# Capítulo 2

# VALIDEZ, CONFIABILIDAD Y AMENAZAS A LA VALIDEZ

Blanca Ariadna Carrillo Ávalos, Melchor Sánchez Mendiola

*"La validez es simple. La validación puede ser difícil."*
*Michael Kane, 2009*

## INTRODUCCIÓN

A lo largo de la vida como docentes realizamos muchas evaluaciones para intentar conocer el nivel de conocimiento o desempeño de nuestros estudiantes. Este proceso implica la elaboración, aplicación e interpretación de exámenes de diferentes tipos: diagnósticos, formativos y sumativos. Independientemente de su finalidad, la meta de cualquier evaluación incluye la identificación del nivel de algún constructo, como conocimiento sobre química orgánica, habilidad para resolver problemas de trigonometría, o competencia de comunicación escrita. Los resultados de las evaluaciones idealmente deben reflejar de una manera precisa y reproducible lo que se pretende evaluar, para poder interpretar de forma racional los resultados de la misma y estar en capacidad de realizar inferencias y tomar decisiones con fundamentos sólidos. Cuando evaluamos a nuestros estudiantes sobre un tema particular, deseamos identificar el proceso y resultados del aprendizaje que permitan inferir el nivel de desempeño en los constructos de interés. Después de aplicar las evaluaciones, obtenemos resultados en forma de puntuaciones que ayudan a tomar decisiones, que conllevan las siguientes interrogantes: ¿estamos evaluando exactamente lo que deseamos evaluar?, ¿qué implican los resultados con respecto al avance académico del alumno?, si se trata de una evaluación sumativa, ¿cuál es la calificación mínima para aprobar el curso?, ¿qué tan reproducible es la medición?, entre muchas otras.

La evaluación en educación es una disciplina cada vez más sofisticada y sustentada en investigación, que requiere incorporar conceptos académicos fundamentales para llevarse a cabo con profesionalismo y solidez metodológica (Instituto Nacional para la Evaluación de la Educación, 2017). El pilar conceptual más importante de la evaluación en educación es la validez, tema del que trata este capítulo. Actualmente el concepto de **validez** ha evolucionado del tradicional "medir lo que se pretende medir", a un modelo más amplio y profundo, en

el que *"se refiere al grado en que la evidencia y la teoría respaldan las interpretaciones de los puntajes de una prueba para usos propuestos de las pruebas"* (AERA, APA y NCME, 2018). Es un conjunto de acciones que se ubican a lo largo del proceso evaluativo, para fundamentar la interpretación de los resultados y así generar inferencias. El análisis de validez, o **validación**, es el proceso mediante el que evaluamos las evidencias que se presentan para determinar cuál es el grado de validez (Cook y Hatala, 2016). Se puede realizar para los diferentes tipos de exámenes, diagnósticos, formativos y sumativos, aunque es particularmente relevante para las evaluaciones sumativas de alto impacto.

Tradicionalmente la validez en educación se clasificaba como "las 3 Cs": validez de contenido, de criterio y de constructo (Cronbach y Meehl, 1955). En la definición actual esta distinción desapareció, ya que el modelo vigente propone diversas fuentes de evidencia que arrojan luz sobre distintos aspectos de la validez, no es que reflejen diferentes tipos de la misma. La validez es un concepto unitario, por lo que se considera que toda la validez es validez de constructo. La palabra **constructo** significa colecciones de conceptos abstractos y principios, inferidos de la conducta y explicados por una teoría educativa o psicológica, es decir, atributos o características que no pueden observarse directamente, por ejemplo: inteligencia emocional, extroversión, conocimientos sobre matemáticas (AERA, APA y NCME, 2018).

La validez es un juicio valorativo holístico e integrador que requiere múltiples fuentes de evidencia para su interpretación, e intenta responder a la pregunta: *"¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen?"* (Downing, 2003). Validez no es una propiedad intrínseca del examen, sino del significado de los resultados en un entorno educativo específico y las inferencias que pueden hacerse a partir de los mismos (Carrillo-Ávalos et al., 2020; Downing, 2003). Por ejemplo, los resultados de los exámenes de admisión a las universidades no deben interpretarse categóricamente como evidencia predictiva de que la persona vaya a tener éxito en la vida, ya que el examen no está diseñado con ese propósito. Quienes desarrollan pruebas e interpretan sus resultados, deben poseer nociones básicas de los conceptos de validez y confiabilidad, para incorporarlos de manera apropiada en el proceso de enseñanza y aprendizaje. Este capítulo presenta un panorama básico de dichos conceptos, además de describir las principales amenazas a la validez.

## VALIDEZ

Anteriormente se hablaba de tres tipos de validez: de contenido, de constructo y de criterio; la de criterio se dividía en validez concurrente y validez predictiva (Cronbach y Meehl, 1955). Posteriormente, a finales del siglo XX, un nuevo marco de referencia de validez fue propuesto y aceptado por las principales organizaciones de evaluación educativa y pruebas psicológicas (American Educational Research Association et al., 2018), incorporando el concepto holístico de validez de constructo. Este modelo establece que, para determinar el grado de validez de los usos e interpretaciones de los resultados de una evaluación, se deben

proveer diversos elementos que lo demuestren (Downing, 2003). Este esquema propone los siguientes elementos como **cinco fuentes de evidencia** de validez (Downing, 2003; Messick, 1989):

1) **Evidencia basada en el contenido de la prueba.** El contenido de la prueba alude a los conocimientos que evalúa; por ejemplo, en el caso de un examen final de un curso de biología evolutiva, son los relacionados con la historia de la teoría de la evolución y sus evidencias, sus bases genéticas, los procesos del cambio evolutivo, el origen de las especies, macro-evolución y la evolución de los homínidos. Es decir, todos los temas que debe saber el alumnado cuando termina de estudiar esta asignatura.

   Para demostrar que los usos e interpretaciones de un examen de esta naturaleza son adecuados, podemos buscar evidencia de que, en efecto, las preguntas del examen versan sobre los temas mencionados. Estas evidencias descansan en cuatro elementos (Sireci y Faulkner-Bond, 2014):

   - *Definición del dominio.* Se trata de la descripción detallada de las áreas del contenido y las habilidades cognitivas que se desean evaluar del constructo definido en el currículo, o de los resultados de la actividad de aprendizaje. En la tabla de especificaciones de la prueba se deben enlistar las subáreas del contenido y los niveles cognitivos que estaremos midiendo (conocimiento, comprensión, aplicación, análisis, síntesis o evaluación), así como los estándares del contenido y los objetivos curriculares.
   - *Representación del dominio*. Con frecuencia se hacen demasiadas preguntas sobre un tema y se dejan de lado otros. Para decidir cuántos ítems, preguntas o reactivos corresponden a cada tema, podemos establecer una tabla de ponderaciones en la que se establece la relevancia de cada uno, a los de mayor importancia se asignan más preguntas en el examen. También se debe buscar la alineación entre el contenido del examen y los estándares para ese conocimiento. Por ejemplo, cada asignatura que cursa un alumno tiene una carta descriptiva en la que se han establecido los objetivos o metas de aprendizaje, lo que se debe saber al terminar de cursar esa materia. En un examen sumativo, las preguntas deberán contemplar tales objetivos y coincidir con la tabla de especificaciones. Otra manera de demostrar alineación es que expertos en el tema opinen sobre el examen, cómo los ítems evalúan los diferentes aspectos del constructo de interés.
   - *Relevancia del dominio*. Se refiere a qué tan importantes son los ítems con respecto al aspecto del constructo que se está midiendo, que se pregunten conceptos importantes y no datos triviales.
   - *Procedimientos apropiados de diseño de la prueba*. Los procedimientos que se llevan a cabo al diseñar la prueba deben servir para asegurarse de que su contenido evalúa fielmente y representa por completo al constructo de interés. Esto puede comprobarse al implementar controles de calidad durante el desarrollo del examen, como revisión de ítems por expertos en contenido para asegurarse de su veracidad técnica,

revisión por expertos en evaluación para verificar que estén bien elaborados, probar los ítems por medio de estudios piloto.

- *Credenciales de los creadores del examen, elaboradores de reactivos y expertos en contenido*. Es importante documentar que las personas que intervienen en el proceso de diseño del examen, definición del constructo, elaboración de los reactivos y análisis de los resultados, tengan las credenciales correspondientes para dar certidumbre a todo el proceso. Si el examen evalúa habilidades de traducción del idioma alemán, deben participar expertos en contenido que dominen dicho constructo; si se trata de elaborar un examen práctico de habilidades de comunicación verbal, deben participar expertos en el tema y profesionales del diseño de este tipo de exámenes.

2) **Evidencia basada en los procesos de respuesta.** Los procesos de respuesta son los procesos mentales que lleva a cabo el sustentante cuando contesta las preguntas de una prueba. Se esperaría que responda cada ítem integrando los conocimientos que se indagan y que no esté tratando de adivinar la respuesta correcta. Si el examen contiene preguntas de opción múltiple (POM) o abiertas, cuando el alumno termine de contestar se puede indagar cómo llegó a la respuesta a través de una entrevista cognitiva, así se puede saber si comprendió los conceptos clave, si hay errores de redacción en las preguntas, cuál fue el razonamiento que utilizó, y si hay posibilidad de que existan falsos positivos (que el alumno haya utilizado un razonamiento erróneo para llegar a la respuesta). Al final, lo que se busca es que el estudiante aplique en verdad los conocimientos adquiridos para resolver los problemas que se proponen en el examen y que luego los pueda aplicar en la vida real. Otra manera de aportar a esta evidencia es a través de modelos matemáticos que evalúen la dificultad y el tiempo que tardan en contestar cada ítem (Padilla y Benítez, 2014).

   Algunos autores agregan en este apartado la familiaridad de los sustentantes con el formato del examen (por ejemplo, evaluación asistida por computadora), la validación de la hoja de respuestas correctas, el control de calidad del reporte de los resultados, entre otros (Downing, 2003).

3) **Evidencia basada en la estructura interna.** La estructura interna presenta tres características básicas: dimensionalidad, funcionamiento diferencial y confiabilidad (Rios y Wells, 2014). Al diseñar la prueba, se debe determinar cuáles dimensiones se desean evaluar sobre el constructo de interés, y esta información se describe en la tabla de especificaciones del examen. Siguiendo el ejemplo de un examen de biología evolutiva, una dimensión sería el proceso cognitivo que siguen los alumnos para resolver problemas de ese tema. Sin embargo, en ocasiones los exámenes contienen ítems que evalúan diferentes dimensiones del mismo constructo, por ejemplo, procesos cognitivos y valores. Para saber cuántas dimensiones posee el examen, se pueden hacer diversos análisis, entre ellos un análisis factorial confirmatorio, el que también permite identificar cuáles ítems valoran la misma dimensión buscando la relación existente entre ellos. Con el resultado de este análisis podemos justificar, por ejemplo, dar el mismo valor a todos los ítems, porque todos evalúan la misma dimensión del mismo constructo.

Las pruebas también deben aportar resultados imparciales, por lo que es útil buscar información con respecto al funcionamiento diferencial de los ítems (DIF, por sus siglas en inglés) (Leenen, 2014; Rios y Wells, 2014). Se trata de que las características de los ítems de la prueba son comparables entre diferentes grupos; es decir, que no habrá diferencia entre los resultados de hombres y mujeres, o por edades, por ejemplo. Para conocer el grado de invarianza de los resultados entre grupos, se pueden realizar pruebas basadas en el funcionamiento diferencial del ítem, como un análisis factorial confirmatorio de grupos múltiples.

Este apartado de evidencia de validez también incluye el análisis estadístico de los reactivos de la prueba (Downing, 2003), para documentar diversos indicadores como grado de dificultad, índice de discriminación, confiabilidad, error estándar de medición, curvas características del ítem, entre muchos otros, para lo cual existen diversos modelos y métodos matemáticos que se describen en otros capítulos de esta obra.

4) **Evidencia basada en las relaciones con otras variables.** En el ejemplo mencionado, los alumnos que hicieron el examen de la unidad de las bases genéticas de la evolución en la asignatura de biología evolutiva, también contestaron uno de genética, en donde se les hicieron preguntas acerca de genética de poblaciones. Ambas pruebas son semejantes en cuanto al constructo evaluado. La validación de la interpretación de los resultados del examen de biología evolutiva podría incluir el análisis de la relación convergente con los resultados del examen de genética, si ambas pruebas valoran constructos similares. Por otro lado, también puede existir una relación divergente cuando se evalúan constructos diferentes. El análisis para establecer ambas correlaciones se puede realizar por medio de la matriz multirasgo-multimétodo (MTMM) de Campbell (Campbell y Fiske, 1959), en el que se establece una matriz de correlaciones entre las dos pruebas en donde se representan al menos dos características y cada una de ellas debe ser medida por lo menos por dos métodos.

   Esta fuente de evidencia proporciona información acerca del grado en que la relación divergente o convergente es coherente con el constructo cuya medición es la base de la interpretación de los resultados de la prueba. Otra evidencia que aporta a las relaciones con otras variables es la relación entre la prueba y el criterio, la que se puede establecer por medio de uno de estos diseños:

   - *Estudio predictivo*. Para conocer el grado de relación entre el resultado de la prueba y el resultado del criterio que se evalúa posteriormente. Por ejemplo, si en el examen de admisión a la universidad se evalúa biología general, podríamos tratar de responder si las calificaciones de biología general en el examen de admisión predicen las de biología evolutiva. En este caso, el criterio sería el resultado de biología evolutiva. Por otro lado, también se pueden hacer estudios para hacer predicciones diferenciales por grupo de edad, sexo, antecedentes académicos, entre otras.
   - *Estudio concurrente*. Se mencionó en el ejemplo que los alumnos llevan las asignaturas de genética y biología evolutiva al mismo tiempo, por lo que las evaluaciones

correspondientes también ocurrieron en fechas cercanas. Si se evalúa un constructo al mismo tiempo, podemos estimar la relación entre las puntuaciones de la prueba y del criterio.

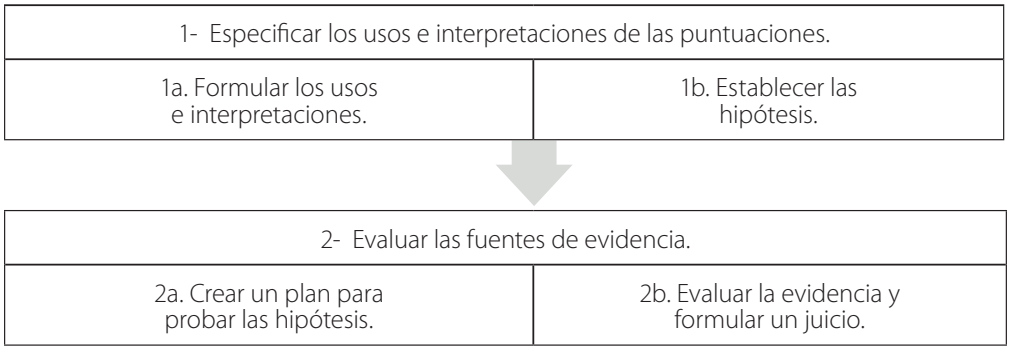
5) **Evidencia basada en las consecuencias de la prueba.** Los resultados de las pruebas, sobre todo las sumativas y de alto impacto como los exámenes de admisión o de titulación, tienen grandes consecuencias para los sustentantes. En el caso de los exámenes de admisión a las universidades, uno de los objetivos puede ser seleccionar a los sustentantes con mayor probabilidad de tener éxito académico (Downing, 2003). Las consecuencias de pruebas sumativas de impacto moderado son importantes también, ya que se van agregando para generar un promedio final que permita el avance en las trayectorias escolares.

   Las fuentes de evidencia de consecuencias de las pruebas se buscan al analizar cuáles son los impactos de los resultados en el avance académico del estudiante, en sus familias, y en la sociedad. En las pruebas cuyos resultados se reportan con base en criterio, la decisión del punto de corte mínimo debe fundamentarse; en el caso de pruebas con referencia a norma también se debe sustentar la decisión de, por ejemplo, por qué admitir a los 100 sustentantes con las puntuaciones más altas. Otras fuentes de evidencia en este rubro pueden ser las consecuencias de aprobar o reprobar, de falsos positivos y falsos negativos, y las consecuencias institucionales. La manera de obtener esta información puede ser a través de entrevistas, grupos focales, cuestionarios, para conocer cuáles son los componentes más importantes de los programas académicos y sus puntos de mayor impacto (Lane, 2014).

## VALIDACIÓN

La validación es un proceso que se debe planear al mismo tiempo que se diseña la prueba, para asegurarse de contar con las fuentes de evidencia necesarias para obtener el mayor grado posible de validez de la interpretación de sus resultados. Una manera de realizar este proceso se sugiere a continuación (Figura 1).

**Figura 1. Esquema de pasos generales para el proceso de validación de una prueba (elaboración propia)**



## 1. Especificar los usos e interpretaciones de las puntuaciones

*1a. Formular los usos e interpretaciones.* Los usos y las interpretaciones de las puntuaciones que se obtienen en una prueba son conceptos diferentes y ambos se deben aclarar desde que inicia el diseño de la prueba. La justificación del uso de las puntuaciones se puede conocer respondiendo a preguntas como: ¿debemos utilizar estas puntuaciones para tomar decisiones sobre quiénes pueden ingresar a un programa de posgrado? Para ello, se deben conocer las características de los usuarios principales, quiénes son las personas que presentan dichas evaluaciones; además, también son de interés las instituciones que las desarrollan, profesores, personal administrativo.

En cuanto a las interpretaciones, se pueden determinar al responder preguntas como: ¿las calificaciones del examen de graduación reflejan las competencias que debe poseer un egresado de este programa de posgrado?, ¿los resultados de esta prueba miden el nivel de desempeño de los estudiantes en esta asignatura?

Los datos que se deben evaluar en este paso se describen en la Tabla 1.

**Tabla 1. Formulación de los usos e interpretaciones de una prueba o examen**

| Dato necesario | Pregunta a contestar | Ejemplo |
|---|---|---|
| El objetivo de la prueba . | ¿Para qué se aplica la prueba? | Establecer el nivel de desempeño en los conocimientos acerca de Biología Evolutiva. |
| Los usuarios propuestos . | ¿Quiénes utilizarán los resultados de la prueba? | La Facultad de Ciencias Biológicas de la universidad. |
| Los usos. | ¿Para qué se utilizarán las puntuaciones? | Las calificaciones serán utilizadas para determinar cuáles alumnos demuestran un desempeño satisfactorio en la asignatura. |
| El constructo medido. | ¿Cuál es constructo que se evalúa con esta prueba? | Los conocimientos en Biología Evolutiva. |
| Las interpretaciones de los resultados. | ¿Los resultados serán interpretados con referencia a norma o a criterio?<br>¿Los resultados de esta prueba miden el nivel de desempeño de los estudiantes en esta asignatura? | Los resultados del examen final de la asignatura se interpretan con referencia a criterio, por lo que quienes obtengan una puntuación mayor a 6.0 tendrán un desempeño satisfactorio. |
| La población examinada. | ¿Quiénes contestarán el examen? | Los alumnos de 7° semestre de la licenciatura en Biología. Porcentaje de hombres y mujeres, rango de edades. |

*1b. Establecer las hipótesis.* Las hipótesis son preguntas que nos podemos hacer acerca de la evaluación que se está elaborando. Deben probarse por medio de las fuentes de evidencia mencionadas. Algunos ejemplos de hipótesis se describen en la Tabla 2.

**Tabla 2. Ejemplos de hipótesis que pueden buscar comprobarse durante el proceso de validación**

| Hipótesis | Fuente de evidencia |
|---|---|
| Los contenidos evaluados coinciden con los objetivos de aprendizaje.<br>Las preguntas utilizadas conforman una muestra adecuada del posible universo de preguntas sobre los mismos temas.<br>La ponderación de cada tema evaluado es correcta.<br>El nivel de complejidad de los ítems utilizados es adecuado con respecto al nivel de estudios. | Evidencia basada en el contenido de la prueba. |
| Los procesos mentales que llevaron a cabo los estudiantes son los esperados. | Evidencia basada en los procesos de respuesta. |
| El examen es unidimensional (si eso era lo planeado). Si es multidimensional, la calificación de cada dimensión es consistente con el constructo evaluado.<br>La confiabilidad de los resultados es adecuada.<br>Si se utilizan las puntuaciones de diferentes partes de la prueba como sub-puntuaciones, se especifica y se justifica cómo se combinan dichas sub-puntuaciones. | Evidencia basada en la estructura interna. |
| Existe relación entre los resultados de otras evaluaciones (que también poseen un nivel de confiabilidad adecuado) y la evaluación sumativa que examinan el mismo tema.<br>Se toman en cuenta las variables confusoras que pueden sesgar los resultados para grupos específicos.<br>Se demuestra que el desempeño en la prueba predice el desempeño en el criterio evaluado. | Evidencia basada en las relaciones con otras variables. |
| Las consecuencias negativas involuntarias no son graves y son menores a las consecuencias positivas. | Evidencia basada en las consecuencias de la prueba. |

## 2. Evaluar las fuentes de evidencia

*2a. Crear un plan para probar las hipótesis.* Con base en las hipótesis seleccionadas, se buscan las fuentes de evidencia y se reúne la información correspondiente.

*2b. Evaluar la evidencia y formular un juicio.* En este último paso se evalúan todas las evidencias en orden y se establece el grado de validez de la interpretación de las puntuaciones de la prueba evaluada. Este grado dependerá de la calidad de las evidencias presentadas y también de las evidencias más importantes, según la prueba.

## AMENAZAS A LA VALIDEZ

Además de analizar las fuentes de evidencia de validez, se sugiere identificar elementos que puedan afectar el grado de validez de los resultados de la evaluación. Este paso es importante porque da fortaleza a las decisiones que se toman con base en los resultados de la prueba. Los elementos que reducen el grado de validez se les denomina amenazas a la validez; se les llama así porque interfieren con la correcta interpretación de las puntuaciones (Carrillo-Ávalos et al., 2020; Downing y Haladyna, 2004). Estas amenazas pueden estar presentes en cualquier tipo de evaluación. En general, se reconocen dos tipos de amenazas a la validez: subrepresentación del constructo (SC) y varianza irrelevante al constructo (VIC) (Downing, 2003; Messick, 1989). Para explicar estos conceptos, partiremos de la teoría clásica de los tests (TCT). De acuerdo con esta, la puntuación que se obtiene a partir de las respuestas de un examen (puntuación observada o total = X) está compuesta por la suma de la puntuación verdadera (V) más el error aleatorio ($E_a$) (de Champlain, 2010; Schuwirth y van der Vleuten, 2011):

$$X = V + E_a$$

De esta manera, los factores que tienen un efecto sistemático sobre la puntuación observada "X" se agregan a la puntuación verdadera "V". Estos factores incluyen tanto al constructo de interés, como a otros factores sistemáticos que se presentan en todas las versiones del examen, pero cuya medición no es el objetivo; por ejemplo, que un reactivo de opción múltiple no tenga una respuesta correcta entre las opciones, por lo que todos los alumnos la contestarán mal. Por otra parte, el error aleatorio ($E_a$) está conformado por todas las condiciones que aportan variabilidad a la puntuación verdadera "V" de manera no sistemática, ya que son diferentes en cada ocasión que se aplica el examen y para cada persona, como el cansancio, el desvelo y el estrés. En realidad, no conocemos el valor del error aleatorio ni el de la puntuación verdadera, aunque con diversos métodos estadísticos se pueden estimar (Haladyna y Downing, 2004).

Con base en lo anterior, la puntuación verdadera se puede dividir en dos partes: la puntuación del constructo de interés (θ), más la puntuación causada por factores sistemáticos ($E_s$). Por lo tanto, la fórmula quedaría así (Haladyna y Downing, 2004):

$$X = \theta + E_s + E_a$$

A partir de esta fórmula podemos definir los dos tipos de amenazas a la validez. Una de ellas se presenta cuando se mide θ a través de ítems que representan *de forma incompleta* el dominio del conocimiento que queremos evaluar: esto es la subrepresentación del constructo. Por otro lado, la varianza irrelevante al constructo se asocia con el error sistemático $E_s$, que es causado por la medición involuntaria de elementos que no son el objetivo de la medición, e interfieren con la medición del constructo de interés y, por lo tanto, con la validez de la interpretación de los resultados (Downing y Haladyna, 2004; Haladyna y Downing, 2004; Mes-

sick, 1989). La varianza indeseable de X debida a $E_a$ también aporta elementos para conformar una amenaza a la validez; sin embargo, sus factores conllevan a una confiabilidad baja, de lo que hablaremos en la siguiente sección de este capítulo.

A continuación, se describen ejemplos de los tipos de amenazas a la validez:

- **Subrepresentación del constructo (SC).** En el ejemplo del examen sumativo de biología evolutiva mencionado, supongamos que consta de 10 preguntas, que constituyen una muestra del universo de todas las preguntas posibles sobre el mismo tema y que miden el mismo constructo. Se considera que 10 ítems son pocos para evaluar a cabalidad los temas involucrados, por lo que no son suficientes para representar adecuadamente al constructo de interés ni a las dimensiones que deseamos evaluar. Otros problemas que podríamos encontrar en esta muestra es que se estén evaluando contenidos triviales, contenidos diferentes a los estudiados, que solo pregunten aspectos de un solo tema o que la confiabilidad sea insuficiente. Es evidente que estos problemas están relacionados con la evidencia de validez de contenido. Si aplicáramos este examen de 10 reactivos, las amenazas a la validez pueden ser: número insuficiente de preguntas, sesgo, concordancia del dominio, baja confiabilidad (Moreno et al., 2004).
- **Varianza irrelevante al constructo.** Esta amenaza tiene origen en el error sistemático $E_s$ causado por una variable irrelevante al constructo de interés. Por ejemplo:

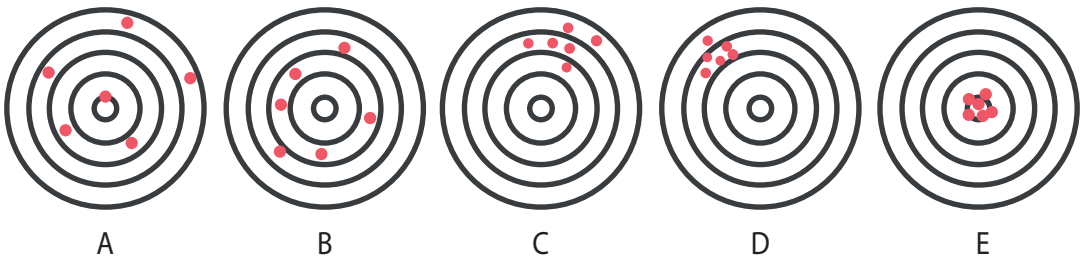
    - *Ítems defectuosos.* La escritura de un reactivo de opción múltiple de calidad debe cumplir con los estándares establecidos por expertos para evitar que aporten varianza irrelevante al constructo (Haladyna et al., 2002; National Board of Medical Examiners, 2016). La baja calidad de una pregunta la hace más difícil de contestar, convirtiéndola en un ítem defectuoso. Otros defectos posibles son que entre las opciones de respuesta se encuentre "todas las anteriores" o "ninguna de las anteriores", o el número de opciones (varios expertos consideran que el número óptimo es de tres a cuatro) (Abad et al., 2001; Haladyna et al., 2019; Rodríguez, 2005).
    - *Formato en negativo.* Varios expertos consideran que escribir preguntas en formato negativo debe evitarse porque se corre el riesgo de escribir un doble negativo (la pregunta y una o más de las opciones contienen términos negativos, por lo que identificar las opciones incorrectas implica negarlas, o sea, indicar que no son correctas) (Chiavaroli, 2017). También existe el riesgo de que el alumno no reconozca la negación en la pregunta, aunque se destaque en negritas "no" o "excepto", y que el proceso de respuesta llevado a cabo para responder no sea el esperado.
    - *Lenguaje.* Una pregunta que presenta una viñeta demasiado extensa o con explicaciones innecesarias ocasiona que el estudiante pierda tiempo leyéndola y en realidad puede evaluar la velocidad de lectura en lugar de medir el constructo deseado (Hicks, 2011; National Board of Medical Examiners, 2016). Por otro lado, las oraciones deben estar bien estructuradas, de manera que la pregunta quede clara.

- *Discordancia con el dominio*. Cuando no hay relación entre lo que se ha estudiado y lo que se está evaluando.
- *Ítems muy fáciles o difíciles*. Reactivos que seguramente todos los alumnos contestarán correctamente, lo que aumentará el promedio del grupo y el individual de manera artificial, lo que interfiere con la validez. Por otro lado, no discrimina entre estudiantes de bajo y alto desempeño.
- *Hacer trampa*. Este es uno de los retos principales de los profesores al aplicar exámenes, pues los estudiantes pueden hacer trampa de muchas formas y obtener ventaja de manera deshonesta: voltear a ver el examen de un compañero, utilizar un "acordeón", conocer las preguntas antes de hacer el examen, e incluso utilizar dispositivos digitales. La presencia de estas conductas genera falsos positivos, e incrementa la varianza sistemática en las puntuaciones.
- *Decisión indefendible de puntuación aprobatoria*. En general, en México los exámenes tienen una calificación mínima aprobatoria de 6.0. Si estos exámenes no son reproducibles o la distribución de las calificaciones es demasiado amplia (calificaciones demasiado altas y demasiado bajas), es difícil justificar por qué el 6.0 es la calificación de pase (Norcini, 2003).
- *Enseñar a la prueba ("teaching to the test")*. Cuando los alumnos han recibido preparación para contestar preguntas de un examen específico, como los exámenes de admisión a las licenciaturas o incluso los exámenes que elabora un profesor en particular, ello añade ventaja a algunos estudiantes y distorsiona el verdadero propósito de la enseñanza (Bond, 2008). Se ha visto que en ocasiones los estudiantes acceden a bancos de preguntas y prefieren aprenderlas de memoria en lugar de aprender los contenidos de los temas estudiados. Este factor es una amenaza a la fuente de evidencia de validez de relación con otras variables, porque no será posible generalizar los resultados de la prueba con el resto de los ítems del universo que miden el constructo deseado.
- *Habilidad para responder exámenes (testwiseness)*. Los alumnos que han presentado una gran cantidad de exámenes escritos a lo largo de su vida, se convierten en expertos para identificar la respuesta correcta en un ítem defectuoso. Por ejemplo, el hecho de que con frecuencia la opción correcta suele ser la más larga, la que tiene mayor detalle y está mejor escrita. Los alumnos que desarrollan estas habilidades como estrategias para responder pruebas, ocasionan que las respuestas no demuestren con veracidad el constructo de interés. Otro concepto relacionado es la "adivinanza educada" (*educated guessing)*, que consiste en que contestan por eliminación de las opciones menos probables, mas no porque realmente sepan la respuesta (Jurado-Núñez y Leenen, 2016).

## CONFIABILIDAD

La confiabilidad es la característica de las evaluaciones que se refiere a que los puntajes sean consistentes de persona a persona, de instrumento a instrumento y de un conjunto de ítems a otro dentro del mismo universo de ítems (Cizek, 2009). Va de la mano de la validez en cuanto a que depende de la interpretación de las puntuaciones de la prueba, y forma parte de la fuente de evidencia de validez basada en la estructura interna (Tavakol y Dennick, 2011). Sin embargo, una prueba puede ser confiable, pero tener validez limitada. Tomemos como ejemplo un tiro al blanco con las marcas de los dardos, haciendo una analogía con las puntuaciones de una prueba de persona a persona, de instrumento a instrumento o de un conjunto de ítems a otro dentro del mismo universo de ítems (Figura 2).

**Figura 2. Esquema de "tiro al blanco" para visualizar los conceptos de validez y confiabilidad**




A. Prueba no confiable, sin validez.

B. Confiabilidad regular, validez regular.

C. Prueba confiable, pero sin validez.

D. Prueba muy confiable, pero sin validez.

E. Prueba confiable y con validez.

En el inciso A las marcas están muy dispersas en una prueba que no es confiable, porque no se concentran cerca del mismo sitio cada vez, ni dan en el centro (que es lo que se acerca más a lo que se desea evaluar). En el inciso B las marcas demuestran regular confiabilidad, pues no están tan dispersas, y validez regular porque, aunque algunas sí están en el centro, otras no. El inciso C demuestra una prueba confiable, pero con poca validez, ya que las marcas se concentran alrededor del mismo sitio, pero no en el centro, al igual que en el inciso D, aunque en este último están más cercanas entre sí. Finalmente, el inciso E demuestra lo deseable en cuanto a confiabilidad y validez: todas las marcas están concentradas en el centro.

**¿Cómo se mide la confiabilidad?** Los coeficientes de confiabilidad expresan la relación entre los puntajes que obtiene el mismo estudiante cuando presenta el mismo examen en dos ocasiones diferentes o en dos partes del mismo examen (Downing, 2004; Fraenkel et al., 2019). Esta relación da una idea de cuánta variación se puede esperar entre las diferentes

ocasiones, de persona a persona o de muestra de ítems a muestra de ítems. A continuación se describen algunas fórmulas para obtenerlos y adquirir información acerca de la consistencia interna de la prueba:

- *Kuder-Richardson*. Se utiliza para pruebas con variables dicotómicas (correcto/incorrecto), y otorga información sobre qué tan bien la prueba mide el constructo de interés (Kuder y Richardson, 1937). para conocer el valor de este coeficiente se pueden utilizar dos fórmulas:
  - KR20 – para pruebas con ítems con dificultad variable.

$$KR20 = \left(\frac{n}{n-1}\right)\left(\frac{1-\sum p \cdot q}{var}\right)$$

n= número de ítems

p= proporción de personas que aprueban el ítem

q= proporción de personas que reprueban el ítem

var= varianza para la prueba

  - KR21 – para pruebas con ítems con la misma dificultad.

$$KR21 = \left(\frac{n}{n-1}\right)\left(1-\frac{M(n-M)}{n \cdot var}\right)$$

n= número de ítems

M= media de la puntuación para la prueba

var= varianza de la prueba.

- *Alfa de Cronbach*. Es conocido como el coeficiente alfa, y se utiliza para pruebas con ítems con más de dos respuestas correctas, por ejemplo, preguntas con respuestas en forma de escala de valoración tipo Likert (Tavakol y Dennick, 2011).

$$\alpha = \frac{N\bar{c}}{\bar{v} + (N-1)\bar{c}}$$

$\propto$= alfa de Cronbach

N= número de ítems

$\bar{c}$= covarianza promedio inter-ítem

$\bar{v}$= varianza promedio

Los valores de los coeficientes de confiabilidad van de 0.00 a 1.00, donde 0 corresponde a una ausencia completa de relación y 1.00 el máximo posible de la relación (Fraenkel et al., 2019).

En el caso de pruebas de mediano impacto, si el valor es menor a 0.5, la consistencia interna es no aceptable; si es de 0.5 a 0.59, es pobre; si es de 0.6 a 0.69, es cuestionable; si es de 0.7 a 0.79, es aceptable; si es de 0.8 a 0.89, es buena; y mayor a 0.9 es excelente. Si la prueba es de alto impacto, el valor mínimo deseado es de 0.8 (Bland y Altman, 1997; Downing, 2004).

## CONCLUSIONES

La validez es uno de los conceptos más importantes en evaluación del y para el aprendizaje. El modelo actual de validez es el resultado de una gran cantidad de investigaciones y discusiones entre expertos en el tema, y se le considera como un concepto holístico que se alimenta de diversas fuentes (contenido, proceso de respuesta, estructura interna, relación con otras variables y consecuencias). Validez se refiere a las inferencias que pueden hacerse con los resultados, más que a las pruebas o exámenes *per se*. Durante el proceso de evaluación deben cuidarse las amenazas a la validez, tanto de subrepresentación como de varianza irrelevante al constructo. La confiabilidad o reproducibilidad de los resultados, se integra como un elemento de fuente de evidencia de validez en el rubro de estructura interna.

## REFERENCIAS


Abad, F. J., Olea, J., y Ponsoda, V. (2001). Analysis of the optimum number alternatives from the Item Response Theory. *Psicothema*, *13*(1), 152–158.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (Original w). American Educational Research Association.

Bland, J., y Altman, D. (1997). Statistics notes: Cronbach's alpha.pdf. *British Medical Journal*, *314*, 572.

Bond, L. (2008). Teaching to the Test: Coaching or Corruption. *New Educator*, *4*(3), 216–223. https://doi.org/10.1080/15476880802234482

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105.

Carrillo-Avalos, B. A., Sánchez-Mendiola, M. y Leenen, I. (2020). El concepto moderno de validez y su uso en educación médica. *Revista de Investigación en Educación Médica*, *9*(33), 98–106. https://doi.org/https://doi.org/10.22201/facmed.20075057e.2020.33.19216

Carrillo-Avalos, B. A., Sánchez-Mendiola, M., y Leenen, I. (2020). Amenazas a la validez en evaluación: implicaciones en educación médica. *Investigación en Educación Médica*, *9*(34), 100–107. https://doi.org/10.22201/facmed.20075057e.2020.34.221

Chiavaroli, N. (2017). Negatively-worded multiple choice questions: An avoidable threat to validity. *Practical Assessment, Research and Evaluation*, *22*(3), 1–14.

Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into Practice*, *48*(1), 63–71. https://doi.org/10.1080/00405840802577627

Cook, D. A., y Hatala, R. (2016). Validation of educational assessments: a primer for simulation and beyond. *Advances in Simulation*, *1*(1), 1–12. https://doi.org/10.1186/s41077-016-0033-y

Cronbach, L. J., y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

de Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, *44*(1), 109–117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. Medical education, *38*(9), 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Downing, S. M., y Haladyna, T. M. (2004). Validity threats: Overcoming interference with proposed interpretations of assessment data. *Medical Education*, *38*(3), 327–333. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to Design and Evaluate Research in Education* (10th ed.). Mc Graw-Hill Education.

Haladyna, T. M., y Downing, S. M. (2004). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice*, *23*(1), 17–27. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.2004.tb00149.x

Haladyna, T. M., Downing, S. M., y Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–334. https://doi.org/10.1207/s15324818ame1503_5

Haladyna, T. M., Rodriguez, M. C., & Stevens, C. (2019). Are Multiple-choice Items Too Fat? *Applied Measurement in Education*, *32*(4), 350–364. https://doi.org/10.1080/08957347.2019.1660348

Hicks, N. A. (2011). Guidelines for identifying and revising culturally biased multiple-choice nursing examination items. *Nurse Educator*, *36*(6), 266–270. https://doi.org/10.1097/NNE.0b013e3182333fd2

Instituto Nacional para la Evaluación de la Educación. (2017). Criterios técnicos para el desarrollo, uso y mantenimiento de instrumentos de evaluación. En *Diario Oficial de la Federación*. https://www.inee.edu.mx/wp-content/uploads/2019/04/P1E104.pdf

Jurado-Núñez, A., y Leenen, I. (2016). Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. *Investigación en Educación Médica*, *5*(17), 55–63. https://doi.org/10.1016/j.riem.2015.07.004

Kane, M. T. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *Validity: Revisions, New Directions and Applications* (pp. 39–64). Information Age Publishing, Inc.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160.

Lane, S. (2014). Evidencia de validez basada en las consecuencias del uso del test. *Psicothema*, *26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Leenen, I. (2014). Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. *Investigación en Educación Médica*, *3*(9), 40–55.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13–103). MacMillan. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x

Moreno, R., Martínez, R. J., y Muñiz, J. (2004). Directrices para la construcción de ítems de elección múltiple. *Psicothema*, *16*(3), 490–497. https://www.redalyc.org/articulo.oa?id=72716324

National Board of Medical Examiners. (2016). *Cómo elaborar preguntas para evaluaciones escritas en las áreas de ciencias básicas y clínicas*. *3*, 1–98.

Norcini, J. J. (2003). Setting standards on educational tests. *Medical Education*, *37*(5), 464–469. https://doi.org/10.1046/j.1365-2923.2003.01495.x

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Rios, J., y Wells, C. (2014). Evidencia de validez basada en la estructura interna. *Psicothema*, *26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. *Medical Teacher*, *33*(10), 783–797. https://doi.org/10.3109/0142159X.2011.611022

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26*(1), 100–107. https://doi.org/10.7334/psicothema2013.256

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd