveles de toma de decisiones.

IV. Representa al conjunto de personas que integran una organización clasificadas por grado o rango de autoridad.

V. Representa al conjunto de personas que integran unas organizaciones clasificadas según el cargo que desempeñan.

A) I – III – IV

B) II –III – IV

C) I – III – V

D) II – III – V

Opciones de respuesta

Corresponden al número de respuestas plausibles que puedan derivarse del enunciado, entre las cuales existe una y solo una que es correcta. Todas las opciones deben parecer posibles respuestas a la pregunta planteada. En general, es recomendable el uso de cuatro opciones (la correcta y tres incorrectas).

Recomendaciones acerca de las opciones de respuesta:

- Deben mantener la misma o similar estructura gramatical y ser concordantes con el enunciado.
- Deben ser lo más directas posibles e independientes unas de otras.
- Evitar las opciones muy diversas y con diferente nivel lógico.
- Evitar el uso de opciones del tipo "Ninguna de las anteriores" o "Todas las anteriores".
- Deben tener una extensión similar.
- Ordenarlas lógica o numéricamente.

La respuesta correcta debe ser completa y claramente correcta y, además, la mejor opción, es decir, debe ser claramente identificable cuando el evaluado conoce bien el concepto por el cual se pregunta.

Recomendaciones acerca de la respuesta correcta (también llamada clave):

- Debe responder a la pregunta planteada tanto en el contenido (lógica conceptual) como en lo formal (lógica gramatical).
- Si no atenta contra la lógica en que fueron ordenadas las respuestas, se debe variar su localización en distintas preguntas.
- Evitar el uso de determinantes específicos (siempre, nunca, completamente, absolutamente) y cuidar que no presente alguna marca textual con el enunciado de la pregunta. Por ejemplo, si se interroga por una actividad que trabaje "conciencia fonológica" y la respuesta correcta es la única que contiene el concepto "fonológico", el evaluado puede seleccionar esa opción sin saber del contenido.

Las opciones incorrectas o distractores están diseñados para atraer a los evaluados con menor dominio o conocimiento sobre lo que se pregunta, por lo que se espera que sean "atractores", presentando errores conceptuales o razonamientos equivocados. Es decir, no son cualquier opción incorrecta, sino que deben estar diseñados para informar acerca de errores típicos.

Recomendaciones acerca de las opciones incorrectas (distractores):

- Deben ser incorrectas, pero plausibles. Evitar el uso de distractores absurdos o irrisorios.
- No deben atraer a través de "trampas" (sutilezas en la forma de preguntar, uso de habilidades diferentes a la que se busca evaluar en el ítem).

En el anexo 1 se entrega una síntesis de los principales criterios de elaboración de ítems de opción múltiple que han sido revisados en esta sección.

Ejemplos de buenos ítems

A continuación, se muestran dos ejemplos de ítems adecuadamente construidos, que cumplen con los criterios de construcción presentados.

Ejemplo A de ítem de opción múltiple

Observe la siguiente imagen:

¿Qué civilización precolombina desarrolló el tipo de cultivo que se muestra en la imagen?

A) Inca.

B) Maya.

C) Olmeca.

D) **Azteca.**

En este ítem se observa un contexto necesario para responder, ya que se requiere observar el tipo de cultivo que se está desarrollando para así identificar que la respuesta correcta es la civilización azteca. Además, el enunciado es claro y preciso, ya que contiene el criterio por el cual el evaluado podrá distinguir la respuesta correcta.

Ejemplo B de ítem de opción múltiple

Lea el siguiente texto:

En el intestino grueso del organismo humano viven ciertas bacterias capaces de transformar restos de alimentos en vitaminas, las que podemos aprovechar para nuestro proceso de nutrición.

¿Qué tipo de relación se establece entre esas bacterias y los seres humanos?

A) Parasitismo.

B) **Mutualismo.**

C) Amensalismo.

D) Comensalismo.

En este ítem se observa que las opciones responden a lo solicitado en el enunciado, ya que todas corresponden a tipos de relaciones entre organismos y son plausibles. Es decir, no corresponden a opciones irrisorias o sin sentido, sino a errores frecuentes en el aprendizaje de este contenido.

III. Construcción de preguntas abiertas o de respuesta construida

Hay ocasiones en las que, para evaluar cierto tipo de habilidades, resulta más pertinente utilizar preguntas abiertas en lugar de ítems de selección múltiple, por ejemplo, cuando se espera que una persona produzca una respuesta, porque importa observar tanto el producto, como el proceso; o en los casos en los cuales, de acuerdo al indicador, no existe una sola respuesta correcta, sino que hay múltiples formas de responder a una tarea. Esto es especialmente importante si se busca abordar indicadores de evaluación que contengan habilidades vinculadas a "crear" o "evaluar" (Anderson y Krathwohl, 2001), ya que implican, por ejemplo, desarrollar un análisis crítico o reflexión sobre algún tema, o bien, generar o elaborar algún producto, a partir de la estructuración u organización de diversos elementos.

Un último aspecto relevante a considerar al momento de decidir si se evaluará a través de preguntas abiertas, es revisar si es factible llevar a cabo un proceso de codificación que asegure confiabilidad, ya que la codificación de las preguntas abiertas involucra la implementación de un procedimiento de corrección con jueces revisores previamente capacitados, lo que tiene una serie de implicaciones, tanto económicas como logísticas, que conviene estar seguros de que sea posible asumir.

Ejemplo de indicadores susceptibles de evaluar mediante pregunta abierta

A continuación, se presentan ejemplos de indicadores para evaluar el desempeño docente en los que, de acuerdo con su foco, sería recomendable emplear una pregunta abierta.

Ejemplos:

- Diseñar estrategias de aprendizaje que consideren los conocimientos previos de los estudiantes para la enseñanza de los objetivos de la disciplina.
- Identificar variadas estrategias de evaluación adecuadas para medir la lectura inicial, considerando los componentes clave del aprendizaje lector (desarrollo de la conciencia fonológica, conciencia de la palabra, fluidez y precisión lectora, desarrollo de la lengua oral y vocabulario, conocimiento de lo impreso).
- Disponer de ejemplos, analogías y explicaciones simples para enseñar conceptos propios de la disciplina.

Descripción de las características de un buen ítem de respuesta abierta

Lo fundamental a tener en consideración cuando se elaboran preguntas abiertas, es que en su formulación debe quedar explícito aquello que se espera que el evaluado realice, vale decir, este debe tener, a partir del reactivo, toda la información necesaria para orientar su respuesta o desempeño. De acuerdo con esto, es preciso considerar que una buena pregunta abierta deberá indicar, como mínimo, cuáles son los criterios con base en los cuales se va a juzgar la calidad de la respuesta y, para ello, deberá considerar siempre en su elaboración, la rúbrica o pauta de corrección con la cual se evaluarán las respuestas.

Recomendaciones para elaborar una pregunta abierta:

- **Usar un lenguaje claro y directo:** Se espera que se cumplan los mismos criterios que se proponen para la formulación de ítems de selección múltiple, cuidando una adecuada aplicación de la gramática y la puntuación, además de ser claro, breve y preciso, evitando complejidades innecesarias en el uso del lenguaje. En este sentido, se propone, en la formulación de las preguntas el empleo de verbos que refieran a acciones precisas que orienten cómo se debe responder, por ejemplo, "Defina", "Calcule", "Compare", "Diseñe", evitando el uso de términos ambiguos como "Escriba acerca de...", "Comente...", sin explicitar criterios sobre lo que debe incluir ese comentario. Asimismo, es muy importante que, en la pregunta se especifique el nivel de detalle que se espera de la respuesta. Por ejemplo, se deberá

decir, "Entregue dos ejemplos...", en lugar de: "Ejemplifique", ya que esta última es una formulación muy general que admite que se entregarán uno, dos, tres o más ejemplos, lo que conllevaría a que algunos examinados presenten respuestas incompletas, solo por la falta de precisión de la instrucción.

- **Evitar la inclusión de elementos innecesarios como parte de la pregunta:** Conviene usar contextos e imágenes solo cuando sean necesarios para responder la pregunta.

Rúbricas o pautas de evaluación de preguntas abiertas

Como se mencionó, en la elaboración de una pregunta abierta se debe considerar siempre la rúbrica con la que será evaluada, ya que entrega las especificaciones respecto de los criterios con los que se juzga el desempeño de quienes responderán la pregunta abierta. Por lo anterior, es fundamental que ambas, pregunta abierta y rúbrica, mantengan completa coherencia, con el fin de asegurar que la instrucción en la pregunta corresponda a lo evaluado.

Las consideraciones específicas para la construcción de rúbricas son descritas en el cuadernillo 7 de esta serie. En términos generales, es posible establecer que una rúbrica plantea diferentes niveles de desempeño –se recomiendan no menos de tres, ni más de cinco– en los que se ofrece una descripción precisa de los criterios que caracterizan las posibles respuestas esperadas para cada nivel, distinguiendo el piso de desempeño esperado, vale decir, el mínimo necesario para quedar calificado en ese nivel. El nivel más bajo generalmente se considera "residual", entendiendo con ello que, en lugar de definir una posible respuesta para el nivel, solo se establece que "no se cumple con las condiciones definidas en el nivel anterior".

Una vez elaborada una pregunta abierta con su rúbrica, es necesario que, al igual que los ítems de selección múltiple, pase por algunas fases de revisión para asegurar su calidad. En este caso, es importante considerar como revisiones imprescindibles las realizadas por el experto en medición y el experto disciplinario. Luego de ello, es fundamental que sea aplicada en un estudio piloto, con el fin de realizar los ajustes que sean necesarios.

Requisitos de una buena rúbrica para evaluar una pregunta abierta:

- Que mantenga coherencia con aquello que se solicita en el enunciado.
- Que logre hacer una distinción clara entre los niveles, estableciendo con precisión las características que diferencian las respuestas de cada nivel.

A continuación, se presenta un ejemplo de una pregunta abierta y su rúbrica de corrección, elaborada para evaluar la producción escrita de estudiantes de nueve años de edad. El indicador a evaluar es: "Crear un cuento a partir de un estímulo".

Ejemplo de pregunta abierta y rúbrica

Escribe un cuento a partir de la imagen y el título.
Recuerda cuidar tu ortografía, usar letra clara y punto cuando corresponda.

El fantasma miedoso

En el ejemplo:

- Se entrega una instrucción directa y clara.
- Se explicitan los criterios que luego son considerados en la rúbrica:
  - Adecuación a la situación comunicativa.
  - Estructura del texto (cuento).
  - Manejo de la lengua (ortografía literal y acentual).

Rúbrica

Dimensiones	Nivel 1	Nivel 2	Nivel 3
D.1 Adecuación a la situación comunicativa: Grado en que la evidencia se atiene a la tarea solicitada.	El texto **no** trata principalmente sobre el personaje de la imagen.	El texto trata principalmente sobre el personaje de la imagen (fantasma), **pero no** sobre la característica que aporta el título.	El texto trata principalmente sobre el personaje de la imagen (fantasma) **y** la característica que aporta el título (miedoso).
D.2 Estructura del texto.	El texto corresponde a una narración que presenta una secuencia de hechos y contiene **solo una** de las tres partes de la estructura del cuento. **O bien**, no existe narración, por ejemplo se presenta una descripción.	El texto corresponde a una narración que presenta una secuencia de hechos y contiene **dos** de las tres partes de la estructura del cuento.	El texto corresponde a una narración que presenta una secuencia de hechos y contiene la estructura de un cuento (comienzo, desarrollo y desenlace).
D.3 Manejo de la lengua: Ortografía literal y acentual.	El texto presenta una ortografía literal y acentual con **recurrentes** errores. (Se observan **seis o más palabras escritas con errores** de ortografía literal o acentual).	El texto presenta una ortografía literal y acentual con **algunos** errores. (Se admiten **cuatro o cinco palabras escritas con errores** de ortografía literal o acentual).	El texto presenta una ortografía literal y acentual **adecuada** para el nivel. (Se admiten **a lo más tres palabras escritas con errores** de ortografía literal o acentual).

Ejercicio 1: Análisis de la construcción de una pregunta abierta[1]

A continuación, se presenta una pregunta abierta con su rúbrica de corrección, elaboradas para evaluar la aplicación de contenidos matemáticos en estudiantes de diez años de edad. Analícela con base en los siguientes criterios: pertinencia del contexto; claridad y precisión del enunciado; concordancia entre los criterios de evaluación incluidos en la rúbrica y el enunciado.

Indicador	Utilizar fracciones para representar diferentes magnitudes.

Andrea tiene un chocolate dividido en ocho cuadraditos iguales.

Le dio a Pablo dos cuadraditos. Pablo le respondió: "Gracias por darme la mitad de tu chocolate".

A) ¿Es correcta la afirmación de Pablo?

Sí ☐ No ☐

B) Explica:

C) Dibuja tu explicación:

[1] El ejercicio resuelto lo podrá revisar en el anexo 2 de este cuadernillo.

Nivel 1	Nivel 2	Nivel 3
No explicita que la afirmación es incorrecta (ya sea en el cuadro de "No" o en la explicación). No cumple con los requisitos mencionados en el nivel 2.	Explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de "No" o en la explicación). Y su explicación **no** alude a una correcta representación de las fracciones involucradas en el problema. Y representa correctamente la fracción de chocolate faltante, en relación con el total. **O bien**, explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de "No" o en la explicación). Y su explicación alude a una correcta representación de las fracciones involucradas en el problema. Y **no** representa correctamente la fracción de chocolate faltante, en relación con el total.	Explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de "No" o en la explicación). Y su explicación alude a una correcta representación de las fracciones involucradas en el problema. Y representa correctamente la fracción de chocolate faltante, en relación con el total.

IV. Proceso de revisión de ítems por equipos y expertos

En esta sección se describe, por una parte, un posible modelo de trabajo para la revisión de ítems y, por otra, los criterios que se deben considerar en este proceso. El modelo presentado considera diferentes instancias de revisión que son aplicadas luego de que los ítems son construidos individualmente por especialistas en el constructo que se espera medir, quienes son capacitados en los criterios de elaboración de ítems descritos en las secciones anteriores.

Instancias de revisión de los ítems

Con el objetivo de asegurar la calidad de los ítems, una vez elaborados y entregados por los especialistas, son sometidos a un riguroso proceso de corrección de contenido (precisión conceptual), estructura lógica de su construcción y formal (sintáctica y ortográfica). Este proceso comienza con la recepción y análisis de los ítems por parte de los

equipos desarrolladores de prueba y continúa con la revisión de jueces expertos externos. Cabe señalar que las instancias de revisión son válidas para todo tipo de ítems (cerrados o abiertos). A continuación, en la figura 3 se detalla el flujo de revisión contemplado bajo este modelo.

FIGURA 3

FLUJO DE REVISIÓN DE ÍTEMS

Revisión interna del equipo desarrollador de la prueba: en una primera instancia los ítems son revisados por un equipo de profesionales, usualmente compuesto por especialistas del área que se está evaluando y profesionales dedicados a la evaluación y medición educacional. Ellos juzgarán si el ítem construido cumple con los criterios de construcción (descritos en el apartado anterior) y con la rigurosidad conceptual del conocimiento que se está evaluando. Con base en los criterios técnicos de revisión, se determina si el ítem cumple o no con los estándares de calidad esperados. Cuando lo hace, se toma la decisión de aprobar el ítem, mientras que, en caso contrario, se puede tomar la decisión de introducir modificaciones (por ejemplo, aclarar más el enunciado o cambiar algún distractor) o rechazarlo. A propósito de los ítems rechazados, cabe señalar que no existe un criterio único para ello. En general, un ítem es rechazado cuando tiene múltiples errores, por ejemplo, cuando supone cambios sustanciales que implicarían su reelaboración, presenta errores disciplinarios o conceptuales, o algún sesgo.

Revisión en comisión o sesión de análisis: en esta instancia se reúne el equipo interno con los elaboradores o constructores, para discutir y realizar en conjunto mejoras a los ítems a partir de modificaciones a su contexto, enunciado u opciones de respuesta. Para ello, el equipo interno previamente selecciona el grupo de ítems a discutir, planteando observaciones y propuestas de mejora frente a aquellos aspectos que requieren algún ajuste.

En la sesión de análisis, se pueden presentar los ítems uno a uno y guiar la conversación a partir de los criterios de construcción señalados antes, para trabajarlos y juzgar entre todos si el ítem es aprobado o rechazado, o bien, cuáles son las modificaciones que es necesario realizar. Es importante que todos los constructores contribuyan al ajuste de los ítems presentados.

Las reuniones se realizan de manera periódica durante el proceso de construcción. Una de sus ventajas es que constituyen una instancia de formación y retroalimentación para los elaboradores.

Revisión del experto en medición: este juez tiene la función de corroborar, nuevamente, si los ítems cumplen o no con los criterios de construcción esperados, aprobarlos o rechazarlos. El foco de revisión de este experto se refiere a los criterios de construcción del ítem, por ello se deben buscar profesionales con experiencia en la construcción y revisión de ítems.

Revisión del experto disciplinario: este juez revisor debe ser experto en los conocimientos y las habilidades evaluadas, ya que su función es asegurarse de que los ítems sean correctos, revisando el rigor conceptual a fin de detectar y corregir las posibles imprecisiones y rechazar aquellos que contengan errores. Como su foco de revisión se refiere a los criterios disciplinares del ítem, se recomienda buscar profesionales con experiencia y conocimiento especializado en la disciplina evaluada.

Revisión de estilo: en esta instancia se realiza una revisión formal, ortográfica y morfosintáctica del ítem. Para ello, se sugiere buscar un profesional del área de lenguaje que revise los ítems en cuanto al uso de ortografía literal, acentual, puntual y a su redacción, buscando la simplicidad y precisión de las oraciones y expresiones empleadas.

Revisión de diseño: una vez que los ítems han sido aprobados en cada una de las etapas mencionadas, se realizan los ajustes de imágenes y elementos gráficos que sean necesarios para asegurar que sean correctas y no presenten imprecisiones.

Criterios para el análisis de ítems

En las diversas instancias planteadas anteriormente, es fundamental que los ítems construidos sean revisados a partir de criterios específicos que aseguren su correcta construcción y rectitud disciplinar para así velar por la evidencia de validez de contenido (Downing y Haladyna, 1997). A continuación, se presentan paso a paso los criterios que guiarán este análisis.

a) Coherencia entre el ítem y el indicador de evaluación: verificar si el ítem evalúa claramente el indicador declarado y consignado en la tabla de especificaciones. Para ello, se sugiere leer el ítem completo y responder a la pregunta: ¿de qué habilidad y conocimiento da cuenta este ítem correctamente respondido? Si la respuesta coincide con lo declarado en el indicador, quiere decir que hay correspondencia entre ambos.

b) Claridad del enunciado: verificar si el enunciado del ítem es lo suficientemente claro y evidente como para asegurar una lectura lo más unívoca posible y si contiene el criterio de corrección de la pregunta. Con tal fin, se sugiere intentar responderlo sin mirar las opciones. El ítem debe tener una respuesta posible sin que haya necesidad de completar el sentido con las opciones de respuesta. Si no fuese así, cabe sospechar que el enunciado no es lo suficientemente claro, o bien, que no está expresado el criterio de corrección en su enunciado.

c) Contexto necesario y sin elementos que distraigan o perturben: verificar si el contexto es necesario, es decir, si entrega información relevante sin la cual no se podría responder el ítem. Si este puede ser respondido omitiendo el contexto, se sugiere eliminarlo. También es preciso constatar que sea apropiado, que su contenido no sea controversial o pueda ofender, perturbar o distraer al evaluado afectando su desempeño. Finalmente, es importante revisar que contenga los elementos suficientes para responder correctamente y que su extensión sea adecuada en relación con la tarea solicitada.

d) Opciones de respuesta concordantes con el enunciado: verificar que las opciones tengan concordancia lógica gramatical con el enunciado, es decir, si responden a este o a otra interrogante; si es así, distinguir qué es lo que realmente se quiere preguntar, contrastando con la opción correcta para determinar si el ajuste se debe hacer en el enunciado o solo en los distractores. Este aspecto no se revisa en preguntas abiertas.

e) Opciones de respuesta bien formuladas: verificar si la opción correcta es claramente la correcta y las restantes son incorrectas, aunque se comportan como distractores plausibles. Para ello, en primer lugar, es relevante constatar que la respuesta presumiblemente correcta se encuentra entre las opciones. Si se considera correcta una opción distinta a la que se presumía, se debe examinar si el error es del enunciado o de la opción correcta. La dificultad en el enunciado puede ser consecuencia de que la pregunta no es lo suficientemente precisa, o de que la opción correcta está planteada con imprecisión o vaguedad y que, por lo tanto, no puede ser identificada como tal. En segundo lugar, se recomienda verificar que los distractores sean **plausibles o atendibles,** es decir, que correspondan a errores habituales frente al enunciado. Si hay distractores poco plausibles o altamente descartables, las probabilidades de que alguien sin conocimiento suficiente conteste correctamente, aumentan. Otro aspecto que se debe constatar es que **los distractores sean incorrectos** desde el punto de vista conceptual. Este aspecto no se revisa en preguntas abiertas.

A partir de los criterios expuestos se analizará un ejemplo de ítem.

Ejemplo de análisis de ítem de opción múltiple	
Indicador de evaluación	Inferir información en un texto no literario.

Los smartphone han entrado en el mundo de los adolescentes por varias de sus características. Pero, sin lugar a dudas, la que más ha impactado es la posibilidad de tomar fotografías y revisarlas instantáneamente, olvidando el antiguo proceso de revelado.

¿Qué ha significado para los jóvenes la cámara en celulares?

a. Permite a los jóvenes guardar momentos de su vida de manera instantánea.

b. Ha producido un cambio en las conductas de los jóvenes por diferentes factores, de los cuales la fotografía digital es solo un factor.

c. No ha revertido un mayor cambio, en las conductas.

d. Ha forzado a los jóvenes por el cambio a celulares con mejores cámaras, aun cuando su uso está subutilizado como recurso social.

Criterio	Análisis
1. Coherencia entre la pregunta y el indicador de evaluación	El ítem no es concordante con el indicador de evaluación, ya que mide extracción de información explícita, no implícita.
2. Claridad del enunciado	Enunciado poco claro, emite un juicio que no está presente en el texto que se utiliza como referencia.
3. Contexto necesario y sin elementos que distraigan o perturben	El contexto es necesario para responder.
4. Opciones de respuesta concordantes con el enunciado	Las opciones no responden en su totalidad gramatical y lógicamente al enunciado.
5. Formulación de las opciones de respuesta	Las opciones, además, son dispares entre sí en términos de su extensión. La respuesta correcta es correcta, pero se obtiene por descarte, porque el enunciado es poco claro y los distractores no son plausibles.

Ejercicio 2: Análisis crítico de ítems cerrados

A continuación, se muestra un ítem de opción múltiple, el cual se sugiere analizar con base en los criterios presentados en la sección anterior. Puede revisar el ejercicio resuelto en el anexo 3 de este cuadernillo.

Indicador de evaluación	Identificar las características de una práctica pedagógica.

La educadora reflexiva tiene la capacidad de analizar su práctica y reconocer posibles problemas, para ello:

a. Implementa acciones de mejora, para mejorar.

b. Nombra los problemas y los trabaja con el equipo.

c. Explica los problemas e invita al equipo a solucionarlos.

d. Identifica los problemas e invita al equipo a definir soluciones.

V. Procesos de pilotaje de pruebas: cuantitativo y cualitativo

Cuando se elabora un instrumento de evaluación, un requisito fundamental es poner a prueba las preguntas antes de aplicar su versión definitiva. Para ello, es importante llevar a cabo una aplicación piloto, cuyo fin es verificar el funcionamiento y calidad de los ítems. Esta indagación puede llevarse a cabo bajo un enfoque cuantitativo o cualitativo. A continuación, se revisará cada una de dichas metodologías.

Estudio cuantitativo: análisis psicométrico

Tiene como finalidad poner a prueba el adecuado funcionamiento de los ítems de un instrumento por medio del análisis de su comportamiento métrico. Una aplicación con esta finalidad deberá contar con una amplia muestra de personas con características similares a la población que será evaluada finalmente con ese instrumento, ya que solo con muestras de esta magnitud es posible obtener resultados cuyos puntajes sean interpretables de manera válida. Proporciona información estadística sobre propiedades específicas de los ítems, como el grado de dificultad, la capacidad discriminativa, la tasa y patrón de omisión y la distribución de respuestas entre los distractores de la pregunta. Analizaremos cada una de estas propiedades a continuación.

- **Grado de dificultad de un ítem:** corresponde a la proporción o porcentaje de sujetos que logró responderlo correctamente respecto del total de personas que rindió una prueba. Se entiende que mientras menor es el porcentaje de respuestas correctas obtenido en la aplicación, más difícil resultó esa pregunta para los evaluados, y viceversa.

¿Qué rango de dificultad es adecuado?

El rango de dificultad requerido va a depender de los propósitos del instrumento, sin embargo, en términos generales, se asume que preguntas demasiado fáciles (cercana a 100%) o demasiado difíciles (bajo 10%), no debiesen incluirse en una prueba.

- **Capacidad discriminativa:** refiere a la efectividad con la que la pregunta diferencia entre personas que poseen diferentes grados de habilidad o conocimiento. Es claro que un ítem funciona de manera adecuada si quienes responden correctamente una pregunta tienen, en promedio, un puntaje total mayor que aquellos que contestaron de forma errónea. Un estimador usado para evaluar la capacidad discriminativa de los ítems es la correlación biserial entre el puntaje de la pregunta y el de la prueba. Se espera que en la medida en que exista una correlación alta y positiva entre la proporción de respuestas correctas en la pregunta y el puntaje obtenido en la prueba completa, el ítem aporte información consistente con lo que esta mide.

¿Cuándo se considera que un ítem posee una discriminación adecuada?

La siguiente tabla muestra los rangos de aceptación de un ítem en cuanto a la correlación biserial (Ebel, 1965):

Discriminación	Interpretación
Mayor o igual a 0,4	El ítem discrimina muy bien.
Entre 0,30 y 0,39	El ítem discrimina bien.
Entre 0,20 y 0,29	El ítem tiene una discriminación suficiente.
Menor a 0,20	El ítem no discrimina.

Preguntas con correlaciones negativas o cercanas a cero se consideran ítems paradójicos, ya que indican que la pregunta fue contestada correctamente, en su mayoría, por examinados que tuvieron un bajo desempeño global en la prueba. Por lo anterior, estos ítems deben omitirse de una prueba definitiva.

- **Tasa de omisión:** indica el porcentaje de los examinados que no respondieron una pregunta. Para interpretar este índice, es preciso considerar las instrucciones que se

entregaron al momento de rendir la prueba, pues, por ejemplo, es esperable que la tasa de omisión de un instrumento aumente si se indicó a los examinados que se descontaría cierto puntaje por la cantidad de preguntas erradas. Es recomendable hacer una comparación entre el patrón de omisión de los grupos de alta y baja habilidad: de este modo, es esperable que la omisión sea mayor en el grupo de menor habilidad, ya que ello evidenciaría el grado de dificultad del ítem. Si, la omisión es equivalente entre ambos grupos o es mayor en el grupo de mayor habilidad, se puede sospechar del funcionamiento del ítem y convendría descartarlo. Es importante, por último, analizar la tasa de omisión considerando el lugar que el ítem ocupa en la prueba, ya que podría entregar información acerca de una longitud excesiva del instrumento, si es que el patrón de omisiones se incrementa hacia el final.

- **Distribución de respuestas entre los distractores:** el porcentaje de elección de los distractores aporta información relevante sobre el funcionamiento del ítem, en tanto permite distinguir si son plausibles, es decir, si atraen a los examinados de menor habilidad. Por ejemplo, si un distractor posee una frecuencia de selección muy alta, tanto por examinados de mejor como de peor desempeño, es importante revisar esa pregunta para evaluar por qué está atrayendo a los buenos evaluados. Si un distractor logra una frecuencia de respuesta alta, y es marcada especialmente por los examinados de mejor desempeño, se hace necesario analizar si esta opción podría juzgarse también correcta e incluso mejor que la asignada como clave, si la opción acertada está mal codificada o si se trata de un error instalado en la población evaluada. Un distractor que mantiene una baja tasa de respuesta, debe cambiarse por otro. Conviene tener en consideración, eso sí, que cuando se trata de un ítem fácil, es esperable que la frecuencia de respuesta de los distractores sea baja.

Estudios cualitativos: basados en los procesos de respuesta

Cuando no es posible obtener una muestra suficiente que permita llevar a cabo un estudio piloto como el descrito, o bien, cuando el foco de interés está en levantar información sobre el proceso cognitivo que se lleva a cabo frente a una determinada tarea, existen como métodos de verificación de la calidad de los ítems los estudios cualitativos basados en los procesos de respuesta. Se llevan a cabo con un número reducido de sujetos con características similares a las de quienes serán evaluados y proporcionan información a partir de las verbalizaciones que realizan acerca de la tarea solicitada, de manera que es posible distinguir el ajuste existente entre el constructo que se quiere medir y la tarea que en efecto realizan los evaluados (Padilla y Benítez, 2014).

Según el momento en que son solicitados, los reportes verbales pueden ser categorizados como reportes concurrentes o retrospectivos. Dentro de la primera categoría se encuentra la técnica de "pensamiento en voz alta" o *think alloud*, en la cual se pide al evaluado que, en el transcurso de elaboración de su respuesta, vaya verbalizando la información en la cual se concentra, evidenciando el proceso cognitivo que lleva a cabo cuando resuelve un determinado problema. En la segunda categoría está la técnica llamada "entrevista cognitiva", en la que el evaluado hace referencia al proceso de respuesta una vez finalizada la tarea, respondiendo un cuestionario específico que orienta el levantamiento de la información sobre el ítem o la tarea.

La información recogida a partir de este tipo de reporte será relevante en tanto permita levantar evidencia de validez a favor o en contra de lo esperado, de acuerdo con lo que el ítem pretende medir. Algunos ejemplos de hallazgos relevantes son:

- Evidencia de que es posible responder el ítem usando un razonamiento que difiere de lo que este pretende medir.
- Evidencia de que alguna palabra del enunciado o del estímulo confunde, o, por el contrario, ayuda a los evaluados a llegar a la respuesta correcta.
- En preguntas de selección múltiple, evidencia de que se eligen distractores por motivos distintos de los que orientaron su construcción.
- Evidencia de errores conceptuales subyacentes a una respuesta o línea argumentativa que no fueron considerados en la pauta de corrección o distractores.
- Evidencia de que el ítem se puede responder por descarte.

Para levantar esta información es imprescindible contar con alguna pauta de entrevista estructurada o semiestructurada, que permita indagar en aquellos aspectos que son relevantes para el propósito del estudio y posibilite responder a sus preguntas guía (ver ejemplos de pautas de este tipo en Howell, Phelps, Croft, Kirui y Gitomer, 2013; Young *et al*., 2014).

Conviene tener en cuenta que, si bien ambos métodos pueden entregar evidencias de validez a un instrumento determinado, no son equiparables en términos de la información que entregan y, por lo tanto, no son comparables, aunque sí complementarios. Ambos tipos de estudio permiten tomar importantes decisiones respecto a los ítems construidos, conduciendo a aprobarlos o rechazarlos, o a modificar sus elementos, ya sea en el enunciado o en alguna de las opciones de respuesta, en pro de asegurar que luego puedan ser parte de un instrumento que permita evaluar el constructo esperado.

VI. Ensamblaje de pruebas definitivas

El ensamblaje de una prueba corresponde al proceso a través del cual se ordenan los ítems que serán incluidos en un instrumento para su aplicación definitiva. Este ordenamiento debe responder a la tabla de especificaciones previamente definida para el instrumento, manteniendo una lógica temática coherente para quien lo responderá. Lo que guía este paso es la "ficha de ensamblaje", que corresponde a la planificación de los pesos porcentuales que se otorga a cada uno de los objetivos de evaluación, temas y subtemas de la tabla de especificaciones y, por ende, del número de ítems asignado a cada elemento, considerando el total de los definidos para la prueba. En la siguiente tabla se detalla el procedimiento paso a paso.

TABLA 1

PROCEDIMIENTO DE ENSAMBLAJE DEL INSTRUMENTO DEFINITIVO

PASOS	ESPECIFICACIONES
1. Ordenar los ítems aprobados según tema, subtema y saber específico.	Se recopilan los ítems que hayan sido aprobados luego de pasar por el proceso de validación.
2. Organizar los ítems al interior de cada tema y subtema.	Secuenciar de acuerdo con el grado de dificultad obtenido en prueba experimental.
3. Seleccionar las preguntas según la cantidad a incluir establecida en la ficha de ensamblaje.	Elegir ítems asegurando cumplir con el porcentaje de preguntas esperado según el nivel de dificultad y los criterios establecidos de acuerdo con su comportamiento psicométrico[2].
4. Imprimir la prueba ensamblada y revisarla.	Se sugiere llevar a cabo revisiones por diferentes personas para asegurar el adecuado tipeo y organización de las preguntas.

Una vez realizado el ensamblaje de la prueba es fundamental revisarlo de forma exhaustiva y dejar registro de cualquier variación que haya existido en la organización de los ítems, respecto de lo requerido por la ficha de ensamblaje.

[2] En ocasiones, las preguntas experimentadas válidas se categorizan según un mejor o peor comportamiento psicométrico: de ser así, se sugiere seleccionar los ítems jerárquicamente, de manera de priorizar la presencia de las mejores preguntas en el instrumento.

Preguntas orientadoras para realizar la revisión del ensamblaje:

- ¿Están los items ordenados por temas y al interior de estos, por subtemas?
- En cada grupo de ítems por subtemas, ¿se procura una lógica ascendente de lo más fácil a lo más difícil?
- Dentro de un mismo subtema, ¿hay variedad de preguntas desde el punto de vista temático, o demasiadas preguntas sobre un mismo aspecto?
- ¿Cubren las preguntas todos los objetivos de evaluación asociados con cada tema?
- ¿Es posible afirmar que ninguna pregunta ayuda total o parcialmente a responder otra de la misma prueba?
- En el caso de preguntas que empleen imágenes en los contextos, ¿estas son claras y nítidas o es necesario reenviar el ítem a diseño?

VII. Establecimiento de puntos de corte (*standard setting*)

Cuando el propósito de evaluar a una población es establecer su desempeño con base en categorías, se hace necesario llevar a cabo un procedimiento que dé la posibilidad de comparar el desempeño actual de los evaluados con un cierto estándar esperado, ya que disponer solo de puntajes como resultado de la medición no es suficiente para cumplir tal objetivo. Este procedimiento, denominado establecimiento de puntos de corte o *standard setting*, permite clasificar a la población de examinados según categorías de desempeño, a partir de su puntuación en la prueba rendida. Por lo tanto, implica que un grupo de personas emita un juicio experto para establecer cuánto es lo mínimo que una persona debiera responder para ubicarse en un cierto nivel de desempeño (Cizek y Bunch, 2007). La cantidad de niveles de desempeño que se quieran distinguir dictará la cantidad de puntos de corte que es necesario establecer. Así, por ejemplo, para identificar cuatro niveles de desempeño se requieren tres puntajes de corte. El juicio, que debe ser realizado por expertos, es regulado por un conjunto de procedimientos estandarizados que permiten objetivar el proceso. Existen varios métodos para establecer puntos de corte. Los más conocidos y utilizados son Bookmark y Angoff, que se describen brevemente a continuación. La elección de uno u otro método está sujeta a varias consideraciones: el propósito del test; la complejidad del conocimiento o habilidades evaluadas; el formato de la prueba; el número de categorías de desempeño que deben identificarse; y los recursos disponibles para implementar el proceso.

Método Bookmark: los jueces trabajan con un cuadernillo que contiene las preguntas del instrumento ordenadas según su grado de dificultad empírica, de acuerdo con los resultados de la aplicación, desde la más fácil a la más difícil.

Su trabajo consiste en revisar primero, individualmente, las preguntas así ordenadas, y en seleccionar la primera que represente un paso o salto cualitativo de un nivel de desempeño a otro. Esta decisión individual es luego discutida en pequeños grupos, para converger finalmente en una sesión plenaria donde se establece el punto de corte final para cada nivel de desempeño.

Teniendo en consideración la descripción de los niveles de desempeño definida para el constructo evaluado, la pregunta que deben responder los jueces para decidir cada punto de corte es: ¿es probable que un examinado mínimamente calificado en este nivel responda correctamente este ítem?

Método Angoff: los jueces trabajan con un cuadernillo que contiene el conjunto de preguntas que conforman la prueba, ordenadas de la misma forma en que fue rendida por los examinados.

Su trabajo consiste en establecer el nivel de desempeño mínimo que se necesita para responder correctamente cada pregunta del cuadernillo. En una primera etapa, los jueces lo realizan de forma individual (*Angoff rating*) y, luego, en una ronda grupal discuten aquellos ítems en los que hay mayor diferencia entre los juicios establecidos. Finalmente, vuelven a adjudicar un nivel de desempeño a todos los ítems en discusión. A partir de lo realizado, la suma de ítems asignados en cada nivel de desempeño entregará el puntaje mínimo para pertenecer a estos.

Considerando la descripción de los niveles de desempeño definida para el constructo evaluado, la pregunta que deben responder los jueces con el fin de determinar el nivel de desempeño de cada ítem es: ¿cuál es el nivel de desempeño mínimo requerido para responder correctamente esta pregunta?

Independientemente del método que se utilice, un elemento crucial para desarrollar un buen procedimiento de establecimiento de puntos de corte es la selección de los jueces y la efectividad del entrenamiento que reciben. Se hace fundamental, entonces, dedicar tiempo suficiente para que los jueces comprendan la tarea y, en este proceso, entregar indicaciones muy específicas sobre lo que se está esperando para cada nivel de desempeño.

Consideraciones finales: ideas fuerza

- Definir adecuadamente el objetivo de evaluación del instrumento y la población a la que está dirigido es fundamental para respaldar la validez de las interpretaciones de los puntajes para los usos previstos.
- Elaborar una buena tabla de especificaciones permitirá delimitar el constructo que se quiere evaluar y guiar la construcción del instrumento.
- Las diferentes fases del proceso de construcción deben cumplirse rigurosamente, ya que cada una es determinante para la siguiente.
- El tipo de preguntas incluidas en una prueba depende del propósito del instrumento y, por ende, del tipo de contenidos y habilidades que se quiere medir.
- Los ítems de selección múltiple son los más usados y permiten evaluar una amplia variedad de contenidos y habilidades. Son fáciles de aplicar, pero difíciles de construir.
- Las preguntas abiertas permiten profundizar en los conocimientos que se quieren medir, pero son más costosas y difíciles de corregir.
- Algunos criterios para la elaboración de buenas preguntas son que el contexto debe ser suficiente y necesario para responder la pregunta, y el enunciado debe expresar de forma clara y directa la tarea.
- En los ítems de selección múltiple la opción correcta tiene que ser clara y completamente correcta. Las opciones incorrectas deben ser plausibles y representar razonamientos equivocados o errores comunes de los evaluados con menor dominio del tema.
- Las preguntas abiertas deben construirse siempre con una rúbrica que entregue los criterios a partir de los cuales se juzgará el desempeño del evaluado. El enunciado y la rúbrica deben mantener completa coherencia.
- Los ítems deben ser sometidos a diversas revisiones para asegurar su calidad según los criterios establecidos. En ellas pueden participar elaboradores de ítems, equipo interno, expertos disciplinarios y expertos de medición.
- Los principales criterios a considerar en la revisión de un ítem son: coherencia entre el ítem y el indicador; claridad del enunciado; contexto necesario y sin elementos que distraigan o perturben; opciones bien formuladas y concordantes con el enunciado.

- El estudio cuantitativo pone a prueba el funcionamiento de los ítems por medio del análisis de su comportamiento métrico, considerando, por ejemplo, el grado de dificultad y la capacidad discriminativa de las preguntas.

- El estudio cualitativo verifica la calidad de los ítems a partir del análisis del proceso cognitivo que se lleva a cabo ante una determinada tarea, a partir de las verbalizaciones del evaluado.

Referencias

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, American Psychological Association, y National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

ANDERSON, L. W., y Krathwohl, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* Nueva York: Longman.

CIZEK, G. J., y Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and evaluating Performance Standards on Tests.* Thousand Oaks: Sage Publications.

DOWNING, S. M., y Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*(1), 61-82.

EBEL, R. L. (1965). *Measuring Educational Achievement.* Englewood: Prentice-Hall.

GIERL, M.J., Bulut, O., Guo, Q., y Zhang, X. (2017). Desarrollar, analizar y utilizar distractores para las pruebas de opción múltiple en educación: una revisión exhaustiva. *Revisión de la investigación educativa, 87*(6), 1082-1116.

HALADYNA, T. M., y Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*(1), 37-50.

HALADYNA, T. M., y Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 2*(1), 51-78.

HALADYNA, T. M., Downing, S. M., y Rodríguez, M. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

HOWELL, H., Phelps, G., Croft, A., Kirui, D. y Gitomer, D. (2013). *Cognitive Interviews as a Tool for Investigating the Validity of Content Knowledge for Teaching Assessments.* Princeton: Educational Testing Service. Recuperado de: http://dx.doi.org/10.1002/j.2333-8504.2013.tb02326.x

MORENO, R., Martínez, R.J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema, 16*(3), 490-497.

MORENO, R., Martínez, R.J., y Muñiz, J. (2015). Guidelines based on validity criteria for development of multiple choice items. *Psicothema*, 27(4), 388-394.

PADILLA, J., y Benítez, I. (2014). Validity evidence based on response proceses. *Psicothema, 26*(1), 136-144.

YOUNG, J., King, T., Cogan, M., Ginsburgh, M., Kotloff, L., Cabrera, J., y Cavalie, C. (2014). *Improving content assessment for english language learners: Studies of the linguistic modification of test items.* (Research report No. RR 14-23). Princeton: Educational Testing Service.

Anexo 1. Criterios para la elaboración de ítems de opción múltiple

Sobre el contenido del ítem	• ¿Tiene concordancia con el indicador de evaluación? • ¿Aborda un aspecto central de la disciplina? • ¿Es claro y conciso? ¿No presenta complejidades innecesarias del lenguaje?
Sobre el contexto y el enunciado	• ¿El contexto es necesario para medir el indicador y responder la pregunta? • ¿El enunciado expresa con claridad la tarea que se solicita al evaluado? • ¿El enunciado constituye una pregunta directa y está planteado en positivo?
Sobre las opciones de respuesta	• ¿Todas las opciones responden a la pregunta contenida en el enunciado? • ¿La opción correcta es indiscutiblemente la única correcta? • ¿Los distractores son respuestas incorrectas, pero plausibles?

Anexo 2. Resolución del ejercicio 1: Análisis de una pregunta abierta

Andrea tiene un chocolate dividido en ocho cuadraditos iguales.

Le dio a Pablo dos cuadraditos. Pablo le respondió: “Gracias por darme la mitad de tu chocolate”.

A) ¿Es correcta la afirmación de Pablo?

Sí ☐ No ☐

B) Explica:

C) Dibuja tu explicación:

Rúbrica

Nivel 1	Nivel 2	Nivel 3
No explicita que la afirmación es incorrecta (ya sea en el cuadro de “No” o en la explicación). No cumple con los requisitos mencionados en el nivel 2	Explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de “No” o en la explicación). Y su explicación **no** alude a una correcta representación de las fracciones involucradas en el problema. Y representa correctamente la fracción de chocolate faltante, en relación con el total. **O bien,** explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de “No” o en la explicación). Y su explicación alude a una correcta representación de las fracciones involucradas en el problema. Y **no** representa correctamente la fracción de chocolate faltante, en relación con el total.	Explicita en alguna parte que la afirmación es incorrecta (ya sea en el cuadro de “No” o en la explicación). Y su explicación alude a una correcta representación de las fracciones involucradas en el problema. Y representa correctamente la fracción de chocolate faltante, en relación con el total.

- **Pertinencia del contexto:** El contexto presentado en la pregunta es adecuado y suficiente. No hay información que no sirva para responder, ni información que falte.

- **Claridad y precisión del enunciado:** La parte A es clara y precisa, ya que corresponde a una pregunta directa acerca de si la afirmación realizada por Pablo, es correcta o no. Sin embargo, la parte B de la pregunta no explicita qué es exactamente lo que debe considerarse en la explicación para que sea correcta. Asimismo, en la parte C la instrucción "Dibuja tu explicación" no deja claro qué fracción se busca que el estudiante represente, sería más preciso señalar: "Dibuja qué parte del total del chocolate Andrea entregó a Pablo".

- **Concordancia entre los criterios de evaluación incluidos en la rúbrica y el enunciado:** Existe concordancia entre la parte A de la pregunta y el primer criterio descrito en el nivel 3, sin embargo, los demás criterios incluidos en el nivel esperado (nivel 3), no se encuentran explicitados en el enunciado, por lo que, si se pretende evaluarlos, deben expresarse claramente en las instrucciones.

Anexo 3. Resolución de ejercicio 2: Análisis crítico de ítems cerrados

Criterio	Análisis
1. Coherencia entre la pregunta y el indicador de evaluación.	La pregunta es coherente con el indicador de evaluación declarado, ya que en su conjunto pondera cuál es una de las características de una práctica pedagógica reflexiva.
2. Claridad del enunciado.	El enunciado no es claro, se podría señalar que se trata de un enunciado no dirigido, es decir, en el que no se presenta con claridad cuál es la tarea que se espera por parte del evaluado y que no contiene un criterio de solución, por lo que es posible que exista más de una respuesta correcta para el planteamiento de este ítem.
3. Contexto necesario y sin elementos que distraigan o perturben.	No aplica este criterio, ya que la pregunta no contiene un contexto para ser respondida.
4. Opciones de respuesta concordantes con el enunciado.	Si bien las opciones de respuesta están todas planteadas con una misma lógica gramatical y semántica, es decir, corresponden a acciones de una práctica pedagógica reflexiva, no es posible observar si son concordantes con el enunciado, porque el enunciado no indica con claridad la tarea que se espera del evaluado.
5. Formulación de las opciones de respuesta.	No es posible identificar una respuesta correcta, porque el enunciado no es claro sobre la tarea a realizar, lo que no permite observar cuál es la respuesta correcta y si los distractores son plausibles o posibles respuestas. Además, las opciones b, c y d contienen una misma idea, con palabras diferentes, por lo cual no son independientes unas de otras y habría más de una respuesta correcta.
Observación general.	Los problemas detectados en este ítem son estructurales y de tal magnitud que no bastaría con realizar cambios específicos para mejorarlo, por lo que debería ser rechazado en un proceso de revisión.

Instituto Nacional para la Evaluación de la Educación
México

Centro UC
Medición - MIDE