



El juicio de expertos en la creación de instrumentos: una revisión narrativa

Recibido: 15 de noviembre de 2024
Evaluado: 25 de marzo de 2025
Publicado: 01 de octubre de 2026

Daniela Zúñiga* 
Claudio Bustos-Navarrete** 
Gamal Cerda*** 

Tipología: artículo de revisión

Resumen

Este artículo realiza una revisión del uso del juicio de expertos en la evaluación de la validez de contenido de instrumentos de medición en ciencias sociales. El proceso implica la reunión de un grupo de expertos en el tema de estudio, así como en metodología de la investigación, para que emitan un juicio calificado sobre aspectos específicos del instrumento analizado (Cabero y Llorente, 2013). El propósito principal de esta investigación es analizar y sintetizar la literatura disponible sobre el proceso de validación por juicio de expertos. Utilizando una metodología cualitativa, se sintetizan los resultados de 80 documentos, incluyendo artículos en revistas científicas, resúmenes, opiniones y cartas al editor. Los resultados revelan diversas posturas en cuanto a las etapas, la selección, los procedimientos, los informes y los tipos de análisis en el juicio de expertos. Estos se discuten en detalle, resaltando su relevancia en el contexto del juicio de expertos en ciencias sociales.

Palabras clave

investigación cualitativa; investigación pedagógica; cuestionario; validez de constructo; validez de contenido

* Doctora. Universidad de Concepción. danizuniga@udec.cl
** Doctor. Universidad de Concepción. clbustos@udec.cl
*** Doctor. Universidad de Concepción. gacerda@udec.cl

Sustainable Consciousness: A Thematic, Disciplinary, Methodological Analysis and the Fundamental Role of Education

Abstract

This article examines the use of expert judgment in assessing the content validity of measurement instruments in social sciences. The process involves convening a group of experts in the subject matter and research methodology to provide qualified opinions on specific aspects of the analyzed instrument (Cabero & Llorente, 2013). The primary aim of this research is to analyze and synthesize the available literature on the validation process through expert judgment. Using a qualitative methodology, the results from 80 documents, including scientific journal articles, abstracts, opinions, and letters to the editor, are synthesized. The findings reveal diverse perspectives regarding the stages, selection, procedures, reporting, and types of analysis in expert judgment. These are discussed in detail, highlighting their relevance in the context of expert judgment in social sciences.

Keywords

qualitative research; educational research; questionnaire; construct validity; content validity

O julgamento de especialistas na criação de instrumentos: uma revisão narrativa

Resumo

Este artigo analisa o uso do julgamento de especialistas na avaliação da validade de conteúdo de instrumentos de medição nas ciências sociais. O processo envolve a reunião de um grupo de especialistas no tema de estudo, bem como em metodologia de pesquisa, para emitir um julgamento qualificado sobre aspectos específicos do instrumento analisado (Cabero & Llorente, 2013). O principal objetivo desta pesquisa é analisar e sintetizar a literatura disponível sobre o processo de validação por julgamento de especialistas. Utilizando uma metodologia qualitativa, sintetizam-se os resultados de 80 documentos, incluindo artigos em revistas científicas, resumos, opiniões e cartas ao editor. Os resultados revelam diversas posições em relação às etapas, seleção, procedimentos, relatórios e tipos de análises no julgamento de especialistas. Esses aspectos são discutidos em detalhe, destacando sua relevância no contexto do julgamento de especialistas nas ciências sociais.

Palavras-chave

pesquisa qualitativa; pesquisa pedagógica; questionário; validade de construto; validade de conteúdo

Para citar este artículo:

Zúñiga, D., Bustos-Navarrete, C. y Cerda, G. (2026). El juicio de expertos en la creación de instrumentos: una revisión narrativa. *Revista Colombiana de Educación*, (101), e22397, <https://doi.org/10.17227/rce.num101-22397>

Introducción

La creación y validación de instrumentos¹ de medición son procesos esenciales en la investigación, ya que la calidad de los datos recopilados depende, en gran medida, de la solidez de las herramientas utilizadas. La validez de una escala es clave para respaldar científicamente las interpretaciones e inferencias derivadas de sus puntuaciones en el contexto de su uso propuesto (American Psychological Association, 1974). La validez es un proceso en constante evolución (American Educational Research Association, 2018), y puede presentarse de diversas formas según los objetivos, la población, las condiciones de administración y los métodos de evaluación utilizados (Aiken, 2003).

Uno de los tipos de evidencia más relevantes es la validez de contenido, que se enfoca en determinar en qué medida el contenido del instrumento se relaciona con el constructo que pretende medir (Martínez-Valdés y Juárez-Hernández, 2019). Se asume que las respuestas a los ítems² de un instrumento bien diseñado son indicativas de lo que serían las respuestas al universo completo de conductas de interés o del fenómeno de estudio (Aiken, 2003; Streiner *et al.*, 2015), por ende, la validez de contenido, refiere al grado en que los ítems o preguntas de un instrumento de medición representan de manera adecuada y completa el constructo que se desea evaluar. Es crucial garantizar que el instrumento constituya una muestra adecuada y representativa de los contenidos que se pretende evaluar (Muñiz, 2003). Esto implica asegurar que los ítems sean representativos de las dimensiones e indicadores pertinentes al objeto de estudio (González-Zamar *et al.*, 2021). Diferentes enfoques y dimensiones pueden requerir adaptaciones específicas para cada contexto (Escobar-Pérez y Cuervo-Martínez, 2008).

El juicio de expertos es uno de los métodos o pasos necesarios para evaluar la validez de contenido (Guillot-Valdés *et al.*, 2022), siendo el proceso principal en la mayoría de los desarrollos de instrumentos (Fernández-Gómez *et al.*, 2020). También, se pueden practicar pasos previos, como la denominada validez racional, que es la construcción de un instrumento derivada de la revisión de literatura; y cuando ella resulta insuficiente, se puede optar por la validez de respuesta, que implica el uso de entrevistas cualitativas para identificar los elementos que pueden constituir un instrumento. El juicio de expertos permite

¹ Entiéndase por instrumento como el espectro de herramientas, *test* o baterías utilizadas para la recolección de datos en una investigación. Estos instrumentos pueden incluir cuestionarios, encuestas, entrevistas estructuradas, escalas de evaluación, y cualquier otra herramienta diseñada para medir variables específicas o recolectar información relevante para el estudio.

² Entiéndase por ítems como las preguntas, observaciones, indicadores, afirmaciones o cualquier otra forma de enunciado que compone un instrumento. Estos ítems son diseñados para capturar datos específicos y proporcionar información detallada sobre las variables o constructos que se están evaluando.

emitir una evaluación a cualquiera de estos dos procesos previos que se hayan realizado (Maldonado-Suárez y Santoyo-Telles, 2024). De hecho, es poco frecuente encontrar métodos alternativos que sean tan efectivos para verificar la validez de contenido (Ding y Hershberger, 2002), y, en la mayoría de los casos, el juicio de expertos se convierte en el indicador principal del proceso de validación de contenido (Fernández-Gómez *et al.*, 2020). En consecuencia, se ha consolidado como uno de los métodos más comunes y recomendados para llevar a cabo en esta etapa crucial del desarrollo del instrumento (Bowen-Mendoza *et al.*, 2022; Ding y Hershberger, 2002; Teistler, 2021).

De acuerdo con Souto-Seijo *et al.* (2020), el juicio de expertos se configura como una de las estrategias más ampliamente empleadas en el ámbito educativo y psicológico (Escobar-Pérez y Cuervo-Martínez, 2008). Esto se debe a que su metodología se basa en la recopilación de juicios calificados emitidos por un grupo de expertos sobre diversos aspectos del instrumento (Cabero y Llorente, 2013). Es esencial que los investigadores definan rigurosamente el constructo, dimensiones e ítems, así como el formato del instrumento antes de someterlo a la evaluación de los expertos (Almanasreh *et al.*, 2019), procedimiento que se puede apoyar en la validez racional o de respuesta. Los expertos deben llevar a cabo un análisis minucioso y emitir juicios sobre la correspondencia entre los elementos y los contenidos evaluados (Pedrosa *et al.*, 2014). La validación por expertos permite obtener información exhaustiva y detallada sobre el objeto de estudio (Luque-Vara *et al.*, 2020), y se destaca por la alta calidad de las respuestas (Fernández-Gómez *et al.*, 2020). Todo ello resalta la relevancia de este proceso dentro del desarrollo de instrumentos de medición en ciencias sociales. No obstante, la diversidad de enfoques y hallazgos respecto a su implementación plantea un desafío importante, ya que la falta de consensos puede dificultar la adecuada comprensión y aplicación del juicio de expertos en la validación de instrumentos.

Es por lo que el objetivo de esta investigación es analizar y sintetizar, a través de una revisión narrativa, el proceso de validación por juicio de expertos en la creación de instrumentos de medición en el contexto de investigaciones en ciencias sociales. Esta revisión permitirá evidenciar los usos y pasos más frecuentes en este proceso, contribuyendo a una toma de decisiones más informada y rigurosa en futuras investigaciones.

Metodología

Diseño de investigación

Esta investigación adopta un diseño de revisión narrativa, un enfoque amplio que facilita la descripción y discusión del estado del arte en un campo específico (Terezinha, 2007). Este estudio se centra en el proceso de juicio de expertos, con el objetivo de proporcionar una actualización sobre las investigaciones que se han llevado a cabo para evaluar la validez de contenido mediante este método.

Criterios de búsqueda y selección

Se definieron términos clave para guiar los criterios de búsqueda en bases de datos online: *"Juicio de expertos/expert judgement"*, *"investigación metodológica/methodological research"*, *"instrumento/instrument"*, *"medición/measurement"*, *"construcción de instrumentos/instrument development"*, *"validez/validity"*, *"validez de contenido/content validity"*, *"validación de instrumentos/validation of instruments"*, *"Education, Scientific Disciplines"*, *"Education y Educational Research"*, *"Psychology, Educational"*, *"sociology"* y *"social sciences and humanities"*.

La búsqueda de documentos se llevó a cabo en dos etapas: la primera entre junio y julio de 2022 y la segunda en marzo 2025. Se buscaron documentos que incluyeran los términos clave en los títulos, resúmenes y palabras clave en Web of Science (WOS), Scopus, Scielo, SAGE Journal, EBSCOhost, entre otras bases de datos.

En la primera etapa, se definió un rango de publicación desde el 2017 hasta el 2022, considerando que un período de al menos cinco años permite capturar tanto estudios recientes como antecedentes relevantes en la evolución del tema. Posteriormente, en la segunda etapa, se amplió la búsqueda para incluir publicaciones desde 2023 hasta 2024, con el objetivo de incorporar avances metodológicos más recientes en la temática. También se realizaron búsquedas específicas que cubren períodos anteriores para proporcionar un contexto más amplio. Las áreas de investigación consultadas variaron según la base de datos pero, en términos generales, se enfocaron en ciencias sociales, psicología y educación. Además, debido a la naturaleza del tema de investigación, se incluyeron estudios en áreas como investigación médica y ciencias de la salud.

Para la selección de documentos, se incluyeron artículos de investigación primarios y secundarios, artículos de opinión, cartas al editor y resúmenes. También se consideraron documentos escritos en inglés o español, así como estudios con metodologías cualitativas, cuantitativas o mixtas. Por otro lado, se

excluyeron los libros y capítulos de libros, las publicaciones en otros idiomas y los estudios que no se ajustaran a las metodologías mencionadas.

Se revisaron los títulos y resúmenes basándose en los criterios de inclusión y exclusión mencionados. Los documentos obtenidos en la búsqueda formal se evaluaron con un “Sí” para avanzar a la lectura completa o un “No” para excluirllos de la revisión. Finalmente, se incorporaron 80 documentos en total, de los cuales 51 provinieron de la búsqueda formal y 29 de la búsqueda informal.

Caracterización de las fuentes consultadas

De los ochenta documentos seleccionados para esta revisión, estos se distribuyen en distintas revistas de investigación. Clasificadas revistas del área de educación/psicología educativa; revistas de investigación médica y las ciencias de la salud; revistas abarcan la informática y la tecnología educativa; revistas especializadas en métodos de investigación y estadística; entre otras.

En cuanto a los años de publicación de los estudios seleccionados, se observa una mayor concentración de publicaciones en el periodo comprendido entre 2024 y 2022, representando casi el 50 % de los artículos en este intervalo de tiempo. En la búsqueda, se agregaron algunos artículos hasta 1990, ya que las publicaciones son pioneras en análisis del el índice de validez de contenido (Aiken, 1985; Cohen, 1968; Fleiss, 1971; Lynn, 1986).

Organización y análisis de la información

Para organizar y analizar la información mediante revisión narrativa, se utilizó la técnica de metasíntesis. Esta metodología cualitativa integra y sintetiza un gran volumen de datos, facilitando la identificación de dominios para la presentación de la información. A través de la categorización, se consolidaron las interpretaciones sobre los criterios y métodos utilizados para evaluar la validez del contenido.

Presentación de resultados

Etapas del proceso de validación por juicio de expertos

El proceso de juicio de expertos implica varias etapas fundamentales para la validación de un instrumento. A continuación, se detallan las etapas recomendadas:

Primero: definición de objetivos y selección de expertos

Los investigadores deben definir los objetivos del juicio de expertos y establecer criterios de reclutamiento para garantizar la validez del proceso. Para ello, es esencial contar con la versión final del instrumento y diseñar una pauta de evaluación con instrucciones claras para los expertos. Una vez definidos estos aspectos, se procede a la selección de los expertos según los criterios establecidos (Escobar-Pérez y Cuervo-Martínez, 2008; Maldonado-Suárez y Santoyo-Telles, 2024). Se recomienda incluir un especialista en lingüística para optimizar la redacción y coherencia del instrumento (Escobar-Pérez y Cuervo-Martínez, 2008).

Segundo: contacto con expertos para realizar evaluación

Se contacta a los expertos con una carta formal que detalla, como mínimo, el objetivo del instrumento, el contexto de estudio, la población y el lugar de aplicación. Este paso es crucial para que los expertos puedan situar y evaluar adecuadamente los ítems del instrumento. También se solicitan antecedentes personales para crear perfiles de los expertos (Escobar-Pérez y Cuervo-Martínez, 2008; Galicia *et al.*, 2017).

Tercero: proceso de evaluación del objeto de estudio

Los expertos emiten su juicio sobre el instrumento utilizando una pauta de evaluación que detalla todos los aspectos a evaluar, las instrucciones, definiciones de dominios, ítems y sus puntajes correspondientes (Guillot-Valdés *et al.*, 2022; Teistler, 2021).

Cuarto: recepción y análisis de la información

Se recopila y analiza la información proporcionada por los expertos. Se realiza un análisis cuantitativo o cualitativo y se elaboran conclusiones. Esta información es clave para la descripción psicométrica de la prueba. Se recomienda una nueva ronda de evaluación para revisar y mejorar el instrumento (Almanasreh *et al.*, 2019).

Proceso de selección de los expertos

El proceso de selección de los expertos requiere cuidadosa consideración. Los expertos deben tener especialización y experiencia en áreas relacionadas con la investigación para evaluar adecuadamente el contenido y la forma de los ítems

(Maldonado-Suárez y Santoyo-Telles, 2024; Sandoval-Henríquez y Badilla-Quintana, 2021; Soriano-Rodríguez, 2015). Aunque muchos estudios no describen explícitamente sus criterios de selección, algunos autores sugieren considerar afinidad académica/laboral y cercanía (Robles y Rojas, 2015), utilizando frecuentemente el muestreo no probabilístico por conveniencia (Fernández-Gómez *et al.*, 2020).

Es importante considerar algunos criterios que incluyen la trayectoria, tanto laboral como académica, años de experiencia, formación y participación previa en investigaciones o actividades formativas; también familiaridad de los expertos con el tema de estudio y la población a la que se dirige la prueba (Arnaiz-Sánchez *et al.*, 2019; Cabero e Infante, 2014; Luque-Vara *et al.*, 2020; Robles y Rojas, 2015).

En el proceso de selección de los participantes, es fundamental considerar su disponibilidad y motivación, ya que estos factores pueden influir en su compromiso y capacidad para llevar a cabo el juicio de expertos de manera efectiva (Cabero e Infante, 2014), del mismo modo, deben tener conocimientos actualizados sobre la materia e investigaciones similares (Herrera-Masó *et al.*, 2023). Además, es crucial prestar atención a la forma de contacto inicial, ya sea a través de correo electrónico, carta formal o relación directa con los investigadores. Establecer una relación que vincule el proceso con los expertos es esencial para garantizar una colaboración efectiva y sostenida a lo largo del estudio.

Cabero e Infante (2014) sugieren complementar procesos cualitativos y cuantitativos para la selección de expertos. El proceso cualitativo, llamado biograma del experto, esquematiza elementos sobre experiencia profesional y producción científica. El proceso cuantitativo, llamado coeficiente de competencia experta, se calcula con la fórmula (Cabero-Almenara *et al.*, 2020; Marín-González *et al.*, 2021; Mengual, 2011):

$$K = \frac{K_c + K_a}{2}$$

El cálculo del coeficiente de competencia experta de acuerdo a Cabero e Infante (2014) se realiza a partir de la opinión evidenciada por el experto sobre:

1. Primero K_c (Coeficiente de conocimiento), aquí el experto valora su propio conocimiento sobre el tema de estudio. En un rango de 0 a 10, donde cero es nulo conocimiento, mientras que 10 es pleno conocimiento sobre el tema de estudio. El valor obtenido se multiplica por 0.1.

2. Segundo K_a (Coeficiente de argumentación), aquí el experto valora su grado de influencia en diferentes áreas de acuerdo con el tema de estudio. Este valor se obtiene a través de una tabla con seis criterios de argumentación adaptada de Dobrov y Smirnov en 1972. El experto debe autoevaluarse (tabla 1), marcando la opción que lo represente de acuerdo con el tema de estudio. De acuerdo con Cruz y Martínez (2018) los valores de la tabla 1 deben permanecer ocultos para el experto.

Tabla 1.

Tabla de valoración de fuentes de la argumentación adaptada de Dobrov y Smirnov (1972)

Fuente de argumentación	Grado de influencia de las competencias		
	Alto	Medio	Bajo
1.- Análisis teórico realizado por usted.	.30	.20	.10
2.- Experiencia práctica.	.50	.40	.20
3.- Generalización de trabajos de autores nacionales.	.05	.05	.05
4.- Generalización de trabajo de autores extranjeros.	.05	.05	.05
5.- Su propio conocimiento del estado del arte a nivel mundial.	.05	.05	.05
6.- Su intuición.	.05	.05	.05

Fuente: Cruz y Martínez (2018).

Número de expertos para el proceso de validación

La determinación del número de expertos necesario para el proceso de validación considera diversos factores para la elección (Robles y Rojas, 2015). Puede incluir la accesibilidad a los expertos y la disponibilidad de personas con conocimiento suficiente en la temática de la investigación (Cabero y Llorente, 2013).

Aunque no hay un estándar común, han propuesto rangos de cantidad de expertos considerados ideales. Hasta 10 expertos (Hyrkäs *et al.*, 2003), un mínimo de 5 expertos (Anggraeni y Mundilarto, 2020), y otros rangos como 3 a 20 expertos (Lynn, 1986), 2 a 20 expertos (Escobar-Pérez y Cuervo-Martínez, 2008), 9 a 24 expertos (Zartha *et al.*, 2017), 5 a 10 expertos (Almanasreh *et al.*, 2019), o 15 a 25 expertos (Bernal-García *et al.*, 2020).

Es importante destacar que la elección del número de expertos debe basarse en la naturaleza de la investigación y las condiciones disponibles, considerando la complejidad del constructo a evaluar y la diversidad de perspectivas necesaria para asegurar una validación adecuada.

Procedimiento para realizar el proceso de juicio con expertos

El proceso también implica considerar diversos aspectos para la ejecución exitosa. La figura 1 presenta un esquema del proceso de decisión en la ejecución del juicio de expertos (Escobar-Pérez y Cuervo-Martínez, 2008; Robles y Rojas, 2015).

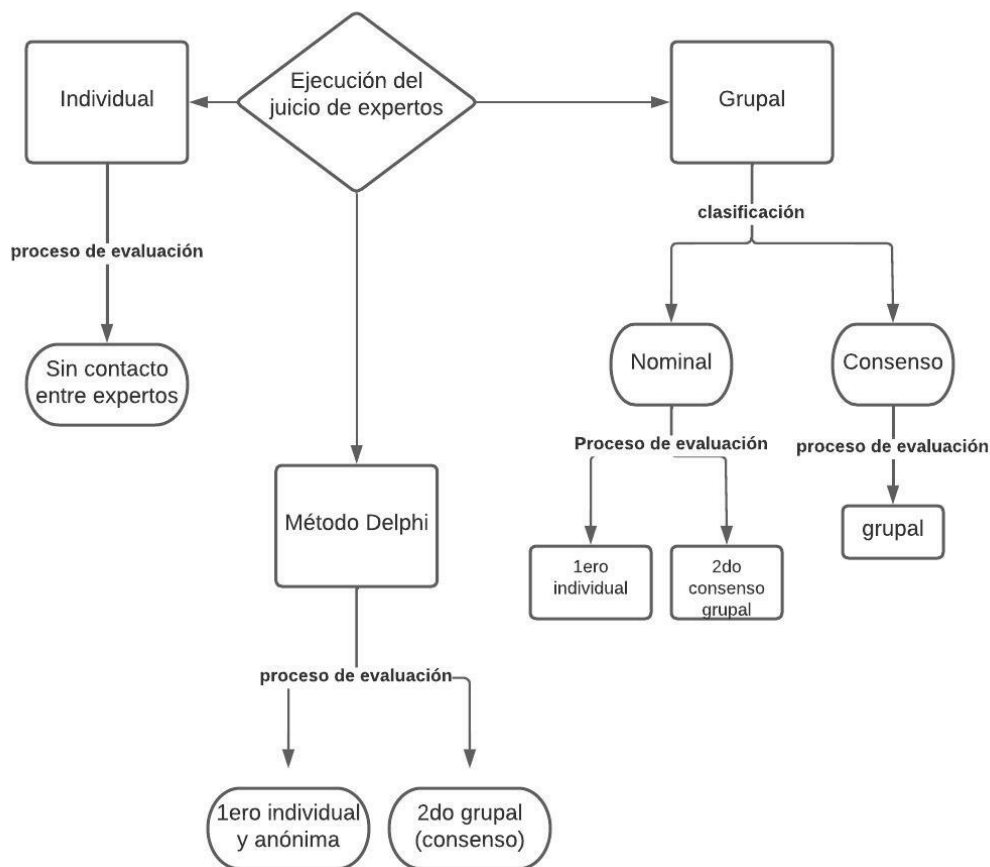


Figura 1.

Esquema de opciones de proceso ejecución para el juicio de expertos.

Fuente: elaboración propia.

En la modalidad individual, la información se obtiene de manera anónima y confidencial, sin contacto directo entre los participantes (Cabero y Llorente, 2013). El investigador es el encargado de reunir y analizar las evaluaciones emitidas por los expertos.

En cuanto a la modalidad grupal, existen dos enfoques. El primero, llamado nominal, incluye una etapa individual en la que las respuestas pueden o no ser anónimas, seguida de una etapa de consenso grupal. La segunda modalidad,

denominada consenso, los expertos trabajan grupalmente durante todo el proceso. Estas diferentes modalidades se deben adoptar de acuerdo a las necesidades específicas de la investigación y al contexto en el que se lleva a cabo.

Por otro lado, en el método *Delphi*, es un proceso estructurado de comunicación, se basa en el juicio colectivo de personas con conocimientos especializados o expertos en el área de interés (Lin *et al.*, 2015). Es un método que, si bien tiene diferentes formas de aplicación, sus elementos clave de acuerdo con Jorm (2015) son:

- Un facilitador organiza el estudio.
- Se recluta a un grupo de individuos con experiencia en el tema de estudio.
- Se compila un cuestionario con afirmaciones que los expertos califican.
- Se recopilan las respuestas utilizando el cuestionario.
- Se proporciona retroalimentación anónima sobre cómo sus respuestas se comparan con las del resto del grupo.
- Los expertos pueden revisar sus respuestas después de recibir retroalimentación.

El método *Delphi* es un método de consenso de expertos depende en gran medida de la evidencia en la que se basa dicho consenso, ya que pueden incluir revisiones sistemáticas, experimentos individuales, estudios cualitativos, experiencias personales, entre otras (Manley, 2013). Su dualidad metodológica permite a los participantes individuales expresar sus opiniones con un alto grado de libertad sobre un problema.

El juicio de expertos nominal y el método *Delphi* son técnicas similares, pero presentan sus diferencias. En el juicio nominal, las etapas iniciales pueden ser individuales y anónimas, culminando en un consenso grupal en el que los expertos interactúan directamente. En contraste, el método *Delphi* mantiene el anonimato y separación durante todo el proceso, utilizando múltiples rondas de cuestionarios para refinar el consenso sin confrontación directa.

Estructura del informe para los expertos

El proceso de juicio de expertos puede emplear metodologías y análisis tanto cualitativos como cuantitativos. En el enfoque cuantitativo se utilizan cuestionarios, matrices de evaluación, listas de chequeo, entre otras herramientas similares (Cabero y Llorente, 2013). En el enfoque cualitativo, las preguntas abiertas son comunes para obtener la opinión de los expertos sobre varios aspectos del instrumento. Se recomienda considerar ambos tipos de datos para

analizar la información en su totalidad, identificar fortalezas y debilidades, y realizar modificaciones necesarias (Luque-Vara *et al.*, 2020).

Las preguntas cualitativas a los expertos pueden incluir comentarios y sugerencias sobre el instrumento creado (Anggraeni y Mundilarto, 2020; Carrasco-Sáez *et al.*, 2019; Guillot-Valdés *et al.*, 2022), mejoras relacionadas con el título, las instrucciones o la creación de nuevos ítems (Arnaiz-Sánchez *et al.*, 2019), y aspectos del instrumento que los investigadores no consultaron previamente (Sandoval-Henríquez y Badilla-Quintana, 2021). También se incluyen solicitudes de observaciones, matizaciones, modificaciones y anulaciones de ítems (González-Zamar *et al.*, 2021).

En el proceso cuantitativo, los tipos de indicadores o criterios de evaluación son diversos. Bernal-García *et al.* (2020) destaca cinco aspectos frecuentemente utilizados en investigaciones, estos son descritos por Garrote y del Carmen-Rojas (2015):

—Suficiencia: grado en que los ítems que pertenecen a una misma dimensión bastan para medirla.

—Claridad: grado en que el ítem es comprensible y sin ambigüedades.

—Coherencia: grado en que el ítem tiene relación lógica con la dimensión que está midiendo.

—Relevancia: grado en que el ítem importante, es decir que debe ser incluido en el instrumento.

—Pertinencia: grado de relación del ítem con el constructo evaluado.

Los estudios de Bowen-Mendoza *et al.* (2022), Sarah *et al.* (2019) y Vínces (2018) coinciden en utilizar los cuatro primeros criterios.

Valoraciones de respuesta

En el proceso de juicio de expertos, se emplean diversos métodos para valorar las respuestas, entre ellos escalas dicotómicas y escalas tipo Likert. En el caso de estas últimas, las investigaciones difieren en el número de alternativas de estas escalas, comúnmente utilizando hasta 5 alternativas (Pedrosa *et al.*, 2014) o entre 5 y 7 alternativas (Penfield y Giacobbi, 2004). El propósito fundamental de este proceso es determinar en qué medida un ítem se ajusta al constructo de interés. Para ello, se busca obtener puntuaciones bajas para ítems de relevancia pobre y altas para aquellos con relevancia muy buena. La concordancia entre expertos es esencial para la evaluación efectiva del instrumento.

Tipos de análisis para respuestas

En su estudio, Almanasreh *et al.* (2019) señala que no existe un método único para establecer la validez de contenido, ni un enfoque definitivo para su análisis y determinación de resultados. A continuación, se presenta un resumen de los procedimientos analíticos propuestos por Pedrosa y colaboradores en el 2014 (tabla 2).

Tabla 2.

Síntesis de métodos para realizar en el juicio de expertos

Año	Autores	Método de análisis
1961	Tucker	Basado en análisis factorial
1975	Lawshe	Relación de validez de contenido (CVR)
1977	Rovinelli y Hambleton	Índice de congruencia ítem-objetivo
1980	Aiken	V de Aiken
1980, 1984	Hambleton	Índice de congruencia
1986	Lynn	El Índice de validez de contenido (IVC)
1992	Sireci y Geisienger	Escalamiento multidimensional y análisis de <i>clusters</i>
1997	Levine, <i>et al.</i>	Minimum qualifications (MQ)
2001	Fitch, <i>et al.</i>	Interpercentile range adjusted for symmetry (IPRAS)
2002	Hernández-Nieto	Coficiente de validez de contenido (CVC)
2003	Rubio, <i>et al.</i>	Índice de validez factorial (FVI)
2012	Claeys, <i>et al.</i>	Average deviation mean (ADM)

Fuente: Pedrosa *et al.* (2014, p. 12).

A continuación, se describen las pruebas que algunos autores emplearon en sus investigaciones para llevar a cabo el análisis de la validez del instrumento mediante la evaluación de expertos.

Análisis para medir el nivel de acuerdo entre jueces

El proceso de determinar el porcentaje de concordancia entre las evaluaciones por ítems y por expertos se ha mantenido como una práctica común y ampliamente utilizada durante varios años (McHugh, 2012).

1. Porcentaje de coincidencia de las evaluaciones: la forma más sencilla de calcular el acuerdo entre jueces se realiza comparando las puntuaciones proporcionadas por múltiples expertos para cada ítem del instrumento, por lo general, se utiliza una escala específica (por ejemplo, tipo Likert). Luego, se calcula el porcentaje de acuerdo entre los expertos para cada ítem, determinando cuántos estuvieron de acuerdo en un criterio específico y expresando ese acuerdo como un porcentaje del total de expertos. Un alto porcentaje indica acuerdo, sugiriendo que el ítem se

percibe de manera consistente, mientras que un bajo porcentaje indica desacuerdo y la necesidad de revisar el ítem.

Una preocupación importante en el análisis del porcentaje de acuerdo absoluto es el grado en que el acuerdo se logra por mero azar, especialmente si existe un sesgo significativo tanto positivo como negativo entre los jueces (Banerjee *et al.*, 1999). Por tanto, se han propuesto indicadores que controlan el efecto del azar en la estimación del acuerdo, destacando el Kappa de Cohen para dos jueces y su extensión a múltiples jueces, el Kappa de Fleiss.

2. Kappa k de Cohen (tabla 3): el coeficiente Kappa mide la concordancia entre dos evaluadores en una escala nominal, lo que significa que se utilizan para clasificar categorías específicas. Su rango va desde -1 hasta +1, donde un valor cercano a +1 indica alto grado de concordancia, y un valor cercano a 0 sugiere que el acuerdo es similar al esperado por casualidad. Un valor negativo, como -1, indica un desacuerdo completo ((Abraira, 2000; Escobar-Pérez y Cuervo-Martínez, 2008). Aunque Kappa puede alcanzar valores negativos, estos son raros y generalmente indican que los evaluadores están en desacuerdo de una manera que es más sistemática y consistente de lo que se esperaría por pura casualidad. La fórmula para calcular Kappa es:

$$\kappa \text{ de Cohen} = \frac{\text{Proporción de acuerdos observados} - \text{Proporción de acuerdos esperados por azar}}{1 - \text{Proporción de acuerdos esperados por azar}}$$

En la tabla 3 se presenta la interpretación de los valores obtenidos.

Tabla 3.

Interpretación valores de la k de Cohen

<i>k de Cohen</i>	Nivel del acuerdo	% de datos que son fiables
<0- .20	Ninguno	0-4 %
.21- .39	Mínimo	4-15 %
.40- .59	Débil	15-35 %
.60- .79	Moderado	35-63 %
.80- .90	Fuerte	64-81 %
Sobre .90	Casi perfecto	82-100 %

Fuente: McHugh (2012, p. 279).

3. Kappa k Ponderado: de acuerdo con Cohen (1968), este coeficiente generaliza el Kappa de Cohen para incorporar grados de acuerdo y desacuerdo en escalas nominales, tratando todas las discrepancias por igual. Proporciona una medida de concordancia corregida

completamente por el azar. Los puntajes varían entre -1 y +1, donde valores cercanos a +1 indican alta concordancia, valores cercanos a 0 sugieren concordancia por azar, y valores negativos indican desacuerdo sistemático. Para la fórmula exacta del Kappa Ponderado y obtener mayor detalle, se recomienda consultar el trabajo original de Cohen (1968).

4. Kappa k de Fleiss: este coeficiente generaliza el Kappa de Cohen para medir el grado de acuerdo entre 3 o más evaluadores. Los valores κ de Fleiss oscilan entre -1 y +1 y se interpretan de acuerdo a la “fuerza” de acuerdo entre jueces (Fernández-Gómez et al., 2020). Valores cercanos a +1 indican alto acuerdo, mientras que valores cercanos a -1 indican mayor desacuerdo. Un valor de 0 indica que el acuerdo observado se debe únicamente al azar (Cerda y Villarroel, 2008). La fórmula exacta para calcular Kappa de Fleiss se detalla en el trabajo de Fleiss (1971).

El Kappa de Fleiss puede tener valores negativos, aunque raros, que indican desacuerdo sistemático entre los evaluadores. La tabla 4 presenta la propuesta de interpretación de los valores obtenidos.

Tabla 4.

Interpretación valores de la k de Fleiss

k de Fleiss	Nivel del acuerdo
<0	Pobre
0,1-0,20	Leve
0,21-0,40	Regular
0,41-0,60	Moderado
0,61-0,80	Sustancial
0,81-1,00	Casi perfecto

Fuente: Luque-Vara et al. (2020, p. 5).

Análisis para medir la calidad del Ítem en el Instrumento:

5. V de Aiken: este estadístico se utiliza para evaluar el nivel de validez de contenido de un ítem por varios jueces o de varios ítems por un solo juez. El coeficiente varía de 0 a 1, donde un mayor valor indica una mayor validez de contenido según el juicio del experto (Aiken, 1985). Por lo general, se espera que el intervalo de confianza inferior sea superior a 0,5. Para el caso de un ítem, la valoración r_i de cada juez i , es transformada a una puntuación $s_i = r_i - \text{mínimo}$. Si n es el número de jueces y c el número de categorías de la escala medición, V de Aiken es:

$$V = \frac{\sum_{i=1}^n s_i}{n(c-1)}$$

Para la valoración de varios ítems por un solo juez, la fórmula es análoga, siendo s_i la valoración del ítem i y n el número de ítems.

6. Índice de Validez de Contenido (IVC): este índice, evalúa cada ítem reflejando su importancia en el instrumento, con valores que varían entre +1 y -1, reflejando la importancia de cada ítem en el instrumento. La fórmula para calcular el IVC es:

$$IVC = \frac{\text{Número de ítems con(a) puntuaciones perfectas o(b) las dos últimas con mayor grado}}{\text{Número total de ítems del instrumento}}$$

Cuando se tiene menos de seis expertos, se recomienda utilizar el acuerdo y el muy acuerdo con las puntuaciones más altas (Almanasreh *et al.*, 2019; Bowen-Mendoza *et al.*, 2022).

7. Índice de Validez de Contenido General (IVCG): este índice se utiliza para asignar una ponderación a la validez de contenido. Según Polit *et al.* (2007), valores superiores a 0.80 sugieren que el instrumento es considerado válido. La fórmula para calcular el IVCG es:

$$IVCG = \frac{\Sigma IVC \text{ de cada experto}}{N \text{ de expertos}}$$

8. Relación de Validez de Contenido (CVR): el Coeficiente de Validez de Contenido (CVR) evalúa el nivel de acuerdo entre expertos sobre la esencialidad de un ítem específico. Los valores del CVR oscilan entre -1 y +1, y los expertos califican cada ítem como "esencial", "útil pero no esencial" o "no necesario" (Bowen-Mendoza *et al.*, 2022). La fórmula para calcular el CVR es:

$$CVR = \frac{N \text{ de expertos que indican el ítem como esencial} - N \text{ total de expertos}}{N \text{ total de expertos}}$$

Conclusiones

Esta revisión narrativa sobre la validez de contenido mediante el juicio de expertos compila hallazgos y metodologías asociadas con este proceso esencial en campos como las ciencias sociales. Se proporciona un resumen claro, reflejando la esencia del juicio de expertos, sus diversas aplicaciones y enfoques analíticos en la construcción de instrumentos de medición.

El análisis revela que, a pesar de las variaciones en su aplicación y análisis, el juicio de expertos sigue siendo fundamental y valorado para validar contenido al crear instrumentos de medición. Esta revisión destaca los aspectos más efectivos de su aplicación, se identifican prácticas y usos actuales, contribuyendo al desarrollo de instrumentos más precisos y confiables.

Los resultados exponen varias posturas sobre las etapas del proceso de juicio de expertos, recomendándose las siguientes: 1) Definir objetivos y seleccionar expertos. 2) Establecer contacto. 3) Evaluar el objeto de estudio. 4) Recibir y analizar la información. Estas metodologías pueden ser individuales, grupales o *Delphi*, se destaca de esta última la versatilidad, debido a su capacidad para abordar enfoques cualitativos y cuantitativos (Villiers *et al.*, 2005).

Para la etapa de reclutamiento, debe seleccionar cuidadosamente a los expertos, tomando en cuenta su especialización y experiencia en áreas relacionadas con la investigación. Los criterios de selección deben abarcar la trayectoria laboral y académica, la familiaridad con el tema de estudio, y la experiencia en investigaciones o actividades formativas relevantes (Robles y Rojas, 2015), se pueden utilizar métodos cualitativos para la selección del experto como el biograma o cuantitativos como el coeficiente de competencia experta (Cabero e Infante, 2014). Respecto a la determinación del número de expertos, no existe un consenso unánime entre investigadores, pero de acuerdo con una revisión de 38 estudios, en promedio considerar entre 6 y 10 expertos es una tendencia frecuente en las investigaciones (anexo 1).

Para la presentación del informe a los expertos, los investigadores pueden considerar tanto datos cualitativos como cuantitativos. En el enfoque cualitativo, se suelen usar preguntas abiertas para recabar la opinión de los expertos sobre distintos aspectos del instrumento. En el enfoque cuantitativo, se pueden emplear herramientas como cuestionarios, matrices de evaluación o listas de chequeo. En este último enfoque, los indicadores de evaluación varían, pero claridad, pertinencia, relevancia y coherencia fueron de los criterios más frecuentes en una revisión de 33 investigaciones (anexo 2). En cuanto a las escalas de respuesta, las escalas tipo Likert con respuestas de 1 a 5 son las más utilizadas en los 32 estudios revisados (anexo 3).

En lo que respecta al análisis de los resultados, los métodos varían, categorizándose principalmente en análisis enfocados en el nivel de acuerdo de los expertos y en la calidad del ítem evaluado. La calidad del ítem puede asegurar la relevancia, precisión y utilidad del instrumento, factores que determinan su efectividad en la medición de los constructos estudiados. Mientras que el acuerdo entre expertos puede proporcionar una perspectiva valiosa, pero no necesariamente garantiza la idoneidad del ítem para capturar los aspectos esenciales del fenómeno estudiado. Es por esto que, priorizar la calidad del ítem

sobre el consenso de los expertos puede contribuir a desarrollar instrumentos más robustos y confiables, capaces de proporcionar medidas más precisas y válidas de los constructos de interés.

Entre las limitaciones de esta revisión narrativa, se reconoce que otros tipos de investigación, como las revisiones sistemáticas y metaanálisis, pueden ofrecer mayor objetividad en la entrega de la información. Una limitación adicional en esta revisión es el enfoque en artículos que se asocian principalmente al área educativa. A pesar de que el juicio de expertos es un método ampliamente reconocido y utilizado en educación, sigue siendo una técnica relevante y válida en el ámbito de las ciencias sociales, lo que justifica las áreas de búsqueda seleccionadas.

Con base en las limitaciones identificadas, en las proyecciones futuras se espera ampliar el alcance del estudio, especialmente, al considerar investigaciones en psicología. Se subraya la necesidad de expandir la revisión a una gama más amplia de campos académicos, en la que el juicio de expertos pueda influir significativamente en el desarrollo y validación de estos instrumentos. Adicionalmente, se espera realizar una revisión sistemática sobre estudios que utilizan el juicio de expertos en el futuro. Esta revisión sistemática, a diferencia de la narrativa, adoptaría una metodología más estructurada y rigurosa en la selección y análisis de los estudios.

En resumen, esta revisión representa un aporte significativo al campo de las ciencias sociales, especialmente en lo que respecta a la construcción de instrumentos de medición en educación. Al proporcionar una panorámica detallada y actualizada del juicio de expertos, esta revisión se posiciona como un recurso esencial para los investigadores que utilicen este tipo de proceso de validez de contenidos en sus estudios. Esta parte, si bien constituye un insumo sustantivo en la configuración de la versión final del instrumento, necesariamente debe complementarse con otros procesos de validez a través de los análisis factoriales. Estos análisis permiten verificar si la elección de los ítems se respalda con el modelo inicial propuesto por el investigador.

Referencias

- Abraira, V. (2000). El índice kappa, notas técnicas. *Semergen*, 27, 247-249.
- Aceves-Villanueva, Y., Aviña-Camacho, I., Martínez-Soto, Y. y Ortega-Andrade, N. A. (2024). Validación de contenido de un modelo para evaluar la calidad de vida en jóvenes universitarios. *Revista Dilemas Contemporáneos*, 12 (edición especial).

- Acuña-Condori, S. P., Palacios-Garay, J. P., Astorayme-Dávalos, G. F., Mendoza-Suárez, C. E., y Gutiérrez-Álvarez de Arriaga, E. R. (2024). Formación continua y competencia digital de profesores de la REI 04, UGEL 02 *Horizontes*, 8(32), 57-66. <https://doi.org/10.33996/revistahorizontes.v8i32.704>
- Agudelo-Gallego, C. M., Montoya-Zuluaga, P. A., Cárdenas-Niño, L., Pino-Montoya, J. W. y Gómez-Tabares, A. S. (2024). Psychometric properties of the Aggression Questionnaire (AQ) in a sample of Colombian children and adolescents. *Suma Psicológica*, 31(2), 34-51. <https://doi.org/10.14349/sumapsi.2024.v31.n2.5>
- Aiken, L. (1985). Three coefficients for analyzing the reliability and validity of ratings, *Educational and Psychological Measurement*, 45, 131-142.
- Aiken, L. (2003). *Test psicológicos y evaluación (11.ª ed.)*. PEARSON Prentice Hall.
- Albán-Romero, L., Mendoza-Jara, M., López-Fernández, R., y Tapia-Bastidas, T. (2023). Recursos didácticos digitales en la presencialidad: dificultades en las buenas prácticas docente. *Mendive*, 21(4), e3632. <https://mendive.upr.edu.cu/index.php/MendiveUPR/article/view/3632>
- Aliaga-Pacora, A., Juárez-Hernández, L., y Herrera-Meza, R. (2021). Diseño y validez de contenido de una rúbrica analítica socioformativa para evaluar competencias investigativas en posgrado. *Apuntes Universitarios*, 11(2), 62-82.
- Almanasreh, E., Moles, R. y Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, 15(2), 214-221. <https://doi.org/10.1016/j.sapharm.2018.03.066>
- American Educational Research Association. (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). American Educational Research Association.
- American Psychological Association. (1974). *Standards for educational and Psychological tests*. American Psychological Association.
- Anggraeni, T. E. y Mundilarto, M. (2020). The Development of Local Wisdom-Based Physics Cognitive Ability Assessment Instrument for Senior High School Students. *Journal Penelitian Dan Pembelajaran IPA*, 6(1), 102. <https://doi.org/10.30870/jppi.v6i1.5718>
- Arnaiz-Sánchez, P., De-Haro Rodríguez, R., Alcaraz García, S. y Caballero García, C. M. (2019). Escuchar las voces del alumnado para construir la inclusión y la equidad educativa: diseño y validación de un cuestionario. *Multidisciplinary Journal of Educational Research*, 9(3), 258. <https://doi.org/10.17583/remie.2019.4613>

- Banerjee, M., Capozzoli, M., McSweeney, L. y Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, 27(1), 3-23. <https://doi.org/10.2307/3315487>
- Bernal-García, M., Salamanca, D., Perez, N., y Quemba, M. (2020). Validez de contenido por juicio de expertos de un instrumento para medir percepciones físico-emocionales en la práctica de disección anatómica. *Educación Médica*, 21(6), 349-356.
- Bowen-Mendoza, L., Pinargote-Ortega, M., Meza, J. y Ventura, S. (2022). Design of peer assessment rubrics for ICT topics. *Journal of Computing in Higher Education*, 34(1), 211-241. <https://doi.org/10.1007/s12528-021-09297-9>
- Cabero, J. y Llorente, M. (2013). Aplicación de juicio de experto como técnica de evaluación de las TIC. *Revista de Tecnología de Información y Comunicación en Educación*, 7(2), 11-22. <http://www.mriuc.bc.uc.edu.ve/bitstream/123456789/1175/1/art01.pdf>
- Cabero, J. e Infante, A. (2014). Empleo del método Delphi y su empleo en la investigación en comunicación y educación. *EDUTEC Revista electrónica de tecnología educativa*, 48, 1-16.
- Cabero-Almenara, J., Romero-Tena, R. y Palacios-Rodríguez, A. (2020). Evaluation of Teacher Digital Competence Frameworks Through Expert Judgement: The Use of the Expert Competence Coefficient. *Journal of New Approaches in Educational Research*, 9(2), 275. <https://doi.org/10.7821/naer.2020.7.578>
- Cajas-Bravo, T. V., Silva-Infantes, M. y Dávila-Morán, R. C. (2023). Percepción docente sobre el uso de la tecnología en la educación superior. *Revista Conrado*, 19(90), 326-335.
- Cañongo, A. A. y Díaz-Barriga, F. (2023). Pensamiento crítico en la virtualidad: validación de un e-caso sobre elección vocacional. *Psicumex*, 13.
- Carrasco-Sáez, J. L., Butter, M. C., Badilla-Quintana, M. G., Pérez, L. J. y Farfán, J. M. (2019). Sociological importance and validation of a questionnaire for the sustainability of personal learning environments (PLE) in 8th grade students of the Biobío Region in Chile. *Sustainability (Switzerland)*, 11(5). <https://doi.org/10.3390/su11051301>
- Castro-Molinares, S. P. (2018). Diseño y validación de un instrumento para evaluar la actitud hacia la investigación formativa en estudiantes universitarios. *Actualidades Pedagógicas*, 1(70), 165-182. <https://doi.org/10.19052/ap.3996>
- Ceballos-Bernal, E. A., Hutchison-Salazar, L. R. y Correa-Bautista, J. E. (2024). Diseño y validación de un cuestionario de percepciones de estudiantes sobre

- la enseñanza a través de analogías en docentes de ciencias del deporte. *Retos*, 57, 494-501. <https://recyt.fecyt.es/index.php/retos/index>
- Cerda, J., y Villarroel, L. (2008). Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. *Revista Chilena de Pediatría*, 79(1), 54-58.
- Costa, M., y Linzan, J. (2022). Gestión Sostenible: Diversificación turística para mejorar la oferta en las Siete Cascadas, cantón Naranjal. *Revista Ciencias Pedagógicas e Innovación*, 9(2), 142-152.
- Cohen, J. (1968). Nominal Scale Agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70 (4), 213-220.
- Cruz, M. y Martínez, M. C. (2018). Origen y desarrollo de un índice de competencia experta: el coeficiente k. *Revista Latinoamericana de Metodología de la Investigación Social*, 18(8), 40-56.
- Cuesta, C., Cossini, F. C. y Politis, D. G. (2023). Adaptación argentina del Cuestionario de Reserva Cognitiva: CRC-a. *Interdisciplinaria*, 40(1), 137-145. <https://doi.org/10.16888/interd.2023.40.1.7>
- Dávila-Morán, R. C., Pasquel-Cajas, A. F., Cribillero-Roca, M. C., Arroyo-Vigil, V. M., y Bustamante-Paredes, R. M. (2023). Competencia digital docente y tecnologías de información y comunicaciones en profesores universitarios. *Revista Conrado*, 19(90), 146-156.
- De Haro-Rodríguez, R. y Abellán, R. M. (2019). Development and validation of a questionnaire to listen to the voices of families of students enrolled in specialized open classrooms. *Publicaciones de la Facultad de Educación y Humanidades del Campus de Melilla*, 49(3), 119-148. <https://doi.org/10.30827/publicaciones.v49i3.11406>
- Ding, C. S. y Hershberger, S. L. (2002). Assessing content validity and content equivalence using structural equation modeling. *Structural Equation Modeling*, 9(2), 283-297. https://doi.org/10.1207/S15328007SEM0902_7
- Dobrov, G. M. y Smirnov, L. P. (1972). Forecasting as a means for scientific and technological policy control. *Technological Forecasting and Social Change*, 4(1), 5-18. [http://dx.doi.org/10.1016/0040-1625\(72\)90043-1](http://dx.doi.org/10.1016/0040-1625(72)90043-1)
- Domínguez-Lara, S. (2017). Construcción de una Escala de Autoeficacia para la Investigación: Primeras Evidencias de Validez. *Revista Digital de Investigación En Docencia Universitaria*, 11(2), 308-322. <https://doi.org/10.19083/ridu.11.514>
- Dorantes-Nova, J. A., Hernández-Mosqueda, J. S. y Tobón.Tobón, S. (2016). Juicio de expertos para la validación de un instrumento de medición del

- síndrome de *burnout* en la docencia. *Ra Ximhai*, 327-346.
<https://doi.org/10.35197/rx.12.01.e3.2016.22.jd>
- Dueñas-Zorrilla, M., Tejada-Fernández, J. y Pozos-Pérez, K. V. (2024). Diseño y validación de una escala de autoevaluación de la competencia digital docente y actitud hacia la innovación educativa del profesorado en servicio. *Revista Complutense de Educación*, 35(2), 237-252.
- Escobar-Pérez, J. y Cuervo-Martínez, Á. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición*, 6, 27-36.
- Fernández-Gómez, E., Martín-Salvador, A., Luque-Vara, T., Sánchez-Ojeda, M. A., Navarro-Prado, S. y Enrique-Mirón, C. (2020). Content validation through expert judgement of an instrument on the nutritional knowledge, beliefs, and habits of pregnant women. *Nutrients*, 12(4).
<https://doi.org/10.3390/nu12041136>
- Figuroa-Oquendo, A. E. (2023). Inteligencia emocional y rendimiento académico en estudiantes de una universidad ecuatoriana. *Revista Chakiñan*, 21, 140-152. <https://doi.org/10.37135/chk.002.21.09>
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 378-382.
- Galicia, L. A., Balderrama, J. A., y Edel, R. (2017). Validez de contenido por juicio de expertos: propuesta de una herramienta virtual. *Apertura*, 9(2), 42-53.
<https://doi.org/10.32870/ap.v9n2.993>
- García, I. y Fernández, S. (2008). procedimiento de aplicación del trabajo creativo en grupo de expertos, *Energética*, 29(2), 46-50.
- García-Magallanes, A. L., Corona-Cruz, K., Hernández-Briones, M. E., Hernández-Ábrego, R. y Villalpa-Martínez, H. (2024). Diseño y validación de contenido por juicio de expertos de un instrumento para evaluar la efectividad del proceso educativo relacionado al emprendimiento en estudiantes de educación superior. *Revista Espacios*, 45(04).
<https://doi.org/10.48082/espacios-a24v45n04p14>
- Garrote, P. R., y del Carmen Rojas, M. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 18, 124-139.
<https://doi.org/10.26378/rnlael918259>
- Gazca-Herrera, L. A., Garizurieta Bernabe, J., y Muñoz Martínez, A. Y. (2023). Estudio de la tutoría académica de un posgrado consolidado y su impacto en la trayectoria y egreso. *Revista RIDE*, 13(26), e492.
<https://doi.org/10.23913/ride.v13i26.1514>

- González-Zamar, M. D., Jiménez, L. O. y Ayala, A. S. (2021). Design and validation of a questionnaire on influence of the university classroom on motivation and sociability. *Education Sciences*, 11(4). <https://doi.org/10.3390/educsci11040183>
- Guillot-Valdés, M., Guillén-Riquelme, A., y Buela-Casal, G. (2022). Content Validity through Expert Judgment for the Depression Clinical Evaluation Test. *International Journal of Clinical and Health Psychology*, 22(2). <https://doi.org/10.1016/j.ijchp.2022.100292>
- Haeruddin, H., Prasetyo, Z. y Pahar, S. (2020). The development of a metacognition instrument for college students to solve physics problems. *International Journal of Instruction*, 13(1), 767-782. <https://doi.org/10.29333/iji.2020.13149>
- Hernández-Briones, M. E., García-Magallanes, A. L., Corona-Cruz, K., Villalpa-Martínez, H. y Hernández-Ábrego, R. (2024). Diseño y validación de contenido por juicio de expertos de un instrumento para evaluar el impacto del COVID-19 en estudiantes de educación superior. *Revista Espacios*, 45(06). <https://doi.org/10.48082/espacios-a24v45n06p03>
- Herrera-Masó, J. R., Calero-Ricardo, J. J., González-Rangel, M.A., Collazo Ramos, M. I. y Travieso-González y. (2023). El método de consulta a expertos en tres niveles de validación. *Revista Habanera de Ciencias Médicas*, 21(1), e4711
- Hyrkäs, K., Appelqvist-Schmidlechner, K y Oksa, L. (2003). Validating an instrument for clinical supervision using an expert panel. *International Journal of nursing studies*, 40 (6), 619-625.
- Jorm, A. (2015). Using the Delphi expert consensus method in mental health research. *Australian and New Zealand Journal of Psychiatry*, 49(10) 887-897. <http://dx.doi.org/10.1177/0004867415600891>
- León-Garrido, A., Barroso-Osuna, J. M., y Llorente-Cejudo, M. C. (2024). Diseño y validación de un instrumento de evaluación para apps móviles musicales a través del juicio de expertos. *RIFIE*, 53(2), 169-179. <https://doi.org/10.17811/rifie.20620>
- Lin, M., Thoma, B., Trueger, N.S., Ankel, F., Sherbino, J., y Chan, T. (2015). Quality indicators for blogs and podcasts used in medical education: modified Delphi consensus recommendations by an international cohort of health professions educators. *Postgrad Med J Published Online*, 0, 1-5. <http://dx.doi.org/10.1136/postgradmedj-2014-133230>
- Luque-Vara, T., Linares-Manrique, M., Fernández-Gómez, E., Martín-Salvador, A., Sánchez-Ojeda, M. A., y Enrique-Mirón, C. (2020). Content validation of an instrument for the assessment of school teachers' levels of knowledge of

- diabetes through expert judgment. *International Journal of Environmental Research and Public Health*, 17(22), 1-13. <https://doi.org/10.3390/ijerph17228605>
- Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-385.
- Maldonado-Suárez, N., y Santoyo-Telles, F. (2024). Validez de contenido por juicio de expertos: Integración cuantitativa y cualitativa en la construcción de instrumentos de medición. *REIRE Revista d'Innovació i Recerca en Educació*, 17(2), 1-19. <https://doi.org/10.1344/reire.46238>
- Mamani-Benito, O. J. (2022). Estresores docentes durante la pandemia COVID-19 en profesores peruanos: adaptación de una escala. *Revista Fuentes*, 2(24), 174-183. <https://doi.org/10.12795/revistafuentes.2022.19870>
- Manley, R. A. (2013). The Policy Delphi: a method for identifying intended and unintended consequences of educational policy. *Policy futures in education*, 11(6), 755-768. <http://dx.doi.org/10.2304/pfie.2013.11.6.755>
- Martínez-Valdés, M. G., y Juárez-Hernández, L. G. (2019). Diseño y validación de un instrumento para evaluar la formación en sostenibilidad en estudiantes de educación superior. *Revista de Investigación Educativa*, 10(19), 37-54. http://www.rediech.org/ojs/2017/index.php/ie_rie_rediech/index
- Marín-González, F., Pérez-González, J., Senior-Naveda, A. y García-Guliany, J. (2021). Validación del diseño de una red de cooperación científico-tecnológica utilizando el coeficiente K para la selección de expertos. *Información Tecnológica*, 32(2), 79-88. <http://dx.doi.org/10.4067/S0718-07642021000200079>
- Matta-Huerta, C. R., Vivar-Bravo, J., Jesús-Carbajal, O., Vela-Yañac, K., Mejía-Castillo, C., Santos-Ku, S. (2023). Aprendizaje autónomo y recursos educativos digitales en estudiantes del I ciclo de una universidad privada de Lima. *Revista Horizontes*, 7(28), 712-727. <https://doi.org/10.33996/revistahorizontes.v7i28.549>
- McHugh, M. L. (2012). Lessons in biostatistics interrater reliability: the kappa statistic. *Biochemica Medica*, 22(3), 276-282. <https://hrcak.srce.hr/89395>
- Mengual, S. (2011). *La importancia percibida por el profesorado y el alumnado sobre la inclusión de la competencia digital en educación Superior* [tesis doctoral]. Universidad de Alicante.
- Muñiz, J. (2003). *Teoría clásica de los test*. Psicología Pirámide.
- Pedrosa, I., Suárez Álvarez, J., y García Cueto, E. (2014). Content validity evidences: Theoretical advances and estimation methods. *Acción Psicológica*, 10(2), 3-18. <https://doi.org/10.5944/ap.10.2.11820>

- Penfield, R. D. y Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science*, 8(4), 213-225. https://doi.org/10.1207/s15327841mpee0804_3
- Polit, D. F., Beck, C. T. y Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459-467. <https://doi.org/10.1002/nur.20199>
- Ponce, I., Juárez, L.G. y Tobón, S. (2020). Construcción y validación de un instrumento para evaluar el abordaje de la sociedad del conocimiento en docentes. *Apuntes Universitarios*, 10(1), 40-65. <https://doi.org/10.17162/revapuntes.v10i1.183>
- Posso-Pacheco, R. J., Córdor-Chicaiza, M. G., Mora-Guerrero, L. M., y Manosalvas, S. L. (2023). Aprendizaje basado en retos: una mirada desde la educación superior. *Revista PODIUM*, 18(2), e1486. <https://podium.upr.edu.cu/index.php/podium/article/view/1486>
- Robles, P., y Rojas, M. del C. (2015). *La validación por juicio de expertos: dos investigaciones cualitativas en Lingüística aplicada*.
- Rosselló, M. R., Pinya, C. y Morcillo, V. (2024). Caracterizar la innovación para validar la práctica educativa, *Revista Colombiana de Educación*, (90), 102-124. <https://doi.org/10.17227/rce.num90-14495>
- Sagaró del Campo, N. M., y Zamora-Matamoros, L. (2023). Validez de las reglas de un análisis estadístico implicative según juicio de expertos. *Revista Cubana de Informática Médica*, 15(2), e645.
- Sánchez-Rosas, J., Dyzenchautz, M., Dominguez-Lara, S., y Hayes, A. (2022). Collective Teacher Self-Efficacy Scale for Elementary School Teachers. *International Journal of Instruction*, 15(1), 985-1002. <https://doi.org/10.29333/iji.2022.15156a>
- Sandoval-Henríquez, F. J., y Badilla-Quintana, M. G. (2021). Measuring stimulation and cognitive reactions in middle schoolers after using immersive technology: Design and validation of the TINMER questionnaire. *Computers and Education*, 166, 104-157. <https://doi.org/10.1016/j.compedu.2021.104157>
- Santos, M. M., Silva, A. A., y Lima, W. (2024). Development and validation of a digital vulnerability identification instrument (DVI-Q) for K-12 students. *Educação em Revista*, 40, e39797. <http://dx.doi.org/10.1590/0102-469839797>

- Sarah, S., Prasetyo, Z. K. y Wilujeng, I. (2019). Development of living values (honesty) assessment instrument in learning process. *International Journal of Instruction*, 12(4), 443-458. <https://doi.org/10.29333/iji.2019.12429a>
- Sarmiento-Peralta, G. G. (2020). Diseño y validación de una escala de autoeficacia del docente universitario. *REDU. Revista de Docencia Universitaria*, 18(2), 131. <https://doi.org/10.4995/redu.2020.14343>
- Sierra-Vite, M. y Ramírez-Díaz, M. H. (2023). Validación por juicio de expertos: Rúbricas de competencias en física. *Revista Internacional de Aprendizaje*, 9(1). <https://doi.org/10.18848/2575-5544/CGP/v09i01/81-95>
- Silva-Hormazábal, M., y Alsina, Á. (2023). Promoviendo el desarrollo profesional docente en STEAM: diseño y validación de un programa de formación. *Revista de Estudios y Experiencias en Educación*, 22(50), 99-120. <https://doi.org/10.21703/rexe.v22i50.1986>
- Streiner, D., Norman, G. y Cairney, J. (2015). *Health measurement scales, a practical guide to their development and use* (5ta ed.). Oxford University Press.
- Soriano-Rodríguez, A. M. (2015). Diseño y validación de instrumentos de medición. *Diá-Logos*, 14, 19-40. <https://doi.org/10.5377/dialogos.v0i14.2202>
- Souto-Seijo, A., Estévez, I., Fustes, V. I., y González-Sanmamed, M. (2020). Between the formal and the non-formal: an analysis from the perspective of continuing teacher training. *Educator*, 56(1), 91-107. <https://doi.org/10.5565/REV/EDUCAR.1095>
- Suárez-Brieva, E., Pardo-Calvache, C., y Ordoñez-Eraso, H. (2024). Instrument to measure the perceptions of software development communities regarding communication, coordination and cooperation: *result of validation through expert judgment*. *INGE CUC*, 20(1), 116-134. <http://dx.doi.org/10.17981/ingecuc.20.1.2024.10>
- Teistler, N. (2021). Development of an Instrument to Assess Pre-Service Teachers' Attitudes on Person-Centered Behavior Toward Students (APBS): Gathering Validity Evidence Based on Test Content. *International Journal of Educational Research*, 110(September 2020), 101878. <https://doi.org/10.1016/j.ijer.2021.101878>
- Terezinha, E. (2007). Editorial: Revisión sistemática x Revisión narrativa. *Acta Paulista de Enfermagem*, 20(2), ix.
- Tinoco-Plasencia, C. J. (2023). Competencias digitales y aprendizaje autónomo en estudiantes de una universidad privada de Lima. *Revista Conrado*, 19(91), 444-454.

- Villiers, M., Villiers, P., y Kent, A. (2005). The Delphi technique in health sciences education research. *Medical Teacher*, 27(7), 639-643. <https://doi.org/10.1080/13611260500069947>.
- Vinces, L. (2018). *Diseño y Validación de una e-Rúbrica para evaluar las Competencias Clínicas Específicas en Diagnóstico Diferencial en Pediatría*. Trabajo de grado Magíster, Universidad Casa Grande.
- Yeomans-Cabrera, M. M. y Martínez-Líbano, J. (2024). Evaluation of Teacher-Induction Programs (E-Tip): validity of the Criteria. *Revista de Estudios y Experiencias en Educación*, 23(51), 213-233. <https://doi.org/10.21703/rexe.v23i51.1706>
- Yévenes-Márquez, J. N., Badilla-Quintana, M. G. y Sandoval-Henríquez, F. J. (2022). Measuring Engagement to Academic Tasks: Design and Validation of the Comp-TA Questionnaire. *Education Research International*, 1-9. <https://doi.org/10.1155/2022/4783994>
- Zartha, J. W., Montes, J. M., Toro, I., Hernández, R. Villada, H. y Hoyos, J. (2017). Delphi Method in Technological Foresight Studies: An Approach to Calculating the Number of Experts and the Application of the Competence Coefficient “k”. *Rev. Biotecnología en el Sector Agropecuario y Agroindustrial*, 15(1), 105-115. [http://dx.doi.org/10.18684/BSAA\(15\)105-115](http://dx.doi.org/10.18684/BSAA(15)105-115)

Agradecimientos

Se agradece el apoyo de ANID/PIA/Fondos para Centros de Excelencia ANID-CIAE-CIA-250005, Chile.

Anexos

Anexo 1. Número de Expertos para el Proceso de Validación

Una revisión de 38 estudios revela una variabilidad significativa en el número de expertos involucrados, con un rango de 3 a 38 expertos por estudio y una media de aproximadamente 10. La mayor concentración de investigaciones realizó con un grupo de entre 6 y 10 expertos.

Tabla 5.

Resumen de estudios consultados para identificar el número de expertos

Número de expertos	Investigaciones
Entre 3 y 5 expertos	Acuña-Condori et al., 2024; Aliaga-Pacora et al., 2021; Bowen-Mendoza et al., 2022; Cajas-Bravo et al., 2023; Castro-Molinares, 2018; Dávila-Morán et al., 2023; Figueroa-Oquendo, 2023; Gazca-Herrera et al., 2023; Hernández-Briones et al., 2024; Matta-Huerta et al., 2023; Posso-Pacheco et al., 2023; Sarmiento-Peralta, 2020; Sierra-Vite y Ramírez-Díaz, 2023; Silva-Hormazábala y Alsina, 2023; Tinoco-Plasencia, 2023
Entre 6 y 10 expertos	Agudelo-Gallego et al., 2024; Albán-Romero et al., 2023; Arnaiz-Sánchez et al., 2019; Bernal-García et al., 2020; Cañongo y Díaz-Barriga, 2023; Carrasco-Sáez et al., 2019; Ceballos-Bernal et al., 2024; Cuesta et al., 2023; De Haro-Rodríguez y Abellán, 2019; Dueñas-Zorrilla et al., 2024; García-Magallanes et al., 2024; González-Zamar et al., 2021; Haeruddin et al., 2020; Mamani-Benito, 2022; Sandoval-Henríquez y Badilla-Quintana, 2021; Suárez-Brieva et al., 2024; Yévenes-Márquez et al., 2022
Entre 11 y 20 expertos	Aceves-Villanueva et al., 2024; Guillot-Valdés et al., 2022; Ponce et al., 2020
Más de 21 expertos	Dorantes-Nova et al., 2016; Rosselló et al., 2024; Yeomans-Cabrera y Martínez-Líbano, 2024

Fuente: elaboración propia.

Anexo 2. Indicadores o criterios de evaluación

En una revisión de 33 artículos de investigación en las áreas de psicología y educación, se evidenció que los principales criterios de búsqueda fueron la Claridad, presente en 25 estudios, seguida por la Relevancia (21) y la Pertinencia (18) y por último la Coherencia en 14 estudios.

Tabla 6.

Tipos de indicadores utilizados al realizar juicio de expertos.

Autores	Criterios	CL	RE	C	P	S	O	OT	Total
Aceves-Villanueva et al., 2024		1	1	1		1			4
Agudelo-Gallego et al., 2024		1							1
Albán-Romero et al., 2023		1			1				2
Aliaga-Pacora et al., 2021			1		1				2
Almanasreh et al., 2019		1	1					1	3
Arnaiz-Sánchez et al., 2019		1			1			1	3
Bernal-García et al., 2020		1	1	1	1	1			5
Cajas-Bravo et al., 2023		1		1					2
Carrasco-Sáez et al., 2019		1	1			1			3
Ceballos-Bernal et al., 2024		1	1	1	1	1			5
Costa y Linzan, 2022					1		1		2
Dávila-Morán et al., 2023		1	1		1				3
Dueñas-Zorrilla et al., 2024			1		1			1	3
Domínguez-Lara, 2017		1	1						2

Fernández-Gómez <i>et al.</i> , 2020	1	1	1					3
García-Magallanes <i>et al.</i> , 2024				1			1	2
Gazca-Herrera <i>et al.</i> , 2023	1	1	1		1			4
González-Zamar <i>et al.</i> , 2021	1		1	1		1		4
Guillot-Valdés <i>et al.</i> , 2022	1	1	1	1				4
Hernández-Briones <i>et al.</i> , 2024	1		1	1			1	4
Mamani-Benito, 2022	1	1		1				3
Ponce <i>et al.</i> , 2020	1			1				2
Posso-Pacheco <i>et al.</i> , 2023	1	1	1					3
Rosselló <i>et al.</i> , 2024	1		1	1				3
Sánchez-Rosas <i>et al.</i> , 2022	1			1				2
Sandoval-Henríquez y Badilla-Quintana, 2021	1	1	1		1			4
Santos <i>et al.</i> , 2024		1				1	1	3
Sarmiento-Peralta, 2020	1		1	1				3
Sierra-Vite y Ramírez-Díaz, 2023		1		1				2
Suárez-Brieva <i>et al.</i> , 2024	1	1	1		1			4
Teistler, 2021	1	1	1					3
Tinoco-Plasencia, 2023	1	1		1				3
Yeomans-Cabrera y Martínez-Líbano, 2024		1						1
Total	26	21	15	19	7	3	6	97

Fuente: elaboración propia.

Nota: CL: Claridad; RE: Relevancia; C: Coherencia; S; P: Pertinencia; Suficiencia; O: Objetividad; OT: Otras.

Anexo 3. Valoraciones de Respuesta

Un análisis de 32 artículos de investigación muestra que el 87,5 % utilizaron escalas tipo Likert en el proceso de validación, mayormente con valoraciones entre 1 y 5 puntos.

Tabla 7.

Tipos de valoración de respuesta en el proceso de validación a través de juicio de expertos.

Autor/es	Valoración de respuesta	Graduación
Domínguez-Lara, 2017	Likert	0-3
Suárez-Brieva <i>et al.</i> , 2024	Likert	0-4
Agudelo-Gallego <i>et al.</i> , 2024	Likert	1-3
Aceves-Villanueva <i>et al.</i> , 2024; Anggraeni y Mundilarto, 2020; Bernal-García <i>et al.</i> , 2020; García-Magallanes <i>et al.</i> , 2024; Garrote y del Carmen-Rojas, 2015; León-Garrido <i>et al.</i> , 2024; Rosselló <i>et al.</i> , 2024; Sarah <i>et al.</i> , 2019; Teistler, 2021	Likert	1-4

Arnaiz-Sánchez <i>et al.</i> , 2019; Cajas-Bravo <i>et al.</i> , 2023; Cañongo y Díaz-Barriga, 2023; Carrasco-Sáez <i>et al.</i> , 2019; Dávila-Morán <i>et al.</i> , 2023; Dueñas-Zorrilla <i>et al.</i> , 2024; Figueroa-Oquendo, 2023; Ponce <i>et al.</i> , 2020; Sagaró del Campo y Zamora-Matamoros, 2023; Sánchez-Rosas <i>et al.</i> , 2022; Santos <i>et al.</i> , 2024; Sierra-Vite y Ramírez-Díaz, 2023; Silva-Hormazábala y Alsina, 2023; Tinoco-Plasencia, 2023; Yeomans-Cabrera y Martínez-Líbano, 2024	Likert	1-5
Ceballos-Bernal <i>et al.</i> , 2024	Likert	1-6
Carrasco-Sáez <i>et al.</i> , 2019; Fernández-Gómez <i>et al.</i> , 2020; Sarmiento-Peralta, 2020; Sandoval-Henríquez y Badilla-Quintana, 2021	Dicotómica	Sí/No

Fuente: elaboración propia.