Instituto de Investigación y Capacitación Profesional del Pacífico [IDICAP P]
Revista Latinoamericana de Educación [RLE]
Vol. 3 N° 3
ISSN: 3028-9769 ISSN-L: 3028-9769
DOI: 10.53595/rle.v3.i3.015

ARTÍCULO

Validación de contenido por expertos: concordancia interjueces y modelo estandarizado para instrumentos de investigación

Expert-based content validation: interjudge agreement and standardized model for research instruments

Reinaldo Antonio Guerrero Chirinos [a*], José Ramón Delgado Fernández [b], Andrés Pacheco Molina [c], Cristina Isabel Vivanco Ureña [d], Jean Pierre Reyes Carrión [e], José Vivanco Ureña [f]

[a b] Universidad Técnica Particular de Loja, Loja, Ecuador
[c] Universidad Técnica de Machala, Machala, Ecuador
[d] Universidad Nacional de Loja, Loja, Ecuador
[e] Universidad Nacional de Educación, Azogues, Ecuador
[f] Unidad Educativa Santa Mariana de Jesús, Loja, Ecuador

Recibido el 20 de mayo del 2025, aceptado el 26 de junio del 2025, en línea el 30 de junio del 2025.

Resumen

La validez de contenido mediante juicios de expertos es necesaria para validar instrumentos de investigación, pero exige métodos sólidos para cuantificar su consenso. Este artículo analiza la validación de contenido de un instrumento según la concordancia entre expertos en distintos ítems, identifica elementos críticos y propone un modelo estandarizado para validación. Se aplicó una metodología cuantitativa de corte transversal, con 25 expertos que evaluaron 14 ítems en escala de Likert. Se determinaron el coeficiente W de Kendall, rangos promedio y se realizó una prueba de hipótesis, revelando una concordancia global baja pero significativa (W=0,208 y p<0,001). Los ítems "Anonimato" y "Organización" registraron mayor consenso (rangos 9,06 y 8,50), mientras que "Experticia" y "Actualidad" mostraron menor acuerdo (rangos 4,46 y 5,02). Se concluye que la concordancia permitió identificar ítems críticos y proponer un modelo de validación en cuatro fases con umbrales estadísticos (W≥0,6), que combina iterativamente métodos cuantitativos y cualitativos.

Palabras clave: metodología, estadística, juicio de valor, investigación, experto.

Abstract

Content validity through expert judgment is essential for validating research instruments but requires solid methods to quantify expert consensus. This article analyzes the content validation of an instrument based on expert agreement across different items, identifies critical elements, and proposes a standardized validation model. A cross-sectional quantitative methodology was applied, involving 25 experts who evaluated 14 items using a Likert scale. Kendall's W coefficient, average ranks, and a hypothesis test were calculated, revealing low but significant overall agreement (W = 0.208; p < 0.001). The items "Anonymity" and "Organization" showed the highest consensus (ranks 9.06 and 8.50), while "Expertise" and "Currency" reflected lower agreement (ranks 4.46 and 5.02). The observed concordance enabled the identification of critical items and the proposal of a four-phase validation model with statistical thresholds (W ≥ 0.6), combining quantitative and qualitative methods iteratively.

Keywords: Methodology, statistics, value judgment, research, expert.

*Autor para correspondencia: raguerrero12@utpl.edu.ec

1. Introducción

La validez de contenido en los instrumentos de investigación es fundamental para preservar la precisión y el rigor científico en la generación de conocimiento, especialmente en el ámbito de la educación y las ciencias sociales, donde los constructos teóricos tienden a ser complejos (Hernández, 2011). Sin embargo, Maldonado y Santoyo (2024) destacan que pocos estudios incorporan estadísticas de concordancia entre jueces, y que son escasos los protocolos estandarizados que permitan cuantificar e interpretar el consenso entre los expertos. Estas debilidades elevan el riesgo de validar instrumentos con sesgos conceptuales y escasa aplicabilidad práctica.

Lo expuesto anteriormente plantea dos desafíos principales. El primero se relaciona con la falta de estandarización en la evaluación de expertos. Aunque algunos estudios, como el de Lao et al. (2016), enfatizan criterios para la selección de los especialistas en la materia, aún existen vacíos en protocolos que integren herramientas estadísticas como el coeficiente de concordancia W de Kendall, y umbrales de decisión bien establecidos. El segundo desafío es la escasa aplicación de métodos cuantitativos y pruebas de hipótesis para distinguir entre la concordancia real y la variabilidad aleatoria (Escobar & Cuervo, 2008), lo que limita la objetividad al momento de identificar los ítems críticos.

Diversas investigaciones recientes (Herrera et al., 2022; Marín et al., 2021; Maskavizan et al., 2023) han propuesto el coeficiente W de Kendall como herramienta para evaluar la concordancia en datos ordinales, pero sin relacionarlo con modelos iterativos de mejora. Este estudio busca cerrar esa brecha metodológica mediante un enfoque integral que combine diagnóstico cuantitativo (concordancia global y rangos promedio), rediseño cualitativo con base en la retroalimentación de expertos y validación final mediante umbrales estadísticos estandarizados ($W \geq 0.6$).

El propósito central de este estudio es evaluar la validación de contenido de un instrumento basada en el consenso de expertos, determinar su nivel de concordancia a través del coeficiente W de Kendall, identificar los ítems críticos mediante rangos promedio y proponer un modelo estructurado de cuatro fases con umbrales estadísticos que fortalezcan la validez de los instrumentos de investigación.

1.1. Sustento teórico

Juicio de expertos

Evaluar la validez de un instrumento de investigación es esencial para asegurar la credibilidad de los resultados. En áreas como la educación y la tecnología, esto es especialmente importante debido a la complejidad de los fenómenos estudiados y sus implicaciones prácticas. La literatura destaca la importancia de evaluar varios tipos de validez para garantizar la calidad de la medición. Por ejemplo, Hernández (2011) sugiere que la validez de contenido se puede evaluar mediante la revisión de expertos y la comparación con otros instrumentos similares.

Seleccionar a los expertos adecuados implica considerar su experiencia en metodología, diseño de instrumentos y el análisis de datos relacionados con el contenido temático. Por tanto, la validez de los expertos es un asunto complejo y multifacético. Aunque los expertos pueden presentar limitaciones en sus valoraciones, su contribución sigue siendo valiosa y puede mejorar significativamente el proceso de toma de decisiones. Además, sus juicios pueden variar influidos por el contexto en el que se desempeñan.

La validez de contenido mediante el juicio de expertos es fundamental en la construcción de instrumentos de medición. En esta etapa, un conjunto de especialistas valoran cada uno de los ítems del instrumento para evaluar si reflejan de una manera adecuada todas las dimensiones de la variable o constructo que se desea medir. Maldonado y Santoyo (2024) proponen una estrategia para este proceso que comprende:

a) Perfiles de los expertos: Se recomienda convocar a investigadores con experiencia en la temática, y que tengan imparcialidad respecto al equipo investigador y al instrumento. Al respecto, se consideraron los criterios propuestos por Lao et al. (2016), que incluyen grado académico alcanzado por el experto, cantidad, visibilidad y calidad de artículos publicados, libros publicados, participación en eventos científicos, premios, consultorías y años de experiencia en el tema.

b) Número de expertos: La participación de múltiples evaluadores es esencial para minimizar la variabilidad individual. Esta variabilidad debe reducirse mediante la estandarización del proceso y, cuando sea posible, mediante capacitación. Maldonado y Santoyo (2024) consideran que no existe consenso sobre el número de integrantes para el panel de expertos, sin embargo, sugieren valorar las posibilidades de contactar a los potenciales candidatos, su disponibilidad de tiempo, y determinar si existe un número suficiente de profesionales que cumplan con el perfil establecido. En este estudio participaron 25 expertos.

c) Evaluación: Una vez identificados y contactados los expertos, se les proporciona el instrumento a evaluar, así como una guía para realizar su valoración.

Coeficiente de concordancia W de Kendall

Barrueta et al. (2022) y Ramírez y Polack (2020) afirman que el coeficiente de concordancia W de Kendall se utiliza con variables de tipo ordinal y mide el grado de acuerdo entre k evaluadores al asignar rangos o calificaciones a un conjunto de N ítems.

Según Maskavizan et al. (2023), el coeficiente W varía entre 0 (nula concordancia) y 1 (máxima concordancia). Asimismo, esos autores proponen un baremo para interpretar el coeficiente W de acuerdo a lo expresado en la Tabla 1.

Tabla 1

Baremo para la interpretación de los valores del coeficiente W de Kendall

Valor	Grado de consenso
$\boldsymbol{W < 0,2}$	Ligero/Insignificante
$\boldsymbol{0,2 \leq W < 0,4}$	Bajo
$\boldsymbol{0,4 \leq W < 0,6}$	Moderado
$\boldsymbol{0,6 \leq W < 0,9}$	Alto
$\boldsymbol{W \geq 0,9}$	Inusualmente alto

Para la prueba de hipótesis, se seguirán las recomendaciones de Escobar y Cuervo (2008), las cuales se muestran en la Tabla 2.

Tabla 2

Prueba de hipótesis

Estadístico	Información que provee	Hipótesis	Interpretación
Coeficiente de concordancia W de Kendall	El grado de concordancia entre varios rangos de n objetos o individuos. Aplicable a estudios interjuicio o confiabilidad interprueba	H_0: Los rangos son independientes, no concuerdan H_1: Hay concordancia significativa entre los rangos	Cuando el nivel de significancia es inferior a 0.05, se rechaza la Ho y se concluye que hay concordancia significativa entre los rangos asignados por los jueces. Además se interpreta la fuerza de la concordancia, que aumenta cuando W se acerca a 1

Rangos promedio

Los rangos promedio son los valores promedio de las posiciones o rangos asignados a cada ítem por todos los expertos que emitieron su juicio. Representan la posición relativa promedio que cada uno de los expertos asignó a cada uno de los ítems. Dicho de otro modo, para cada uno de los N ítems evaluados, se calcula el promedio de los rangos que cada uno de los k expertos le ha asignado (Barrueta et al., 2022).

En esta investigación, para el cálculo de los rangos promedio, todos los expertos evalúan todos los ítems con una escala Likert del 1 al 3. Para cada uno de los 25 expertos se asignan rangos a cada uno de los 14 ítems según las puntaciones dadas, de modo que los ítems con mayor y menor puntuación reciben, respectivamente, el rango más alto y más bajo, que son 14 y 1. Si dos o más ítems tienen el mismo puntaje, se les asigna el promedio de sus rangos. Finalmente se suman los rangos de cada ítem en las 25 evaluaciones y se divide entre el número de jueces (25) para obtener el rango promedio por ítem.

Es pertinente observar que un rango promedio más alto sugiere que los jueces ubicaron ese ítem entre los más favorablemente evaluados o mejor formulados y viceversa; un rango promedio más bajo, indica que el ítem es menos adecuado o más irrelevante.

Cabe destacar que, si todos los expertos coinciden que un ítem es el mejor evaluado, recibirá un rango promedio de 14; de la misma forma, sí coinciden que un ítem es el peor evaluado, su rango promedio será 1. El rango promedio refleja la posición relativa del ítem respecto a los demás, no su calidad absoluta. Esto lleva a proponer un baremo para la interpretación de los rangos promedio, el cual se muestra en la Tabla 3.

Tabla 3

Baremo para la interpretación de los rangos promedio

Valor	Grado de consenso
11 – 14	Valoración muy alta (fuertemente deseado)
9 – 10,9	Valoración alta (bien valorado pero sin ser el mejor)
7,5 – 8,9	Valoración media-alta (aceptable, por encima del promedio)
6 – 7,4	Valoración media-baja (por debajo del promedio)
3 – 5,9	Valoración baja (con problemas o baja aceptación)
1 – 2,9	**Valoración muy baja (fuertemente rechazado)**

2. Métodos

Esta investigación se desarrolló bajo un enfoque cuantitativo porque se basó en la recolección y análisis de datos numéricos y se centró en la medición objetiva, la estadística y el análisis cuantitativo de los datos recopilados (Cueva et al., 2023). Además, es de tipo transversal porque se realizó en un momento específico (Cvetkovic et al., 2021).

La investigación se llevó a cabo en diferentes centros educativos en España y Ecuador, donde trabajan expertos en investigación científica del área de educación primaria, secundaria y universitaria. En total, la muestra estuvo compuesta por 25 docentes, de los cuales 16 son hombres y 9 mujeres. Además, 3 participantes tienen título de magister y los 22 restantes, título de doctor. La selección de los expertos fue una parte crucial del proceso, y se consideraron los criterios propuestos por Lao et al. (2016), que incluyen grado académico alcanzado por el experto, cantidad, visibilidad y calidad de artículos publicados, libros publicados, participación en eventos científicos, premios, consultorías y años de experiencia en el tema.

Antes de comenzar la investigación, se presentó la propuesta a los equipos directivos de cada centro educativo, quienes aprobaron su ejecución. Asimismo, se obtuvo el consentimiento informado

de todos los participantes, quienes fueron informados sobre los objetivos del estudio, su carácter voluntario, la confidencialidad de sus respuestas y el anonimato de sus datos. Se garantizó la no existencia de conflictos de interés y se aseguró la imparcialidad de los evaluadores. El protocolo de investigación fue revisado y aprobado por un comité ético universitario, de conformidad con las regulaciones vigentes en materia de investigación educativa con sujetos humanos.

El estudio fue implementado por el grupo de investigación del Departamento de Ciencias de la Educación de la UTPL. Todos los participantes recibieron las mismas instrucciones de aplicación, así como algunas orientaciones sobre el uso del software estadístico SPSS, versión 26.

Se creó un instrumento de investigación para analizar la validez de contenido de un instrumento mediante el consenso de los juicios emitidos por expertos en educación y tecnología. Este instrumento contiene 14 ítems con respuestas de medida ordinal, que se valoraron utilizando una escala de Likert de 3 alternativas, asignando el valor 1 a "No", 2 a "A veces" y 3 a "Sí".

Las primeras seis preguntas corresponden a datos generales (1. “Se informa de la autoría de la investigación”; 2. “Se solicita la colaboración al destinatario”; 3. “Se justifica la relevancia de la investigación”; 4. “Se motiva al destinatario a que colabore”; 5. “Se agradece la colaboración del destinatario”; 6. “Se asegura el anonimato y la confidencialidad”).

Las ocho preguntas restantes estuvieron referidas a la validación en sí (7. Claridad: “¿El lenguaje utilizado en la redacción de las preguntas es claro y adecuado para la población objetivo?”; 8. Coherencia: “¿Las preguntas elaboradas tienen relación con el título y con los diferentes aspectos?”; 9. Metodología: “¿El instrumento elaborado responde al objetivo de la investigación?”; 10. Suficiencia: “¿La calidad y la cantidad de preguntas son adecuadas?”; 11. Experticia: “¿Existe una relación del contenido del instrumento con el conocimiento de los encuestados?”; 12. Organización: “¿Existe una secuencia lógica y ordenada de las preguntas?”; 13. Pertinencia: “¿Considera que las opciones empleadas son correctas para medir los diferentes aspectos de la investigación?”; 14. Actualidad: “¿Considera de actualidad el tema tratado?”).

Para evaluar la validez del instrumento de investigación, se llevaron a cabo varios pasos. En primer lugar, se recopilaron los instrumentos validados por los 25 profesionales expertos en educación y/o tecnología educativa. Una vez obtenidos los resultados, se determinaron los rangos promedio de cada ítem para jerarquizarlos e interpretarlos, se aplicó el coeficiente W de Kendall para evaluar la concordancia global y se realizó una prueba de hipótesis para evitar que el valor de W sea atribuible al azar..

3. Resultados y discusión

3.1 Rangos promedio

La primera parte de este análisis de resultados corresponde al análisis de los rangos promedio, los cuales son los valores promedio de los rangos asignados a cada uno de los 14 ítems por cada uno de los 25 expertos que emitieron su juicio.

En SPSS versión 26, al aplicar la secuencia Analizar > Pruebas no paramétricas > Cuadros de diálogo antiguos > k muestras relacionadas, seleccionar los 14 ítems y la prueba W de Kendall, y darle clic al botón Aceptar, se obtiene la lista de rangos promedio por cada ítem, la cual se muestra en la Tabla 4.

En atención a los resultados obtenidos en la Tabla 4 y al baremo mostrado en la Tabla 3, se analizan los rangos promedio de cada uno de los ítems. Cabe destacar que ninguno de los ítems obtuvo rangos promedio de 10 o mayor, lo que significa que ninguno de ellos se ubicó en la categoría de ítem bien evaluado por los expertos.

1. Comenzando el análisis indivual de cada uno de ellos, el rango promedio r = 8,34 del ítem “Se informa de la autoría de la investigación” indica que es bien valorado por los jueces,

aunque se puede reforzar. Se sugiere que deben agregarse algunas credenciales las cuales pudieran ser afiliación institucional o una breve biografía.

2. El ítem "Se solicita la colaboración al destinatario" es aceptable, por encima del promedio (r = 7,84). Cumple su función, pero no es de los más destacados. Esto muestra que la invitación a participar pudiera ser clara o poco persuasiva. En este caso, se sugiere incorporar los beneficios de participar, así como ser más específico en aspectos como propósito, acción requerida o tiempo.

Tabla 4

Rangos promedio correspondientes a cada ítem

Rangos		
	Rango promedio (r)	Categoría según el baremo
Se informa de la autoría de la investigación	8,34	Media-alta
Se solicita la colaboración al destinatario	7,84	Media-alta
Se justifica la relevancia de la investigación	7,82	Media-alta
Se motiva al destinatario a que colabore	7,62	Media-alta
Se agradece la colaboración del destinatario	7,62	Media-alta
Se asegura el anonimato y la confidencialidad	9,06	Alta
Claridad: ¿El lenguaje utilizado en la redacción de las preguntas es claro y adecuado para la población objetivo?	8,46	Media-alta
Coherencia: ¿Las preguntas elaboradas tienen relación con el título y con los diferentes aspectos de investigación	8,34	Media-alta
Metodología: ¿El instrumento elaborado responde al objetivo de la investigación?	7,78	Media-alta
Suficiencia: ¿La calidad y la cantidad de preguntas son adecuadas?	6,94	Media-baja
Experticia: ¿Existe una relación del contenido del instrumento con el conocimiento de los encuestados?	4,46	Baja
Organización: ¿Existe una secuencia lógica y ordenada de las preguntas?	8,50	Media-alta
Pertinencia: ¿Considera que las opciones empleadas son correctas para medir los diferentes aspectos de la investigación?	7,20	Media-baja
Actualidad: ¿Considera de actualidad el tema tratado?	5,02	Baja

3. El rango promedio r = 7,82 del ítem "Se justifica la relevancia de la investigación" lo evalúa como bien recibido, pero no sobresaliente. Se puede inferir que la fundamentación es válida, pero desvinculada con problemas urgentes. Al respecto, se propone la citación de estudios recientes que evidencien vacíos.

4. “Se motiva al destinatario a que colabore” presenta un rango promedio r = 7,62 que lo ubica ligeramente por encima del promedio, sin un claro impacto emocional. Cumple, pero se debe reforzar la motivación al destinatario.

5. Similar al ítem anterior, “Se agradece la colaboración del destinatario” presenta el mismo rango promedio (r = 7,62) que lo ubica como bien valorado, pero que se puede enriquecer el lenguaje de gratitud. Se sugiere modificar un posible agradecimiento genérico, por uno más personalizado.

6. El rango (r = 9,06) del ítem “Se asegura el anonimato y la confidencialidad” lo interpreta como bien valorado por los jueces. Esto sugiere que los procesos de anonimato y confidencialidad son claros y robustos, pero se debe verificar la coherencia con posibles regulaciones locales.

7. El ítem “Claridad” presenta rango promedio r = 8,46 que lo ubica con buena valoración, lo que muestra una pregunta relativamente bien redactada, con lenguaje comprensible, sin tecnicismos, pero con posibles discrepancias en términos específicos. En este caso se debe realizar una prueba de legibilidad y sustituir posibles términos muy académicos por otros más accesibles.

8. “Coherencia”, con un r = 8,34 se considera bien estructurado y pertinente, da a entender que las preguntas se relacionan con el título y los diferentes aspectos de la investigación, pero no en su máxima coherencia. Esto requiere una revisión para asegurar que cada pregunta se alinee con el título y demás aspectos.

9. “Metodología” es un ítem con un rango promedio r = 7,78 que lleva a entender que el instrumento no responde completamente al objetivo de la investigación y se deben realizar los ajustes mediante la incorporación o eliminación de preguntas.

10. El ítem “Suficiencia” tiene un rango promedio de r = 6,94. Esto sugiere que el número de preguntas puede ser adecuado, pero con redundancias. Requiere revisión en términos de extensión y profundidad. Ante esto, se sugiere realizar un análisis de redundancia.

11. El ítem “Experticia” es r = 4,46. Esto sugiere la existencia de preguntas que no se ajustan al conocimiento de la población objetivo y riesgo de no respuesta o datos inválidos. En ese sentido, se hace necesario rediseñar los ítems mediante grupos focales con muestra representativa, nivelar el lenguaje técnico según la educación promedio y, similarmente con el ítem “Claridad”, aplicar una prueba de legibilidad.

12. El r = 8,50 para el ítem “Organización” lo hace considerar con buena percepción en cuanto a la secuencia lógica de preguntas, pero que se debe hacer una revisión a la estructura y transición de las preguntas para realizar mejoras menores.

13. El rango promedio r = 7,20 del ítem “Pertinencia” lo ubica en la categoría media-baja y hace que requiera atención. Esto puede indicar problemas en las opciones de respuesta debido a que no son claras, suficientes o adecuadas, o dudas sobre si miden constructos adecuadamente y, en ese caso, se deben revisar las escalas de respuesta, por ejemplo, cambiar escala de Likert por diferencial semántico o validar con análisis factorial confirmatorio.

14. Finalmente, el ítem “Actualidad”, tiene un rango promedio bajo (r = 5,02) que indica la existencia de referencias teóricas obsoletas, elementos actuales con desfases o abordaje de problemáticas que no son actuales. Entre otras acciones, se sugiere incorporar tendencias emergentes o actualizar el marco teórico con literatura menor o igual a 5 años.

3.2. Análisis de la concordancia interjueces

El análisis de la concordancia interjueces se basó en los datos arrojados en la Tabla 5 y se plantearán las siguientes hipótesis:

H_0: No hay concordancia entre los 25 expertos en la evaluación de los 14 ítems (W = 0).

H_1: Existe concordancia significativa entre los expertos (W > 0).

Tabla 5
Estadísticos coeficiente de concordancia W de Kendall

Estadísticos de prueba	
N	25
W de Kendall[a]	,208
Chi-cuadrado	67,682
gl	13
Sig. asintótica	,000
a. Coeficiente de concordancia de Kendall	

3.3. Resultados de la variable 1

El valor W de Kendall es 0,208 (0,2≤W<0,4) e indica que el grado de consenso entre las evaluaciones de los 25 jueces sobre los 14 ítems es bajo. Esto sugiere un acuerdo real entre los jueces, pero insuficiente para validar el instrumento sin hacerle mejoras o modificaciones (ver Tabla 1). En el análisis realizado en la sección 3.1 se señaló que esas mejoras son menores en algunos ítems y sustanciales en otros.

La significancia o p-valor = 0,000 < 0,05, por lo que se rechaza la hipótesis nula y se confirma que la concordancia, aunque se considera baja, es estadísticamente significativa.

El valor de χ^2 es una medida de la discrepancia entre el consenso observado (W = 0,208) y el consenso nulo (W = 0), cuanto mayor sea χ^2, hay mayor evidencia de que existe algún nivel de acuerdo (no aleatorio) entre expertos. El valor crítico χ^2 al 99,9 % y 13 grados de libertad es 34, 53 y el valor observado es 67,682 (Walpole et al., 2012). Como el valor observado supera ampliamente al valor crítico, confirma que la probabilidad de obtener este resultado por azar es menor que 0,001 (Sig. = 0,000).

El elevado valor chi-cuadrado ($\chi^2 = 67.68$, $p < 0,001$) confirma que el consenso, aunque débil en magnitud (W = 0,208), es no aleatorio y suficientemente sólido para guiar mejoras específicas. En consecuencia, no se debe validar el instrumento en esta instancia.

El bajo valor del coeficiente de concordancia de Kendall (W = 0,208) puede atribuirse a diversos factores. Por un lado, la redacción de algunos ítems pudo haber generado ambigüedad en sus interpretaciones, especialmente en aquellos que involucran términos técnicos o conceptos abstractos. Por otro lado, los expertos provenían de contextos geográficos, institucionales y disciplinarios variados, lo cual hace más probable que apliquen diferentes criterios de evaluación. Esta diversidad, si bien enriquece el análisis, también puede haber limitado el consenso entre ellos. Las consecuencias prácticas de esta baja concordancia incluyen una mayor posibilidad de interpretar de forma divergente las dimensiones que se pretenden medir, lo que podría comprometer la validez del instrumento si no se aplican correctivos a tales discrepancias.

Estudios como los de Maskavizan et al. (2023) en enfermería y Marín et al. (2021) en cooperación científico-tecnológica, evidencian que valores bajos de W son comunes en rondas preliminares, pero pueden incrementarse significativamente cuando se aplican procesos iterativos de rediseño y retroalimentación. Dichas investigaciones respaldan la utilidad del coeficiente W como herramienta para el diagnóstico y la mejora continua, validando de esta manera su empleo en el presente estudio.

3.4. Modelo estandarizado de validación

Con base en los análisis realizados previamente, se propone un modelo estandarizado de cuatro fases, donde cada una de ellas cuenta con procedimientos específicos. Cada una de estas fases son las siguientes:

Fase 1 (evaluación inicial): en esta fase se establecen bases metodológicas sólidas y conforma un grupo de expertos calificado con los criterios explícitos propuestos por Lao et al. (2016) y un mínimo de 15.

Fase 2 (análisis cuantitativo): el objetivo de esta fase es cuantificar el consenso entre los jueces expertos y priorizar ítems críticos. Esta etapa incluye el cálculo e interpretación del coeficiente W de Kendall, la prueba de hipótesis para determinar la significancia de la concordancia y el cálculo de rangos promedio para asignar la posición relativa de cada ítem, identificar los que sean críticos y sugerir posibles mejoras.

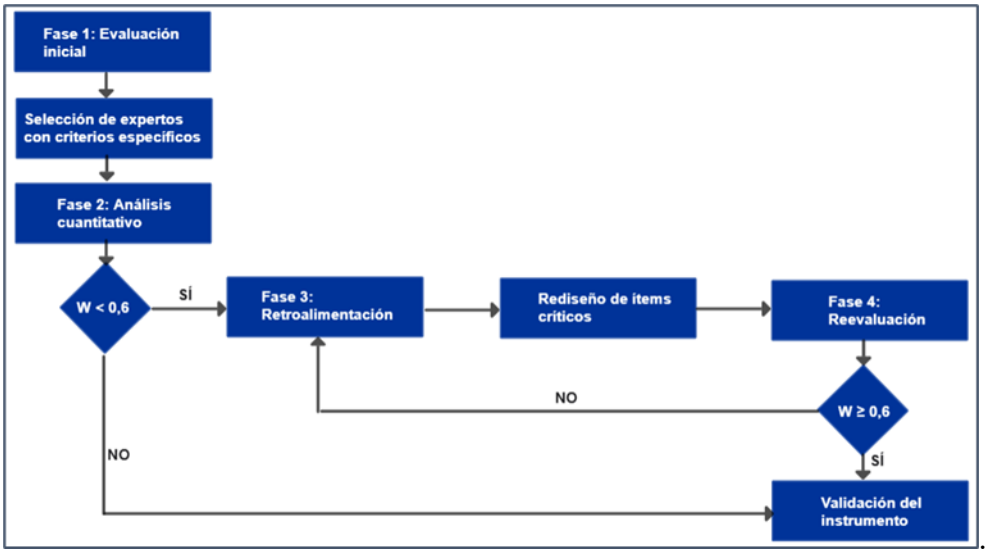
Fase 3 (retroalimentación cualitativa): aquí se busca rediseñar ítems críticos mediante los aportes especializados de los expertos. Esto incluye sesiones de trabajo con los especialistas para sugerir cambios e introducir rediseños en ítems críticos de modo que se obtenga una nueva versión mejorada del instrumento, con cambios totalmente justificados.

Fase 4 (reevaluación): la intención de esta fase es validar la versión final del instrumento. Para ello, los mismos expertos evalúan solo los ítems rediseñados, se realiza nuevamente la fase 2. Si $W \geq 0{,}6$, el instrumento se valida y se emite el certificado correspondiente; si $W < 0{,}6$, se deben repetir las fases 3 y 4.

En la Figura 1 se muestra un esquema del modelo estandarizado propuesto para la validación de instrumentos de investigación.

Figura 1

Mapa conceptual del modelo estandarizado de validación



4. Conclusiones

En la presente investigación se logró analizar la validación de contenido del instrumento mediante consenso de expertos en diferentes ítems, determinando que la concordancia es baja ($W = 0,208$), pero significativa ($p < 0,001$), haciendo evidente que los expertos coinciden en la jerarquización de los ítems, pero no en la intensidad de sus valoraciones. El $\chi^2 = 67,682$ ($p < 0,001$) confirma que el consenso no es aleatorio, a pesar del bajo valor de la W de Kendall. Esa baja magnitud de W exige una ronda de ajustes como mínimo.

Asimismo, se identificaron los ítems críticos "Experticia" y "Actualidad" mediante rangos promedio, los cuales requieren rediseño urgente al ubicarse en niveles bajos del baremo. El valor $r = 4,46$ del ítem "Experticia" señala desconexión con el conocimiento de la población y el de "Actualidad", $r = 5,02$, es un claro indicador de obsolescencia temática. Los ítems "suficiencia" y "Pertinencia" se ubican en la categoría media-baja e indican que las preguntas y opciones deben revisarse y reformularse. En el otro extremo, "Anonimato y confidencialidad" ($r = 9,06$) y "Organización" son fortalezas susceptibles de ser replicadas en versiones posteriores del instrumento.

En cuanto al modelo estandarizado, este busca resolver las dificultades identificadas mediante un proceso estructurado en cuatro fases. La primera consiste en la definición de estándares claros y precisos para la selección de expertos. En la segunda fase se desarrolla el análisis cuantitativo, estableciendo umbrales de decisión como $W < 0,6$, que implica rediseño obligatorio, y rangos promedio inferiores a 7,5, que identifican ítems prioritarios por su criticidad. La tercera fase se enfoca en la retroalimentación cualitativa, en la cual se reformulan los ítems críticos con el fin de mejorar la calidad del instrumento. Finalmente, la cuarta fase contempla la validación de la versión revisada del instrumento, considerándose exitosa cuando se alcanza un coeficiente W igual o superior a 0,6.

Entre las ventajas inmediatas del modelo estandarizado propuesto, está la de evitar validaciones superficiales al integrar análisis cuantitativo (coeficiente de concordancia W de Kendall y prueba de hipótesis con chi-cuadrado) y cualitativo. Además, se puede sugerir este modelo como referencia para validar instrumentos de investigación, exigiendo $W \geq 0,6$ y rangos promedio mayores que 7,5 en todos los ítems. Sin embargo, su implementación demanda tiempo, recursos y disponibilidad sostenida de expertos, factores que podrían representar barreras operativas en investigaciones con cronogramas ajustados.

Como líneas de investigación futuras o sugerencias para estudios posteriores, entre otras se recomienda aplicar este modelo en otras disciplinas, como salud, psicología o ingeniería educativa, para analizar su adaptabilidad metodológica; estudiar y comparar la eficacia del coeficiente W de Kendall con la de otros estadísticos de concordancia; validar instrumentos en contextos geográficos y culturales diversos para examinar si el modelo propuesto mantiene su efectividad; analizar cómo varían el grado de concordancia y la estabilidad de los resultados según el número de expertos, lo cual puede ofrecer recomendaciones prácticas para equipos con recursos limitados; desarrollar simulaciones que permitan estimar el impacto de diferentes escenarios metodológicos sobre el coeficiente W, tales como variaciones en el número de ítems, tipo de escala, dispersión de respuestas, entre otras, para anticipar necesidades de rediseño.

Referencias

Barrueta, N., Peña, S. y Fernández, E. (2022). El estadígrafo Kendall y su aplicación. Un ejemplo práctico. A3Manos, *Revista de la Universidad Cubana de Diseño*, 9(16), 1-9. https://portal.amelica.org/ameli/journal/784/7843889004/7843889004.pdf DOI no disponible.

Cueva, T., Córdova, O., Gonzáles, J., Flores, F. y Balmaceda, C. (2023). *Métodos mixtos de investigación para principiantes*. Editorial INUDI PERÚ. https://doi.org/10.35622/inudi.b.106

Cvetkovic, A., Maguiña, J., Soto, A., Lama, J. y Correa, L. (2021). Estudios transversales. *Revista de la Facultad de Medicina Humana*, 21(1), 179-185. http://dx.doi.org/10.25176/rfmh.v21i1.3069

Escobar, J. y Cuervo, A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Revista Avances en Medición*, 6(1), 27-36. https://gc.scalahed.com/recursos/files/r161r/w25645w/Juicio_de_expertos_u4.pdf DOI no disponible.

Hernández, R. (2011). *Instrumentos de Recolección de Datos en Ciencias Sociales y Ciencias Biomédicas: Validez y Confiabilidad*. Universidad de los Andes, Venezuela.

Herrera, J., Calero, J., González, M., Collazo, M. y Travieso, Y. (2022). El método de consulta a expertos en tres niveles de validación. *Revista Habanera de Ciencias Médicas*, 21(1), 1-11. https://www.redalyc.org/journal/1804/180473621013/180473621013.pdf DOI no disponible.

Lao, Y., Pérez, M. y Marrero, F. (2016). Procedimiento para la selección de la comunidad de expertos con técnicas multicriterio. *Centro de Información y Gestión Tecnológica de Holguín*, 22(1), 1-16. https://www.redalyc.org/pdf/1815/181543577003.pdf DOI no disponible.

Maldonado, N. y Santoyo, F. (2024). Validesa de contingut per judici d'experts: integració quantitativa i qualitativa en la construcción d'instruments de mesura. *REIRE Revista d'Innovació i Recerca en Educació*, 17(2), 1-19. https://doi.org/10.1344/reire.46238

Marín, F., Pérez, J., Senior, A. y García, J. (2021). Validación del diseño de una red de cooperación científico-tecnológica utilizando el coeficiente K para la selección de expertos. *Revista Información Tecnológica*, 32(2), 79-88. http://dx.doi.org/10.4067/S0718-07642021000200079

Maskavizan, A., Poco, A. y Calzolari, A. (2023). Aspectos prácticos del uso del coeficiente de concordancia W de Kendall para el jueceo de cuestionarios en enfermería. Revista Arandu Poty, 2(2), 23-32. https://www.researchgate.net/publication/381655794_aspectos_practicos_del_uso_del_coeficiente_de_con-_cordancia_w_de_kendall_para_el_jueceo_de_cuestio-_narios_en_enfermeria DOI no disponible.

Walpole, R., Myers, R. y Myers, S. (2012). *Probabilidad y estadística para ingeniería y ciencias*. Editorial Pearson. Novena edición.

Ramírez, A. y Polack, A. (2020). Estadística inferencia. Elección de una muestra estadística no paramétrica en investigación científica. *Revista Horizonte de la Ciencia*, 10(19), 191-208. DOI: https://doi.org/10.26490/uncp.horizonteciencia.2020.19.597