

7

Cuadernillo técnico
de evaluación educativa

Desarrollo de instrumentos de evaluación: tareas de desempeño y rúbricas

7

Cuadernillo técnico
de evaluación educativa

Desarrollo de instrumentos de evaluación: tareas de desempeño y rúbricas



Centro UC
Medición - MIDE



Textos de
divulgación

INEE

Instituto Nacional para la
Evaluación de la Educación
México

Desarrollo de instrumentos de evaluación: tareas de desempeño y rúbricas

© Centro de Medición MIDE UC

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

© Instituto Nacional para la Evaluación de la Educación INEE

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

Autoras

Paloma Fernández Vázquez, MIDE UC
Ángela Haquin Bravo, 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
¿Qué son las evaluaciones de desempeño?	4
¿Cómo se evalúan los desempeños?	4
¿Qué pasos se deben seguir para diseñar una evaluación de desempeño?	6
Rúbricas	12
¿Qué es una rúbrica?	13
¿En qué se diferencian las rúbricas de otras pautas de corrección?	14
¿Qué ventajas tiene utilizar rúbricas?	16
¿Qué tipos de rúbricas existen?	16
Pasos para construir una rúbrica	18
Consideraciones finales: ideas fuerza	26
Referencias	27

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: tareas de desempeño y rúbricas

Resumen

En este cuadernillo se profundiza en las evaluaciones de desempeño: qué son, cuál es su naturaleza, qué las diferencia del resto de las evaluaciones y, a través de diferentes ejemplos, se va desarrollando cada uno de los pasos que dan cuenta de su proceso de construcción. Además, se abordan las rúbricas como instrumento de calificación para describir un desempeño en distintos niveles de logro y las principales etapas que se deben llevar a cabo para elaborarlas. También se establecen las diferencias con otros instrumentos como las listas de cotejo y las escalas de apreciación, y se analizan las ventajas y/o desventajas de utilizar cada uno de ellos según diferentes propósitos. Finalmente, como parte del proceso de validación, se profundiza en las tareas asociadas al pilotaje de la evaluación de desempeño y de su respectiva rúbrica.

Introducción

Para comenzar y comprender de mejor manera la necesidad de utilizar evaluaciones de desempeño en ciertos contextos, analizaremos el siguiente ejemplo:

Ejemplo de evaluación de desempeño: Trabajo colaborativo entre profesionales

El equipo directivo y los y las docentes de una escuela han implementado el trabajo colaborativo entre profesionales como una metodología para alcanzar objetivos comunes y construir conocimiento en la escuela. Dialogar en torno a necesidades e intereses de tipo pedagógico permitirá fortalecer el desempeño de sus integrantes y, a largo plazo, impactar en la mejora de los aprendizajes de los y las estudiantes.

Ya ha pasado un año y quieren evaluar la implementación de esta metodología para poder hacer ajustes. Consideran que aún es muy pronto para evaluar el impacto a nivel de las prácticas de docentes y aprendizaje de estudiantes, pero sí creen que existen avances en la calidad del diálogo entre los participantes.

- ¿De qué forma podrían recoger evidencia útil para evaluar la calidad del diálogo?
- ¿Qué factores podrían influir al tomar la decisión?

Cuando lo que se busca evaluar corresponde a un desempeño, determinar cómo se recogerá la información es una decisión fundamental. En el ejemplo anterior, la comunidad educativa quiere evaluar el diálogo pedagógico entre profesionales. Para este propósito, una prueba podría llevar a interpretaciones erróneas frente a los resultados que se obtengan, ya que no se busca evaluar cuánto sabe cada participante respecto del trabajo colaborativo o del diálogo entre profesionales, sino con qué calidad se ejecuta o pone en práctica. En este caso, lo más adecuado sería evaluar a través de la observación, ya sea durante la ejecución del desempeño (por ejemplo, observando una o más sesiones de trabajo), o bien, a partir de algún producto de esta ejecución (por ejemplo, mediante un reporte del diálogo llevado a cabo en las sesiones).

I. ¿Qué son las evaluaciones de desempeño?

Son evaluaciones que se emplean cuando el **constructo** que se quiere evaluar implica un **saber hacer** que se evidencia a través de la ejecución de una tarea o actividad en particular, o bien, un **saber ser** que se traduce en el comportamiento de los sujetos en una determinada situación. En estos casos, lo que importa no es cuánto sabe un sujeto respecto de un área del conocimiento, sino cómo se enfrenta a una determinada situación o con qué calidad ejecuta o pone en práctica aquello que sabe. Se utilizan evaluaciones de desempeño cuando lo que se quiere evaluar refiere, por ejemplo, a:

Un **constructo** es una construcción teórica no observable directamente, por ejemplo, comprensión lectora, autoeficacia, colaboración. Los constructos (también denominados variables latentes) suelen contraponerse a las variables observables, las cuales, como dice su nombre, sí pueden observarse directamente, por ejemplo, peso corporal o estatura (Corral-Verdugo, 2001).

- Procedimientos de naturaleza motriz, con una acción corporal observable (por ejemplo, practicar un deporte o tocar un instrumento).
- Procedimientos más cognitivos que se utilizan al operar con símbolos o representaciones (por ejemplo, desarrollar una teoría, escribir una noticia, comprobar una hipótesis, etc.).
- Acciones o prácticas inherentes a una determinada labor, por ejemplo, el desempeño docente.
- Actitudes, hábitos o valores, es decir, aspectos más afectivos asociados al comportamiento de los sujetos en una determinada situación (por ejemplo, actuar de acuerdo con los valores y normas de convivencia cívica, respetar la diversidad cultural, étnica y religiosa, valorar el entorno natural, tener hábitos de cuidado del medio ambiente, entre otras).

II. ¿Cómo se evalúan los desempeños?

Dada su naturaleza, el medio más adecuado para la evaluación de desempeños es la **observación** del proceso, o bien, del producto de una determinada tarea o actividad. Es posible realizar esta observación en distintas situaciones:

Evaluaciones que refieren a desempeños espontáneos. Corresponden a la evaluación del desempeño en situaciones que no están previamente diseñadas, pero que, al ocurrir, permiten apreciar el desempeño de las personas y recoger información acerca de aquello que interesa. Al observar sistemáticamente actividades que forman parte de las rutinas o de la programación habitual de los evaluados, podemos levantar información útil.

Por ejemplo, la observación de celebraciones institucionales, reuniones de profesores y apoderados, pueden otorgar información valiosa sobre distintos constructos, como el clima laboral y la relación entre los distintos actores.

Evaluaciones que refieren a desempeños guiados. Corresponden a la evaluación del desempeño de un sujeto en situaciones especialmente diseñadas para ello. Las evaluaciones guiadas buscan apreciar el “mejor desempeño” posible de el/la evaluado/a, por ello se caracterizan por proveer al sujeto de instrucciones específicas y de un tiempo acotado para mostrar su desempeño.

Son evaluaciones que refieren a desempeños guiados: la realización de portafolios, formulación de proyectos, ensayos, disertaciones o debates, experiencias de laboratorio o taller, entre otras.

Ejemplo de evaluación de desempeño: Evaluación Docente en Chile

Para evaluar el cumplimiento de los estándares expresados en el *Marco para la Buena Enseñanza*¹, el Sistema de Evaluación del Desempeño Profesional Docente de Chile utiliza, entre otros instrumentos, un Portafolio. Este instrumento permite recoger evidencias concretas del quehacer de cada docente, posibilitando que dé cuenta de su desempeño en las tareas centrales de su profesión.

Para la presentación del Portafolio, cada docente recibe un manual que contiene instrucciones específicas dirigidas a elaborar los productos solicitados, entre ellos, la grabación de una clase. Los profesores tienen indicaciones sobre lo que se espera observar durante dicha clase, incluyendo orientaciones didácticas según la asignatura o materia en la que se desempeñan.



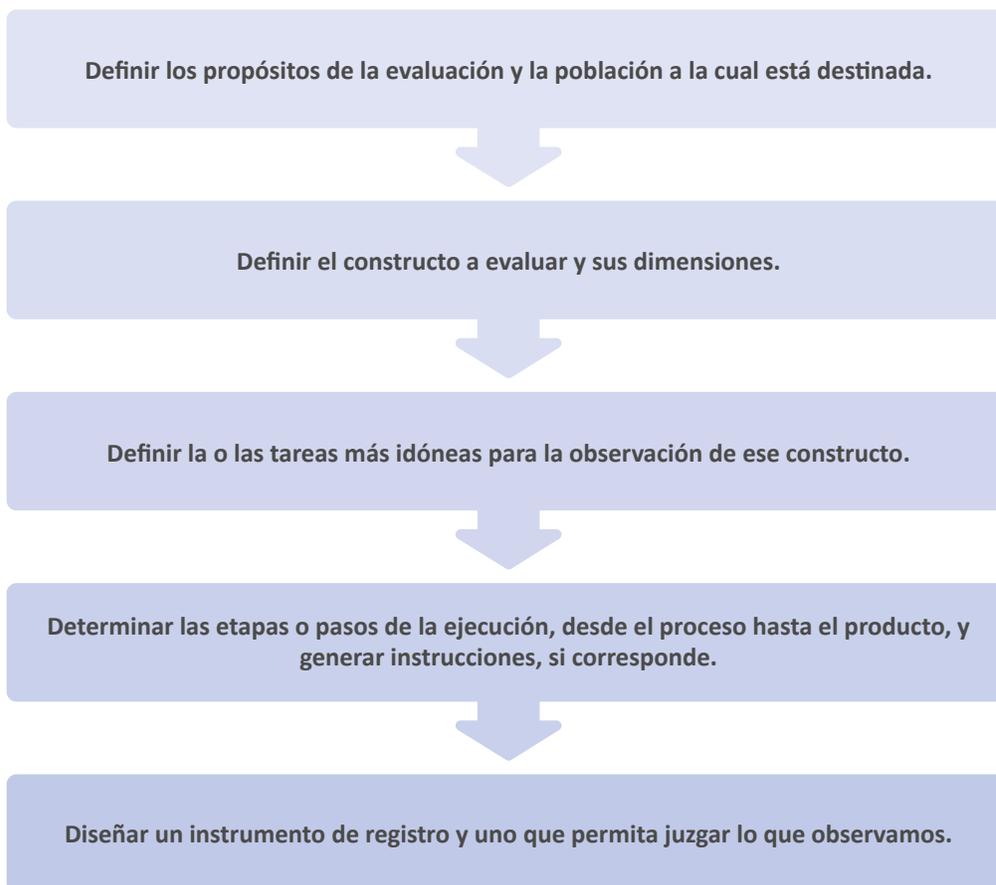
¹ El *Marco para la Buena Enseñanza* establece lo que las y los docentes chilenos deben conocer y saber hacer en el aula y en la escuela.

III. ¿Qué pasos se deben seguir para diseñar una evaluación de desempeño?

El diseño de una evaluación de desempeño es un proceso que comprende diferentes etapas o pasos. En el siguiente esquema se presentan aquellos que se han considerado centrales. Es importante tener en cuenta que, en la práctica, este proceso suele no ser lineal sino cíclico. Las decisiones tomadas en una etapa se vuelven a revisar en etapas posteriores, con el propósito de asegurar coherencia y hacer ajustes, en caso de ser necesario.

Durante este proceso, además, se debe ir acumulando evidencia que proporcione una base científica a la interpretación de los resultados que se obtengan. A esto se denomina **proceso de validación**, el cual es un proceso continuo, de construcción y evaluación de argumentos a favor y en contra de la interpretación que desea hacerse de los resultados y su relevancia para los usos que se han propuesto (tal como se explica en el cuadernillo 2 de esta serie).

FIGURA 1
PASOS DEL DISEÑO DE UNA EVALUACIÓN DE DESEMPEÑO



A continuación, se explica cada uno de los pasos para diseñar una tarea que permita evaluar un desempeño.

Paso 1: Definir los propósitos de la evaluación y la población a la cual está destinada

Como en toda evaluación, en primer lugar, se debe establecer claramente cómo se tiene previsto que se interpreten y en consecuencia se utilicen los resultados (American Educational Research Association [AERA], American Psychological Association [APA] y National Council on Measurement in Education [NCME], 2014).

Validez es el grado en que la evidencia y la teoría respaldan las interpretaciones de los resultados de una evaluación en función de los usos propuestos para ellos (AERA, APA y NCME, 2014).

Esta primera definición deberá guiar los siguientes pasos del diseño de la evaluación, en los que se debe resguardar que las **interpretaciones que se hagan de la información recogida sean válidas en función de los usos previstos para dicha información.**

También es importante definir claramente la **población** para la que está prevista la evaluación (AERA, APA y NCME, 2014), de modo que todos los pasos del proceso de diseño tengan en consideración las características de esta. Asimismo, se debe poner atención a la posible presencia de **subgrupos** relevantes que puedan verse perjudicados o favorecidos arbitrariamente por ciertas características del

Población hace referencia al conjunto de sujetos (u objetos) que comparten una característica determinada en común que los identifica (Fontes de Gracias *et al.*, 2010).

diseño de la tarea, lo que implicaría hacer ajustes o bien interpretar de modo distinto la información que se recoja a partir de la evaluación.

**Ejemplo de propósito de evaluación y población a la que se dirige:
Ensayo sobre violencia en los estadios**

Para evaluar la competencia de expresión escrita en sus estudiantes, un docente de secundaria plantea como tarea de finalización del año escolar la elaboración de un ensayo sobre la violencia en los estadios. Al analizar los resultados, el profesor observa que hay un grupo de alumnos que destaca por sobre el resto, y que coincide con quienes juegan habitualmente fútbol en la escuela.

En el ejemplo se puede observar cómo la tarea solicitada podría estar beneficiando al grupo de estudiantes que tiene mayor dominio del tema a partir del cual se debe realizar el ensayo. El conocimiento previo del tema, en este caso correspondería a una característica ajena e irrelevante respecto del constructo que el docente está evaluando (competencia de comunicación escrita).

Una tarea es de calidad cuando los resultados pueden ser interpretados solo en función de la competencia que se está evaluando y no de otros aspectos irrelevantes para ese constructo que puedan beneficiar o perjudicar a subgrupos particulares de la población.

**Ejemplo de propósito de evaluación y población a la que se dirige:
Comunicación oral en lengua Maya**

En una escuela se desea evaluar la competencia de comunicación oral en lengua Maya en un grupo de estudiantes que toma clases de esta lengua, para determinar las fortalezas y debilidades de la nueva metodología de enseñanza empleada desde hace un año. Los docentes han pensado en solicitar a los estudiantes exponer en lengua Maya sobre algún tema referido a esta cultura. Entre ellos, hay algunos/as estudiantes más inmersos en dicha cultura que otros y también con mayor exposición a su lengua.

En este ejemplo, posiblemente se encuentren diferencias en los resultados entre los estudiantes más expuestos a la cultura en cuestión y aquellos que no lo están. En contraste con el caso anterior, detectar tales diferencias no necesariamente representa una debilidad de la tarea diseñada, pues el grado de exposición a la lengua sí se relaciona con el constructo que está evaluando (no es una característica irrelevante) y, por lo tanto, hay que considerarlo al interpretar los resultados.

En conclusión, al reconocer la presencia de subgrupos relevantes al diseñar una tarea de desempeño, se deben atender dos aspectos: por un lado, resguardar que los subgrupos no se vean perjudicados o favorecidos arbitrariamente por ciertas características irrelevantes para el constructo. Por otro, si es que existen características relevantes asociadas al constructo que puedan incidir en los resultados, es preciso tenerlas en cuenta al momento de realizar la interpretación de estos.

Paso 2: Definir el constructo a evaluar y sus dimensiones

Otro punto central en el diseño de una evaluación de desempeño es determinar con claridad lo que se quiere evaluar. Considerando que los constructos son variables complejas, es fundamental definirlos y especificar claramente las dimensiones o aspectos que comprenden (AERA, APA y NCME, 2014). Cuando el constructo es muy complejo, puede

requerirse, además, de subdimensiones. La definición del constructo y sus dimensiones sirve de marco conceptual para la evaluación y debe contemplar, entre otras fuentes, la **revisión de literatura** y la **consulta a expertos**. Esta etapa se explica en profundidad en el cuadernillo 3 de esta serie.

Ejemplo de dimensiones: Evaluar el rol del tutor de docentes novatos

Para evaluar la eficacia del cargo de tutores *seniors* en un grupo de escuelas, se realizan distintas entrevistas a docentes novatos que han recibido tutorías. Luego de la revisión bibliográfica y de la realización de algunos grupos focales con expertos en el tema, se ha definido que el constructo “rol del tutor de docentes novatos” se entenderá del siguiente modo: “Un tutor es quien ofrece acompañamiento efectivo en la tarea pedagógica de docentes novatos, promoviendo su reflexión sobre la práctica y constituyéndose en un apoyo emocional y técnico en la búsqueda de respuestas ante los desafíos de su profesión”.

Desde esta definición del constructo, el equipo ha descrito las siguientes dimensiones:

- **Acompañamiento emocional del tutor:** Dimensión que comprende todas aquellas acciones orientadas a brindar un apoyo y ayuda emocional en el proceso de inserción a la práctica de los docentes novatos. Por ejemplo, contención y apoyo afectivo al enfrentar situaciones que desconoce o situaciones que le generen frustración o desmotivación, etcétera.
- **Acompañamiento pedagógico-técnico:** Dimensión que comprende todas aquellas acciones de retroalimentación orientadas a informar al docente novato de sus logros y áreas de desarrollo, entregar directrices u orientaciones sobre cómo seguir aprendiendo.
- **Promoción del autoaprendizaje:** Dimensión que comprende todas aquellas acciones de mediación que se orientan a promover la propia reflexión del docente novato y el análisis crítico de sus propias decisiones pedagógicas.

Como una forma de juzgar la calidad y pertinencia de las dimensiones que se describen con relación al constructo, siempre es importante responder las siguientes preguntas:

- ¿Todas las dimensiones son concordantes con el constructo que se definió?
- ¿Las dimensiones representan lo nuclear del constructo?
- ¿Se agregan aspectos que no fueron considerados en la definición del constructo?

En el ejemplo, las tres dimensiones descritas son coherentes con el constructo: “rol del tutor de docentes novatos”, ya que consideran los tres aspectos centrales de la definición construida a partir de la revisión bibliográfica y la consulta a expertos en el tema.

Paso 3: Definir la o las tareas más idóneas para la observación de ese constructo

Una vez que se ha determinado con claridad y suficiencia en qué consiste el desempeño a evaluar, es necesario determinar qué tareas permitirán obtener evidencia de ese constructo. Al hacerlo, es importante tener en cuenta las definiciones que se han establecido previamente respecto de los propósitos de la evaluación, los sujetos evaluados y las dimensiones que comprende el constructo a evaluar. En cuanto a esto último, la

tarea que proponamos debe estar acorde con la extensión y alcance del constructo, evitando dos aspectos (Koretz, 2008):

- **Infrarrepresentación del constructo:** Es decir, el grado en el cual la evaluación no logra capturar aspectos importantes del constructo que se busca medir, lo que se traduce en resultados que no lo representan totalmente.

Por ejemplo, si lo que se busca evaluar y, en consecuencia, reportar, es la calidad de la docencia en una escuela, el tener en cuenta solo las planificaciones de los maestros y maestras podría considerarse una infrarrepresentación del constructo, pues la evaluación no lograría capturar otros aspectos importantes que definen la calidad del ejercicio docente.

- **Irrelevancia de constructo:** Refiere al grado en el cual los resultados se ven afectados por procesos externos al fin previsto de la evaluación.

Por ejemplo, si lo que se busca evaluar es la capacidad de alumnos de primaria para resolver problemas matemáticos, el emplear problemas de gran extensión, con vocabulario que los y las estudiantes aún no dominan completamente, podría derivar en irrelevancia de constructo, en tanto los resultados se verían afectados a causa de su comprensión lectora y, por lo tanto, no reflejarían su capacidad para resolver problemas matemáticos.

Otro punto importante a considerar es la **factibilidad** de llevar a cabo las acciones que implican la realización de la tarea. Esto supone poner atención en las condiciones prácticas, situacionales o institucionales. Por ejemplo, cuando un distrito decide evaluar el desempeño de los y las docentes a través de la observación de aula, una decisión relevante es determinar las veces que será observado cada docente. Ello no solo dependerá de fundamentos técnicos respecto de cuántas observaciones es más aconsejable realizar, sino también de aspectos más prácticos, como puede ser la cantidad de observadores capacitados de los que se dispondrá y el plazo con el que se cuenta para realizar todas las observaciones.

Paso 4: Determinar las etapas o pasos de la ejecución, desde el proceso hasta el producto, y generar instrucciones, si corresponde

Cuando se diseña una tarea para evaluar el desempeño de uno o más sujetos, se debe procurar que las instrucciones que se les entreguen posibiliten a todos ellos dar cuenta de tal desempeño. A su vez, es preciso procurar que la formulación de las instrucciones no se convierta en una fuente de error, es decir, conviene evitar que la variabilidad en los

resultados sea producto de variables externas al desempeño del sujeto, como pueden ser la ausencia de una determinada instrucción o protocolo, o la presencia de instrucciones imprecisas, confusas, ambiguas o incompletas. Las instrucciones de una tarea deben ser lo suficientemente precisas y completas para que los resultados sean interpretados adecuadamente en función del constructo que se busca evaluar. Asimismo, en caso de que se requieran protocolos para aplicar la evaluación, deben ser claros de modo que se aseguren las mismas condiciones para todos los evaluados.

Ejemplos de tareas para evaluar desempeños: Instrucciones y protocolos

En cada uno de los siguientes casos, ¿qué mejoras se podrían realizar en el diseño de la tarea, instrucción o protocolo de aplicación para que no constituya una fuente de error?

- Un distrito quiere evaluar habilidades pedagógicas generales en los docentes de tres de las escuelas que tiene a cargo. Con tal fin, contrata a dos profesionales, quienes son capacitados para observar clases y aplicar una rúbrica con los distintos aspectos a evaluar. Luego de una primera visita a las escuelas, uno de los profesionales comenta que solo pudo observar veinte minutos de la clase de uno de los docentes, mientras que de las clases de los otros logró observar cuarenta minutos.
- Un docente entregó la siguiente instrucción a sus estudiantes de primer año de secundaria para evaluar la redacción de textos argumentativos: “Escriban un texto en el que entreguen su opinión sobre la utilización de los teléfonos celulares en las clases”. Mientras realizaban la tarea, varios alumnos le preguntaron: ¿De cuántas páginas? ¿Cuántos argumentos debo entregar? ¿Puedo poner un argumento a favor y dos en contra?

En estos casos, la forma en que se aplica el protocolo o se entregan las instrucciones podría constituirse en una fuente de error. En el primer ejemplo, las diferencias en la duración de la observación de clases podrían afectar los resultados que se obtengan. En el segundo ejemplo, la ausencia de instrucciones centrales para desarrollar la tarea podría implicar que los resultados no permitan evaluar adecuadamente la competencia de redacción de textos argumentativos, dadas las dificultades de los estudiantes para definir cómo realizar la tarea.

Por lo tanto, en el primer caso es fundamental incluir en el protocolo el tiempo esperado de observación y, en el segundo, completar la instrucción haciendo referencia al tipo de texto que se espera y a los requisitos mínimos para que pueda ser evaluada su competencia argumentativa.

Tenga en cuenta que si bien se sugiere estandarizar las condiciones para aplicar una evaluación, hay casos en los que es necesario flexibilizar estas condiciones con el fin de proporcionar oportunidades equivalentes a los evaluados que tienen características diferentes. Por ejemplo, alguna discapacidad o dificultades en el manejo de la lengua

pueden demandar flexibilidad en las condiciones de aplicación (como dar más tiempo para realizar la tarea).

Paso 5: Diseñar un instrumento de registro y uno que permita juzgar lo que observamos

Las tareas de desempeño se evalúan mediante la observación, es decir, existe "otro" que debe observar al sujeto durante la ejecución de la tarea, o bien, observar el producto de esta (ya sea escrito, gráfico, concreto, etc.). Es entonces la observación de este sujeto o juez la que nos lleva a obtener el resultado de la evaluación.

Un **sesgo** corresponde a cualquier influencia consciente o inconsciente (creencias, preferencias, experiencias, etc.) que puede afectar el modo en que se corrige un desempeño.

Cuando se observa para evaluar un desempeño en función de la definición previa de un constructo y sus dimensiones, la observación que se realice debe ser estructurada. Los resultados que de ella se obtengan deben estar asociados con la competencia, habilidad, destreza, etcétera, que se busca medir, y no a otras variables, como podrían ser los **sesgos** del evaluador o juez.

Para minimizar el error causado por la presencia de un juez, es importante contar con instrumentos que permitan registrar lo que observamos (en el caso de que la observación se realice durante la ejecución misma), e instrumentos que nos faciliten juzgar lo que observamos y determinar el grado en que están presentes los aspectos a medir. Existen distintos tipos de instrumentos, como listas de cotejo, escalas de apreciación y rúbricas.

IV. Rúbricas

A continuación, se aborda la rúbrica como un instrumento de evaluación que describe un desempeño en distintos niveles de logro. A través de diferentes ejemplos, se desarrollan los pasos que se deben llevar a cabo para su construcción y se reflexiona sobre las ventajas de usar este tipo de instrumento por sobre una lista de cotejo o escala de apreciación.

Además, como parte del proceso de validación, se profundiza en las tareas asociadas al estudio piloto de una evaluación de desempeño y de su respectiva rúbrica.

**Ejemplo de utilización de rúbrica:
Evaluación de la comunicación escrita en estudiantes universitarios**

Como una manera de asegurar el dominio de habilidades básicas asociadas con la producción escrita, una prestigiosa universidad decidió aplicar de forma obligatoria un examen de comunicación escrita a todos los estudiantes que cursan su primer año. El Departamento de Lingüística de la universidad, a cargo del examen, ha optado por solicitar la escritura de un ensayo. Durante el mes de marzo aplicarán esta evaluación, de tal manera que se tenga al menos un mes para la corrección y análisis de los resultados. La idea es tener estos en el mes de mayo, de modo que los estudiantes que no obtengan un nivel de logro aceptable puedan integrarse a los talleres de reforzamiento de comunicación escrita que se iniciarán ese mismo mes.

Por ahora, el equipo a cargo ha pensado que quienes corregirán el examen serán egresados de la carrera de Lengua y Literatura, sin embargo, aún están discutiendo cuál es el instrumento de corrección más idóneo para cubrir sus necesidades.

¿Qué instrumento podrían utilizar para evaluar el ensayo?

Cuando lo que se quiere evaluar corresponde a un desempeño, el empleo de pautas de corrección como rúbricas, listas de cotejo o escalas de apreciación ayuda a disminuir el error de medición causado por la presencia de un juez, al clarificar los criterios a usar al corregir. En este caso, para optar por uno u otro es conveniente considerar, entre otros aspectos, cuál de ellos aporta mayor objetividad, puede ser aplicado en un mes y permite retroalimentar con mayor información a quienes se encarguen de diseñar los cursos de reforzamiento. En este contexto, la rúbrica permite distinguir diferentes niveles de logro y describir con detalle en qué consiste cada uno de ellos, lo que ayudaría a que los jueces que actúen como correctores de los ensayos no apliquen sus propios criterios, sino que se guíen por lo especificado en cada nivel. Esta misma propiedad de las rúbricas favorece también que quienes diseñen el programa de reforzamiento tengan información precisa sobre el nivel de desarrollo de cada dimensión asociada con la comunicación escrita.

V. ¿Qué es una rúbrica?

Una rúbrica es un instrumento de calificación que describe en qué consiste un desempeño en distintos niveles de logro. A diferencia de otras pautas, la rúbrica especifica claramente cada uno de los niveles y, de este modo, entrega orientaciones sobre cómo debe apreciar el desempeño un juez.

La rúbrica es una herramienta que permite juzgar el desempeño o el fenómeno estudiado con base en las dimensiones definidas como relevantes en una situación

evaluativa. Es decir, permite guiar el proceso de apreciación y fundamentar el juicio evaluativo. Para hacerlo, se compone de los siguientes elementos:

1. Una o más **dimensiones** que sirven de base para juzgar el desempeño frente a la tarea y que se relacionan directamente con el constructo que se desea evaluar.
2. Una **escala de valoración** con la que se evalúa cada dimensión (cuantitativa o cualitativa).
3. Una **descripción para cada nivel** de la escala (nivel de desempeño).

Ejemplo de elementos de una rúbrica: Entrevista a docentes				
<p>Como parte del Sistema de Evaluación Docente en Chile se utiliza una entrevista a los docentes del sector municipal que consiste en una instancia de conversación profesional entre pares. A continuación, se presenta una de las preguntas de este instrumento y la rúbrica con la que fue evaluada el año 2014 (Ministerio de Educación, 2014):</p> <p>Pregunta de la entrevista: ¿Qué hace usted cuando los alumnos no cumplen alguna norma de convivencia?</p>				
Niveles de logro				
	Insatisfactorio	Básico	Competente	Destacado
Dimensión: Estrategias para abordar las normas de convivencia	No cumple la condición del nivel “Básico” o solo deriva la situación de incumplimiento a otra persona, sin hacerse cargo de ella en algún momento. O, en su respuesta, no alude a la práctica.	Aborda el quiebre de las normas, pero no queda claro que lo haga con un sentido formativo.	Aborda el quiebre de las normas con sentido formativo, otorgándole un sentido o “razón de ser” que va más allá de las consecuencias directas de su transgresión.	Cumple con la condición del “Competente” y además genera estrategias para acordar con el o los alumnos las consecuencias de haber transgredido las normas.

VI. ¿En qué se diferencian las rúbricas de otras pautas de corrección?

Las rúbricas, al igual que otros instrumentos como las listas de cotejo y las escalas de apreciación, también sirven para juzgar un desempeño, sin embargo, es importante no confundirlos. Veamos los ejemplos de una lista de cotejo y de una escala de apreciación.

**Ejemplo de utilización de lista de cotejo y de escala de apreciación:
Evaluación de habilidades cognitivas en el aula**

El equipo directivo de una escuela quiere observar a sus docentes para diagnosticar la promoción de habilidades cognitivas superiores en el aula. ¿Cuál de los siguientes instrumentos escogería para el propósito que se plantea?

Lista de cotejo

Indicadores	Sí	No
Plantea tareas que apuntan al desarrollo de habilidades de elaboración de información.		
Plantea tareas que apuntan al trabajo autónomo de los estudiantes.		
Promueve que los estudiantes analicen sus propios errores.		

Las **listas de cotejo** se caracterizan por presentar un listado de indicadores cuya presencia o ausencia interesa determinar. En consecuencia, la respuesta es dicotómica. Las listas de cotejo suelen utilizarse para sistematizar observaciones y pueden servir como herramientas de apoyo en la aplicación de una rúbrica, en tanto permiten enlistar los distintos elementos que se deben observar, facilitando su registro en forma ordenada y metódica.

Escala de apreciación

Indicadores	Logrado	Mediana-mente logrado	No logrado
Plantea tareas que apuntan al desarrollo de habilidades de elaboración de información.			
Plantea tareas que apuntan al trabajo autónomo de los estudiantes.			
Promueve que los estudiantes analicen sus propios errores.			

Las **escalas de apreciación** también presentan un listado de indicadores que son juzgados de acuerdo con una escala que permite identificar el grado en el cual se ha representado la cualidad o característica. A diferencia de las rúbricas, no describen los niveles de desempeño correspondientes a esa graduación.

El uso de cualquiera de estos instrumentos aporta a disminuir el error de medición por la presencia de un juez que evalúa el desempeño de los docentes. Pensando en el propósito del equipo directivo, que quiere diagnosticar a sus docentes en el trabajo de estas habilidades, la escala de apreciación presenta más ventajas, ya que aporta más distinciones en los niveles de desempeño que la lista de cotejo. Sin embargo, si el equipo

quisiera también determinar con mayor precisión las diferencias en el desempeño de los docentes para cada indicador en cuestión, o bien, si las observaciones son realizadas por distintos actores, sería importante aclarar, al menos, cuál es la diferencia entre los niveles “Logrado”, “Medianamente logrado” y “No logrado” en cada uno de los indicadores. Si esto no se hace, el director, el jefe técnico pedagógico y el coordinador de ciclo podrían aplicar la escala con diferentes grados de exigencia. Para que ello no ocurra, en esta situación sería conveniente emplear una rúbrica.

VII. ¿Qué ventajas tiene utilizar rúbricas?

- Aumentan la objetividad de la evaluación, pues cada nivel de desempeño está claramente definido. Esto contribuye a disminuir la probabilidad de errores de medición a causa de la aplicación de sesgos, garantizando mayor consistencia entre evaluadores.
- Son instrumentos de calificación muy versátiles, pues se utilizan en diferentes contextos, desde la educación inicial hasta la universitaria, o bien en tareas de muy diverso tipo (desde evaluar procedimientos cognitivos hasta actitudes, hábitos o valores).
- El solo hecho de construirlas implica un ejercicio reflexivo y consensuado para la selección y definición de las dimensiones que serán observadas, y para la descripción de cada nivel de logro.
- Cuando son utilizadas con el fin de retroalimentar al evaluado, aportan al sentido formativo de la evaluación, ya que proveen al sujeto evaluado de claras especificaciones respecto de los criterios con los cuales se juzga su desempeño. A su vez, favorece la retroalimentación precisa al sujeto evaluado, pues le indica claramente el nivel de logro alcanzado en cada dimensión.

VIII. ¿Qué tipos de rúbricas existen?

Rúbricas analíticas	Rúbricas holísticas
<p>En una rúbrica analítica, se evalúan por separado cada una de las dimensiones de un constructo, para lo cual se plantean descripciones distintas de los niveles, en cada una de ellas (Moskal, 2000). Cuando se utiliza este tipo de rúbricas, los evaluadores emiten un juicio por separado y específico de cada una de tales dimensiones.</p>	<p>En una rúbrica holística, el evaluador emite un juicio sobre la totalidad del proceso o producto, sin juzgar por separado las dimensiones que lo componen. Se utiliza cuando hay una superposición entre las dimensiones y no es posible separar el constructo, o bien cuando se busca dar cuenta de una apreciación general, bajo el supuesto de que “el todo es más que la suma de las partes” (Moskal, 2000).</p>

A continuación, para una misma situación de evaluación del ámbito escolar, revisaremos un ejemplo de rúbrica analítica y uno de holística, planteadas para la tarea de presentar un trabajo.

Ejemplo de rúbrica holística: Evaluación de una presentación oral	
Nivel	Descripción
4	La presentación del trabajo es clara, consistente y denota manejo profundo del tema. Se utilizan diversos recursos para mantener interesada a la audiencia durante toda la disertación y se responden todas las preguntas en forma correcta.
3	La presentación del trabajo es clara y denota que se comprenden los aspectos centrales del tema. La mayor parte del tiempo la audiencia se muestra interesada.
2	Solo una parte de la presentación es clara, ciertos fragmentos son confusos. Durante muchos momentos la audiencia se distrae por completo y cuesta retomar su atención.
1	No hay claridad en la presentación de las ideas, se aprecia escaso manejo del tema. La audiencia no presta atención a la exposición.

Ejemplo de rúbrica analítica: Evaluación de una presentación oral				
	4	3	2	1
Manejo del tema	El presentador denota un muy buen manejo del tema , lo que se manifiesta en su precisión conceptual, su capacidad para explicar de distintos modos y dar respuesta a todas las preguntas de la audiencia.	El presentador denota un adecuado manejo del tema , lo que se manifiesta en su precisión conceptual y su capacidad para dar respuesta a la mayoría de las preguntas de la audiencia.	El presentador denota un manejo incipiente del tema , lo que se manifiesta en un uso adecuado, aunque a veces algo impreciso, de los conceptos centrales y en que solo da respuesta a algunas preguntas de la audiencia.	El presentador denota un escaso o nulo manejo del tema (no hay un uso adecuado de los conceptos ni puede responder preguntas de la audiencia).
Claridad de la presentación	Se aprecia claridad y consistencia durante toda la presentación.	Se aprecia claridad y consistencia durante la mayor parte de la presentación.	Se aprecia claridad y consistencia solo en la mitad de la presentación.	No hay claridad ni consistencia en la presentación.
Uso de recursos del lenguaje	Se maneja con fluidez y soltura el volumen de voz, la postura corporal y el contacto visual para mantener a la audiencia interesada durante toda la presentación.	Se maneja con fluidez y soltura el volumen de voz, la postura corporal y el contacto visual para mantener a la audiencia interesada durante la mayor parte de la presentación.	Se aprecia poca fluidez en el manejo del volumen de voz, de la postura corporal y del contacto visual, de modo que la audiencia pierde interés y cuesta retomarlos.	No se maneja bien el volumen de voz, la postura corporal ni contacto visual con la audiencia, de modo que esta se muestra muy desinteresada en la presentación.

Ambos tipos de rúbricas tienen ventajas y desventajas. Por una parte, las holísticas son más sencillas de construir y más rápidas de aplicar, pues se realiza la evaluación general de un desempeño sin ponderar detalles que puedan resultar muy específicos. Por esta misma razón, tanto para el evaluador como para el evaluado son poco informativas respecto de las fortalezas y debilidades observadas, ya que no registran detalladamente en cuáles aspectos o dimensiones el sujeto ha logrado lo esperado y en cuáles no. Por su parte, las rúbricas analíticas son más complejas de construir y más lentas de aplicar (se realizan varios juicios por separado), ya que se gradúan los distintos aspectos que componen una dimensión. Sin embargo, son enormemente más informativas, pues permiten un diagnóstico más claro de las fortalezas y debilidades observadas en los sujetos, ya que se pueden analizar los resultados de forma separada por cada dimensión. A su vez, si se utilizan para retroalimentar, indican con precisión al sujeto evaluado cuáles son los aspectos mejor logrados y cuáles no en cada dimensión que compone su desempeño.

IX. Pasos para construir una rúbrica

La construcción de las rúbricas debe hacerse en función de las decisiones tomadas al elaborar una tarea de desempeño. Al igual que en el caso de esta última, el proceso también comprende distintos pasos que en la práctica no siempre se desarrollan de manera lineal, pues cada decisión se debe revisar a la luz de la nueva evidencia recogida. En la figura 2, mostramos cada uno de los pasos y, posteriormente, describimos en qué consisten.

FIGURA 2
PASOS PARA CONSTRUIR UNA RÚBRICA



Paso 1: Determinar las dimensiones a evaluar

Si bien, se trata de un paso que debió abordarse al momento de definir el constructo en una tarea de desempeño, como la construcción es un proceso cíclico, es necesario volver a revisar las dimensiones definidas. Para ello, lo importante es comprobar que tanto la cantidad de dimensiones como la descripción de cada una logran representar el constructo y, a la vez, están en coherencia directa con la tarea o desempeño al que se aplicará la rúbrica.

Además, considerando el propósito y el contexto de evaluación, se debe decidir qué tipo de rúbrica es más adecuado utilizar: analítica u holística.

Paso 2: Determinar el número de niveles de desempeño

Una vez que se han acordado las dimensiones, se debe estudiar cuántos niveles de logro o de desempeño resultan apropiados para la rúbrica. Esta decisión depende del grado de complejidad de la tarea (lo que puede dar cabida a varios niveles de logro), del alcance del constructo y del conocimiento que posee el evaluador respecto de las distintas formas en que se manifiestan los desempeños. En general, la experiencia recomienda:

- Determinar al menos tres niveles, pues la distinción solo en dos niveles puede resultar demasiado gruesa y no aportar a la discriminación entre varios desempeños. La diversificación en más de cinco niveles tampoco se recomienda, ya que es muy poco probable que pasado ese número se hagan diferencias razonables o significativas.
- Considerar que las escalas impares ofrecen la tentación del punto medio, es decir, que muchas veces los jueces aplicadores de la rúbrica tienden a asignar los niveles de desempeño que se encuentran en el medio.
- Reflexionar sobre las etiquetas que se usarán para nombrar los niveles, pues mediante estas se entregan señales relacionadas con el propósito de la medición. Es decir, no es lo mismo, en una evaluación de desempeño utilizar las etiquetas: "insuficiente - elemental - adecuado", que "inicial - intermedio - avanzado".
- Elaborar una conceptualización general de cada nivel que permita orientar la descripción posterior en cada una de las dimensiones. Por ejemplo, si se ha decidido utilizar una rúbrica que distinga los niveles sobresaliente, logrado, mínimamente logrado y no logrado, se puede contar con una definición genérica como la que sigue:

Ejemplo de definición genérica de cuatro niveles de desempeño	
Sobresaliente	Corresponde a un nivel de desempeño que destaca o sobresale con respecto a lo esperado.
Logrado	Es el nivel de desempeño que describe lo esperado o adecuado para una dimensión.
Mínimamente logrado	Corresponde a un nivel de desempeño que cumple con aspectos básicos de una dimensión. Si bien cumple en forma irregular con lo esperado y requiere de mayor capacitación, hay evidencia de potencial en la dimensión evaluada.
No logrado	Corresponde a un nivel de desempeño claramente insuficiente con respecto a lo esperado en la dimensión. Señala debilidades importantes.

Como se puede observar en este caso, la opción de contar con cuatro niveles y su descripción implica el acuerdo de que el nivel "logrado" no es el más alto de la graduación, sino que se ha optado por incluir "sobresaliente", nivel que agrupa todas aquellas conductas que van más allá de lo solicitado o destacan por algún valor especial.

Paso 3: Describir los niveles de desempeño

La descripción de cada nivel se debe hacer en función de aquellas conductas o acciones que son observables y pueden ser apreciadas de manera directa durante la realización de una tarea o a partir del producto que se obtiene de ella. Por esta razón, la descripción tiene que ser:

- **Observable / medible:** Debe ser posible de apreciar directamente a través de la observación, por lo cual es preciso que refiera a conductas o acciones que operacionalizan los procesos internos en evaluación.
- **Coherente con la dimensión a la que pertenece:** Es decir, corresponder a una muestra inequívoca de la dimensión de la cual se desprende.
- **Relevante en su contenido:** Es indispensable que aborde aspectos centrales de la dimensión a la cual pertenece, aun cuando no la represente del todo.
- **Coherente con la tarea:** O sea, considerar la situación o tarea a partir de la cual se recogerá la información o evidencia.

La descripción de los niveles es una etapa compleja, pues se debe mantener la consistencia de las conductas que se están graduando al caracterizar cada nivel. Lo que varía es la calidad con la que se presentan las conductas (frecuencia de aparición u otro atributo que se considere relevante para esa dimensión).

Como la rúbrica muchas veces será aplicada por distintos jueces y no siempre se cuenta con períodos de capacitación en su uso, es fundamental que al describir cada nivel se utilice un **lenguaje claro y directo** que no esté sujeto a interpretaciones. Términos como “generalmente”, “muchas veces”, “suficiente”, “de calidad” deben evitarse y, en caso de emplearse es necesario especificar su alcance. Hay que considerar que cualquier diferencia en la interpretación de un nivel de desempeño generada por algún término confuso de la rúbrica, puede afectar la **confiabilidad** en la apreciación de ese desempeño. Es decir, “medir con niveles de exigencia distintos” implica emitir interpretaciones erradas respecto del constructo.

Para **describir en qué consiste** cada uno de los niveles de desempeño, se recomienda seguir una “lógica de pisos”, es decir, que en cada nivel se describan las conductas mínimas requeridas para pertenecer a ese nivel, lo cual implica que se puedan observar desempeños que sean un poco mejor que lo descrito para un nivel, pero no lo suficiente como para pertenecer al nivel siguiente. Esta forma de describir los niveles es fundamental al aplicar la rúbrica. Considérese lo siguiente:

- Al comenzar la descripción, se debe partir del nivel esperado, es decir, describir aquel nivel que englobe lo que significa realizar correctamente esa tarea en esa dimensión.
- Luego, se recomienda continuar describiendo los niveles inferiores, pensando qué aspectos del nivel esperado pueden ir presentándose con irregularidad o con menor calidad.
- El nivel más bajo suele ser una categoría residual, donde cabe todo lo que no cumple con las condiciones del anterior. Se puede optar por no describir nada específico y solo señalar que “no cumple con las condiciones del nivel siguiente”, o bien, detallar las formas que cumplirían con el nivel más bajo.

Finalmente, una manera de aportar evidencia de validez de contenido de las rúbricas es apoyar el proceso con la revisión de expertos, de tal manera que se reúna evidencia que ayude a validar la graduación. Los expertos tienen que juzgar la calidad de la rúbrica, respondiendo preguntas como:

- ¿La rúbrica entrega información relevante y central de la dimensión que evalúa?
- ¿La definición de cada nivel de desempeño es la adecuada?
- ¿Lo que se describe en cada nivel corresponde a la definición general del nivel?

Ejemplo de rúbrica: Evaluación del rol de los tutores

En el contexto de evaluar a través de una rúbrica el rol de los tutores, se propone la siguiente graduación para una de las dimensiones que componen el constructo:

Dimensión acompañamiento pedagógico técnico del tutor: Dimensión que comprende todas aquellas acciones de retroalimentación orientadas a informar al docente novato de sus logros y áreas de desarrollo, entregar directrices u orientaciones sobre cómo seguir aprendiendo.

No logrado	Mínimamente logrado	Logrado	Sobresaliente
No cumple con lo indicado en el nivel "Mínimamente logrado". O bien, al hacer la retroalimentación utiliza un trato irrespetuoso o manifiesta bajas expectativas sobre el futuro del docente novato.	Si bien entrega retroalimentación al docente novato, pues lo ayuda a reconocer sus áreas logradas y en desarrollo, no le entrega herramientas concretas para mejorar.	Entrega retroalimentación que resulta efectiva para el docente novato: • Lo ayuda a reconocer sus áreas logradas y en desarrollo. • Y, le entrega herramientas concretas para mejorar.	Cumple con todo lo señalado en el nivel "Logrado" y, además, sobresale porque: • Al hacer la retroalimentación ayuda al docente novato a organizar o priorizar las acciones para mejorar.

Algunas preguntas relevantes para evaluar la calidad de una rúbrica como la anterior son las siguientes:

- ¿La descripción de los niveles es clara?
- ¿La descripción de los niveles apunta a lo central de la dimensión y excluye aspectos irrelevantes?
- ¿El desempeño en cada nivel está planteado en términos observables?
- ¿Hay consistencia de contenido entre los niveles de desempeño, es decir, todos los niveles apuntan a graduar un mismo aspecto? ¿Hay una gradualidad equivalente entre los niveles de desempeño?

Otro aspecto a considerar al describir cada nivel, tiene que ver con la organización gráfica de la información. Considere que para ello es útil:

- Ubicar en una misma posición aquello que es común entre niveles de desempeño, lo cual ayuda a los evaluadores a aplicar la pauta.
- Cuando un mismo elemento tiene características distintas en los niveles de desempeño, se deben marcar gráficamente dichas diferencias (por ejemplo, subrayándolas).

Nivel 1	Nivel 2	Nivel 3	Nivel 4
Elemento A	Elemento A Elemento B2	Elemento A Elemento B1 Elemento C2	Elemento A Elemento B1 Elemento C1

Cálculo de puntaje

Si bien lo importante de una rúbrica es la información cualitativa que entrega, es decir, la descripción detallada de los niveles de desempeño asignados a la evidencia de un sujeto, si fuera necesario resumir o combinar distintas dimensiones para obtener un único puntaje, se debe considerar lo siguiente:

- En términos ideales, el puntaje debe surgir del cálculo más simple posible (por ejemplo, promedio de las dimensiones).
- También se pueden ponderar de modo distinto las dimensiones, según la importancia que quiera otorgarse a cada una.

Paso 4: Validación

Los estudios de pilotaje forman parte del proceso de validación de una evaluación. Consisten en poner a prueba la tarea y el instrumento que se usará para juzgar el desempeño. El propósito de esto es evaluar la adecuación de su diseño, en función de la interpretación y uso que se busca dar a los resultados. Esto es especialmente relevante para aquellas evaluaciones con altas consecuencias para los evaluados.

Puesta a prueba de la tarea, sus instrucciones y protocolos

Una vez que se ha determinado la tarea que se utilizará para evaluar un desempeño, se han construido las instrucciones para quienes serán evaluados y se han definido los protocolos de aplicación (en caso de ser necesario), es conveniente **poner a prueba** todos estos elementos con el fin de asegurar que la información recogida a través de la realización de la tarea permite evaluar el desempeño en cuestión.

Para ello, se deben considerar sujetos que, si bien no formarán parte de la población que se evaluará, comparten características centrales. Idealmente, se les solicita que realicen la tarea en condiciones similares a las que tendrán los sujetos en evaluación. Si esto no es posible, se pide a los sujetos que desarrollen una parte de la tarea de especial interés para los

diseñadores de la evaluación. En ambos casos, puede resultar especialmente enriquecedor instarlos a explicar las razones que los llevaron a **responder** de una u otra forma a la tarea, de modo que sea posible analizar el proceso de respuesta de los individuos y determinar si es coherente con lo que se buscaba.

Algunas preguntas que pueden hacerse al momento de poner a prueba la tarea, son:

- ¿Fueron la formulación de la tarea y sus instrucciones suficientemente claras y completas para los sujetos?
- ¿Qué estrategia están usando los sujetos para abordar la tarea?
- ¿Están los sujetos respondiendo por las razones que se espera?
- Las instrucciones, ¿**suscitan o provocan** los procesos propios del constructo que se quiere evaluar?

La información que se recoja puede conducir a reconsiderar el formato de la tarea, ajustar las instrucciones o modificar los protocolos de aplicación.

Puesta a prueba de la rúbrica

Implica la capacitación de jueces en el uso de la rúbrica y la posterior aplicación de esta a un conjunto delimitado de **muestras** del desempeño que se está evaluando, las que pueden corresponder a evidencia producto del pilotaje de la tarea. Tanto el proceso de aplicación de la rúbrica como los resultados obtenidos, entregan información central para determinar si es necesario realizar ajustes o modificaciones en ella.

La **muestra** corresponde a una parte o subconjunto de elementos de una población que representa las características de la misma (Fontes de Gracia *et al.*, 2010, p. 282).

El análisis de los resultados de la aplicación nos permite tener una primera aproximación al comportamiento estadístico de la rúbrica (por ejemplo, desempeño promedio, distribución de los distintos niveles, etc.).

Los resultados también nos pueden entregar evidencia preliminar respecto de la **confiabilidad** entre evaluadores. Por supuesto, la calidad de estas aproximaciones depende de la conformación de la muestra, tanto en magnitud como en similitud con la población real a evaluar.

La **confiabilidad** entre los evaluadores hace referencia al nivel de coherencia con el que dos o más de ellos califican el desempeño de los evaluados.

Es importante señalar que, antes de aplicar el instrumento en forma piloto, los diseñadores suelen tener **hipótesis** respecto del comportamiento que tendrá, con base en experiencias previas, estudios o evaluaciones similares. Estas hipótesis se contrastan con los resultados obtenidos y se buscan interpretaciones que les den sentido a estos últimos o cuestionen algún aspecto de su diseño.

Las opiniones de los jueces sobre el proceso de aplicar la rúbrica (dificultades, dudas, etc.) reportan valiosa información cualitativa respecto de su claridad y cómo mejorarla. Por ejemplo, se les puede consultar a propósito de la presencia de términos ambiguos o diferencias poco nítidas entre niveles de desempeño, que conlleven la necesidad de sustituir un determinado término o incluir aclaraciones.

La información recogida puede conducir a reconsiderar el formato de la tarea, ajustar las instrucciones o modificar los protocolos de aplicación.

Consideraciones finales: ideas fuerza

Para concluir, algunas ideas centrales que hemos revisado son las siguientes:

- Las evaluaciones basadas en desempeños o ejecuciones se utilizan cuando el constructo a evaluar se relaciona con un **saber hacer** o un **saber ser**. En estos casos, la observación es la forma más adecuada de recoger información sobre cómo se ejecuta este desempeño.
- Cuando se elabora una tarea para evaluar desempeños, junto con definir el propósito, la población y el constructo a evaluar, es fundamental reflexionar sobre cuál es la tarea más idónea para observar ese constructo (en ese contexto y para esa población).
- Las instrucciones o los protocolos que acompañan la realización de una tarea cobran vital importancia para evitar que la variabilidad de resultados sea producto de aspectos externos al desempeño del sujeto.
- Dado que las evaluaciones a través de tareas de desempeño necesitan de otro que observe esa ejecución, es importante evitar los sesgos del evaluador o juez mediante el empleo de instrumentos que permitan juzgar lo que se está observando, por ejemplo, las rúbricas.
- Las rúbricas, a diferencia de otros instrumentos, como las listas de cotejo y las escalas de apreciación, aumentan la objetividad de la evaluación, ya que cada nivel de logro está claramente definido. Además, permiten retroalimentar con mayor precisión y suficiencia a los evaluados.
- Como parte del proceso de validación de las tareas de desempeño y sus respectivas rúbricas, es necesario poner a prueba tanto las instrucciones o protocolos de la tarea como la aplicación de la rúbrica a distintas muestras de desempeño.

Referencias

- AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.
- CORRAL-VERDUGO, V. (2001). El significado de "variables latentes" en psicología. *Acta Comportamental: Revista Latina de Análisis del Comportamiento*, 9(1), 85-98. Recuperado de <http://www.journals.unam.mx/index.php/acom/article/view/14634>
- FONTES DE GRACIA, S., García-Gallego, C., Quintanilla Cobián, L., Rodríguez Fernández, R., Rubio de Lemus, P., y Sarriá Sánchez, E. (2010). *Fundamentos de investigación en psicología*. Madrid: UNED.
- KORETZ, D. (2008). *Measuring Up: What Educational Testing Really tells Us*. Cambridge: Harvard University Press.
- MINISTERIO DE EDUCACIÓN, Chile. (2014). *Preguntas y rúbricas de la Entrevista por un Evaluador Par 2014*. Recuperado de: <http://www.docentemas.cl/descargas/instrumentos-complementarios/Preguntas-EEP-2014.pdf>
- MOSKAL, B. M. (2000). Scoring rubrics: what, when and how? *Practical Assessment, Research & Evaluation*, 7(3). Recuperado de <http://PAREonline.net/getvn.asp?v=7&n=3>.

