ASPECTOS PRÁCTICOS DEL USO DEL COEFICIENTE DE CONCORDANCIA W DE KENDALL PARA EL JUECEO DE CUESTIONARIOS EN ENFERMERÍA

Practical aspects of the use of the concordance Kendall's coefficient W for the judging of nursing questionnaires

Maskavizan, A. J. [1,2], Poco, A. N. [1], Calzolari, A. [3*]

*Autor para correspondencia. Correo electrónico: aldo.calzolari@hospitalitaliano.org.ar, acalzolari30@yahoo.com.ar

[1] Universidad de Concepción del Uruguay, Concepción del Uruguay, Entre Ríos, Argentina.
[2] Becaria CONICET, Universidad Tecnológica Nacional, Facultad Regional Concepción del Uruguay, E. Ríos, Argentina.
[3] Instituto Universitario Hospital Italiano Buenos Aires, Departamento Posgrado, Argentina.

Recepción: 07/07/ 2023; Aceptación: 15/10/2023

RESUMEN

La adaptación de un cuestionario es un proceso necesario en diversas áreas disciplinares, que en forma creciente se está aplicando en la Enfermería. Siempre se plantea el interrogante acerca de la calidad y cantidad de las preguntas que lo conforman, su redacción o su contenido. Previo al análisis de validez de contenido, es necesario determinar el grado de acuerdo que existe entre un grupo de personas expertas que evalúan un instrumento. Para esto existen diferentes herramientas, entre las que se destaca el coeficiente de concordancia W de Kendall. En este trabajo se analizó el uso de este coeficiente para medir el acuerdo entre un grupo de personas evaluadoras, con detalle del procedimiento de cálculo mediante distintos softwares. Se describieron casos que pueden ocurrir en la práctica, con análisis de resultados de la prueba y la metodología para extraer conclusiones de ellos, de modo que no se requiera un profundo manejo estadístico para conceptualizar los resultados. Se muestra la utilidad del coeficiente, métodos para su cálculo de manera sencilla mediante los softwares PSPP® o SPSS® y un complemento del software Excel®. Asimismo, se proponen soluciones de interpretación que ocurren en la práctica de enfermería y se exponen las limitaciones que posee.

Palabras clave: Investigación en Enfermería; Diseño de Encuestas; Cuestionarios; Jueceo; Concordancia entre jueces; Evidencia de validez

ABSTRACT

The adaptation of a questionnaire is a necessary process in various disciplinary areas, which is increasingly being applied in Nursing. Concerns still arise about the quality and quantity of the questions that compose it, its formulation or content. Prior to the content validity analysis, it is necessary to determine the degree of agreement that exists among a group of experts who evaluate an instrument. There are different tools for this, among which Kendall's coefficient of concordance W stands out. In this paper, the use of this coefficient to measure agreement among a group of evaluators was analyzed, detailing the calculation procedure using different software. Cases that can occur in practice were described, with analysis of test results and the methodology to draw conclusions from them, so that deep statistical management is not required to conceptualize the results. It shows the utility of the coefficient, methods for its calculation in a simple way using PSPP® or SPSS® software and an Excel® software complement. Likewise, interpretation solutions that occur in nursing practice are proposed and its limitations are exposed.

Keywords: Nursing Research; Survey Design; Questionnaires; judge analysis; Concordance between judges; Evidence of validity

Se puede citar: Maskavizan, A. J., Poco, A. N., Calzolari, A. (2023). Aspectos prácticos del uso del coeficiente de concordancia W de Kendall para el jueceo de cuestionarios en enfermería. Arandu Poty; 2(2):23-32.

INTRODUCCIÓN

La validación de un instrumento (cuestionario u otro) es un proceso necesario en diversas áreas disciplinares. Gradualmente, su uso se ha ido incrementando en la disciplina de Enfermería, para enlace de turnos (Yáñez Corral & Zárate Grajales, 2016), rol materno (Santos-Díaz & cols., 2021), competencias de enfermería (Castañeda Hidalgo y cols., 2018), comportamientos de cuidado (Morales Castillo & cols., 2016), material didáctico (Velázquez Moreno, González Velázquez & Peña León, 2020), entre muchas otras. Existe siempre el interrogante acerca de la selección de las preguntas que lo conforman, su redacción o su contenido. En algunas situaciones, se desea aplicar un instrumento ya reconocido por la comunidad científica en otros contextos, en un idioma diferente o en disímiles marcos culturales, por lo que éste se debe traducir y adaptar al contexto. Este proceso implica disponer de elementos de juicio lo suficientemente confiables como para que el instrumento sea de utilidad en cada situación. Si el cuestionario está redactado en otro idioma, previo a su empleo es necesario realizar proceso de traducción, adaptación transcultural y retrotraducción (Beatty & cols., 2020). Si el cuestionario ya está traducido al idioma en que se lo va a emplear, es necesario llevar a cabo un proceso de revisión lingüística previo a disponer de evidencias de validez, según lo propuesto por la Comisión Internacional de Tests (Muñiz, Elosúa & Hambleton, 2013).

Beatty y cols. (2020) discuten diversas opciones y analizan las ventajas y desventajas de cada método de obtener dichas evidencias de validación. Uno es el método de revisión por un grupo de personas expertas en una determinada temática como medio de certificación de la potencia de una herramienta diseñada para medir una característica, llamado de jueces/juezas (Escobar-Pérez & Cuervo-Martínez 2008), que representa un sistema apto para un cuestionario breve. En ocasiones, sus resultados poseen evidencia de validez reducida, ya que el grupo puede no ser representativo y puede haber mayor liderazgo de las opiniones de una persona en detrimento de las de otras. Algunos autores/as no aplican ningún estadístico (Barrientos Sánchez, Hernández Cantoral & Hernández Zavala, 2013; Domínguez Martínez & Cortés Escárcega 2019; Uribe Romero & Rivas Espinosa, 2010). Esto puede representar una debilidad del instrumento y, por ende, de los resultados que se obtengan de su aplicación. Por el contrario, revisar el instrumento mediante la técnica de jueceo da confianza que el instrumento mide realmente lo que se pretende medir.

Una de las formas de aumentar la robustez de la consulta a jueces, para luego ser aplicada a diferentes poblaciones y disponer de evidencias de validez a las conclusiones extraídas del estudio, consiste en aplicar criterios de análisis estadístico, por medio de distintos coeficientes. Uno de estos análisis la validez de contenido o relevancia de los ítems, lo cual puede concretarse mediante el cálculo del coeficiente V de Aiken (Aiken, 1985).

Previo a este análisis, es necesario determinar el grado de acuerdo que existe entre los jueces y juezas expertos que han evaluado cada uno de los ítems que integran el instrumento. Para esto existen diferentes herramientas, entre las que se destaca el coeficiente de concordancia W de Kendall (Arízaga Piedra, Monge Loja & Muñoz Pauta, 2019). Este coeficiente informa el grado de asociación entre las puntuaciones que aporta cada juez (Siegel & Castellán, 1995), es decir, en qué medida las opiniones coinciden. No informa acerca de si una pregunta es clara o no lo es, únicamente si están de acuerdo (o no).

El coeficiente W se utiliza ampliamente en áreas de Enfermería (Antomás, Huarte & Gárriz, 2011; Goyzueta & Consuelo Cervantes, 2020; Padilla Reyes & Guerra Cari, 2016), Educación (Álvarez Gómez & cols., 2019; Matzumura Kasano & cols., 2018; Valdiviezo Rodríguez & Larzabal Fernández, 2017), Psicología (Zapata Zabala, Cárdenas Niño & Cuartas Arias, 2019) y también en otras, como Biología, Medicina, y, en general, en todas las disciplinas en las cuales las variables que se manejan no son cuantitativas y no responden a la estadística paramétrica (Bailón García & cols., 2015). Asimismo, resulta una herramienta muy útil para la validación de instrumentos y escalas de medición.

Desde su descripción inicial en 1938 por Kendall (1938) y en base a las dificultades que presenta su utilización y cálculo, hay artículos y libros que describen y clarifican su uso (entre otros Álvarez Cáceres, 1995; Howell, 2010; Legendre, 2005; Siegel & Castellán, 1995; Sprent & Smeeton, 2001). No obstante, estos documentos contemplan únicamente algunos aspectos del coeficiente W, que exige a quien los lee conocimiento matemático o estadístico, asumiendo un conjunto de saberes de los que muchas veces la persona usuaria carece. Al realizar trabajos de investigación con cuestionarios, la gran mayoría conoce en profundidad el tema específico sobre el cual se investiga. Sin embargo, uno de los

problemas es que hay menor familiaridad con el proceso estadístico necesario para que esa encuesta pueda ser utilizada en forma confiable, mediante procedimientos de validación. Hay dificultades de tipo disciplinar, como desconocimiento de criterios estadísticos o qué coeficientes conviene utilizar, dificultades en el manejo de paquetes de software e interpretación de los resultados de pruebas para medir evidencias de validez (Lissitz, 2009).

Por tanto, los objetivos de este artículo fueron describir los procedimientos para el cálculo del coeficiente W de modo sencillo, mostrar los casos que pueden ocurrir en la práctica y proveer una guía de interpretación de resultados.

METODOLOGÍA

Software: Se analizaron 10 softwares estadísticos, tanto de acceso gratuito como de pago: Epidat®, PSPP 1.2.0, SPSS® v25, InfoStat®, InStat+®, Medcalc® v19.2.6, Minitab® v18, Stata® v12 (sus referencias se encuentran en las respectivas páginas Web), Kendall W (Legendre, 2005) y el complemento RealStats® para Excel® (Zaiontz, 2020). No se empleó el programa R (https://www.r-project.org) debido a que para su uso es necesario disponer de nociones de programación, además de los conocimientos disciplinares y de estadística.

Análisis: Se prepararon planillas con resultados dispares de respuestas entre jueces/juezas, las que se analizaron con diferentes softwares para los casos presentados. Se determinó cuáles de los programas estudiados poseen la función para el cálculo del coeficiente W, con o sin empates, esto es, valores repetidos en las puntuaciones. Los pasos a seguir para el cálculo de este coeficiente solamente se describen para el software gratuito PSPP y su equivalente el software de pago SPSS®, sí como el complemento gratuito *Realstat* del Excel® (Zaiontz, 2020). Para el resto de los softwares, solo se revisó si disponían o no de la función, la que debe ser empleada siguiendo los manuales correspondientes.

Determinación del coeficiente de W con PSPP o SPSS® para cada pregunta del instrumento:

a. Carga de datos: PSPP se puede descargar de modo gratuito en español. SPSS® es de pago y pueden obtenerlo en español o inglés. Con cualquiera de los dos softwares, abrir un archivo nuevo y desde la pestaña Vista de Datos (View Data) (ubicada abajo a la izquierda en la pantalla) copiar los datos de una planilla Excel® o cargarlos en forma manual. Esto genera de forma automática las variables o características a evaluar en forma de columnas.
b. Desde la pestaña Vista de Variables (Variable View) es posible modificar las especificaciones de las variables. En la columna Nombre (Name) puede escribirse una abreviatura y en el recuadro Etiqueta (Tag) se puede indicar el nombre completo de la misma, si éste es muy extenso. En la columna Tipo (Type) se selecciona Numérico (Numeric) y en la columna Medida (Measure) se debe seleccionar una, usualmente Escala (Scale). La primera variable contendrá la identificación de quienes evalúan. Para la primera columna el Tipo es Cadena (Chain) y la Medida es Nominal (Nominal).
c. Cálculo del coeficiente W: En el menú de la parte superior de la pantalla de PSPP Analizar – Pruebas No-Paramétricas, se elige K Muestras Emparejadas (en SPSS® se accede a través del menú Analizar – Pruebas No-Paramétricas – Cuadros de diálogo antiguos y se elige K muestras relacionadas). Se abre una ventana en la cual se deben incorporar las variables al sistema, siguiendo la dirección de la flecha, hacia el recuadro Variables de test (o de prueba). Una vez cargadas las variables a analizar, se activa la opción W de Kendall y se selecciona OK (Aceptar, [Accept] en SPSS®). Los comandos pueden variar ligeramente según la versión del software instalada.
d. En una nueva ventana emergente, denominada Visor de resultados se muestran las variables y sus rangos promedios. Debajo, se observan las Pruebas Estadísticas (en SPSS®, los Estadísticos de prueba) N: la cantidad de jueces, el coeficiente W de Kendall, el valor del estadístico Chi-cuadrado χ^2, los grados de libertad (gl o df – Degrees of Freedom) y la Significación Asintótica (Asintotic Significance) o p-valor. Los datos del coeficiente W y el p-valor se copian del programa para ser guardados en una planilla de cálculo o documento.

Determinación del coeficiente de W con el complemento RealStats®: las instrucciones corresponden a la instalación del complemento para Excel® 2007 o posterior (Valdiviezo-Rodríguez & Larzabal-Fernández, 2017). En el programa, se selecciona *Archivo/Opciones* y luego *Complementos/Administrar: Complementos de Excel*. El complemento *Solver* debe estar activado para que el paquete RealStats® funcione correctamente. Esta misma ventana permite instalarlo mediante *Examinar,* localizar el archivo *xrealstats.xlam* descargado y *Aceptar*. Se requiere que tanto el complemento *Solver* como *Xrealstats* queden seleccionados. Para poder utilizar el complemento se deberá reiniciar Excel®.

En todos los casos, el valor declarado de significación fue de 0,05.

Jueceo:

Se realizaron ensayos con 8 jueces (designados A-H) para 4 categorías de análisis: (a) Coherencia, que evalúa si el ítem tiene una relación lógica con la dimensión o indicador que está midiendo. (b) Relevancia, que determina si el ítem es indispensable para interpretar el fenómeno y, por tanto, si debe ser incluido o no en el instrumento. (c) Claridad, que hace referencia a la facilidad de comprensión del ítem, es decir, si este es adecuado de forma sintáctica y semántica y (d) Suficiencia, que evalúa si todos los ítems incluidos en una misma dimensión bastan para obtener la medida de esta (Escobar-Pérez & Cuervo-Martínez, 2008).

DESARROLLO

Se presentan casos que pueden aparecer en la práctica en Enfermería u otras áreas disciplinares, tomando como ejemplo ocho personas expertas (A-H) y cuatro variables (Suficiencia-Claridad-Coherencia-Relevancia), con un determinado número de preguntas. Cada variable debe ser puntuada de 1 (mínimo) a 4 (máximo).

1. Caso con concordancia significativa

El coeficiente W varía entre 1 (máxima concordancia) y 0 (nula concordancia). El cálculo para la pregunta de la Tabla 1, con 8 jueces y 4 preguntas, siguiendo los pasos descriptos en Metodología con el software PSPP o SPSS® resultó de 0,67 con un p-valor de 0,001. Al margen del valor del coeficiente W, el *p*-valor menor que el nivel de significación convenido de 0,05, indica que existe acuerdo entre las respuestas dadas por quienes evaluaron el cuestionario.

Note que cada columna tiene cierto patrón de uniformidad: la columna "Relevancia", por ejemplo, con 5 valores máximos (4) y 2 valores altos (3); o la columna "Claridad", con 5 valores muy bajos (1) y 2 valores bajos (2). En ambos casos, se percibe un cierto grado de "acuerdo": quienes realizaron el jueceo otorgan valoraciones similares. Los coeficientes de concordancia miden justamente esto: grado de acuerdo entre pares; no si las respuestas son elevadas o bajas (para ello se utilizan otros tests, como la V de Aiken (Aiken, 1985).

Tabla 1. Caso de aplicación del coeficiente de W de Kendall con concordancia significativa (W= 0,67; p=0,001).

Juez/a	Suficiencia	Claridad	Coherencia	Relevancia
A	4	1	2	4
B	3	2	3	4
C	4	3	2	3
D	3	1	3	4
E	4	1	2	2
F	4	2	3	4
G	4	1	2	4
H	2	1	3	3

Si bien en el ejemplo de la Tabla 1 se muestran los resultados para una sola pregunta, este procedimiento debe realizarse para cada una de las preguntas del instrumento.

Schmidt propone la utilización de una tabla para interpretar el grado de consenso (Schmidt, 1997). Sugiere interpretar el coeficiente W de la siguiente manera: para valores del coeficiente W de Kendall de 0,1, el acuerdo es muy débil, para 0,3, es débil, con 0,5, es moderado, 0,7 indica un alto grado de acuerdo y 0,9, un acuerdo inusualmente alto. En general, se acepta que valores de W iguales o superiores a 0,8 son indicativos de concordancia, sin necesidad de observar el *p*-valor. Por otro lado, en parte basados en dicha tabla, hay artículos que han expresado concordancia con valores del coeficiente W menores, de hasta 0,5, sin hacer uso del *p*-valor (Okwesili, Mazzuchi & Sarkani, 2016). La Tabla 2 muestra ejemplos de artículos, de Enfermería y otras disciplinas, donde se han utilizado diversos valores de corte para aceptar la concordancia empleando este coeficiente. Como se observa, hay ejemplos con W relativamente elevados, con *p*-valor significativo (renglones 1-3). Los renglones 4-6 muestran un valor "moderado" según Schmidt (1997) y significación estadística, en tanto que el ejemplo mostrado en el renglón 7 señala un valor más cercano a 0,5 con un *p*-valor no significativo. En este contexto, obtener un valor de W de 0,41 (renglón 8) y considerarlo como indicador de concordancia débil a moderada sin evaluar el *p*-valor correspondiente, puede inducir a una conclusión incompleta o errónea. Los últimos cuatro renglones muestran valores decrecientes de W, incluso uno muy pequeño, de solo 0,068, todos estadísticamente significativos.

Consideramos que el empleo del *p*-valor como indicador de la fortaleza del coeficiente W otorga una mayor confiabilidad al proceso de validación

Tabla 2. Valor de coeficiente de concordancia W de Kendall y *p*-valor de diversas fuentes.

	Disciplina	N°*	WK**	p	Referencia
1	Psicología	5	0,966	0,000	Zapata Zabala, Cárdenas Niño & Cuartas Arias, 2019
2	Enfermería	6	0,806	0,000	Goyzueta & Consuelo Cervantes, 2020
3	Enfermería	13	0,755	0,000	Padilla Reyes & Guerra Cari, 2016
4	Educación	30	0,638	< 0,05	Álvarez Gómez y cols., 2019
5	Psicología	14	0,63	0,000	Arízaga Piedra, Monge Loja & Muñoz Pauta, 2019
6	Enfermería	9	0,625	0,030	Antomás, Huarte & Gárriz, 2011
7	Educación	5	0,478	0,126 NS	Valdiviezo-Rodríguez & Larzabal-Fernández, 2017
8	Ingeniería	15	0,41	ND	Okwesili, Mazzuchi & Sarkani, 2016
9	Ingeniería	4	0,366	0,034	Bailón García y cols., 2015
10	Medicina	10	0,28	< 0,05	Abdel-Tawab y cols., 2011
11	Veterinaria	9	0,11	0,04	Bracke & cols., 2008
12	Educación	7	0,068	< 0,05	Parejo Jiménez & cols., 2020

(*) Número de jueces/as o rangos. (**) W de Kendall. ND: No determinado; NS: No significativo.

de un cuestionario. En efecto, puede haber valores elevados o bajos de W, que presenten un nivel de significación tal que permita concluir que existe asociación en los puntajes otorgados. En estos casos, además, queda a criterio del grupo de investigación resolver qué hacer ante valores de W bajos, pero estadísticamente significativos. De León García & cols. (2016) plantean dos condiciones para la determinación de la concordancia, a través de la evaluación del *p*-valor con un nivel de significancia establecido por convenio de 0,05 y un valor del coeficiente W superior a 0,4 para asegurar concordancia moderada. Por tal motivo, a pesar de obtener un *p*-valor menor a 0,05, elimina aquellas variables en las cuales no se alcanzaba el consenso preestablecido. En este caso particular, la intención del grupo investigador era eliminar variables en las cuales no parecía haber demasiado consenso y por eso la inclusión del criterio del coeficiente W mayor a 0,4. Para otros casos, en cambio, el *p*-valor menor a 0,05 es criterio suficiente, sin importar el valor de W.

2. Caso con baja concordancia

Para el caso del ejemplo de la Tabla 3, se realizaron los mismos pasos que en el ejemplo de la Tabla 1. El coeficiente W resultó de 0,18, y el *p*-valor, 0,228, no significativo. Note cómo, en contraposición a lo mostrado para las columnas de la Tabla 1, aquí las puntuaciones aparecen dispares. En este caso, el nivel de acuerdo es bajo y, por lo tanto, se debe tomar una decisión al respecto, debido a que no existe una recomendación unánime de quienes emitieron un juicio sobre cómo abordar esta problemática.

Tabla 3. Caso de aplicación del coeficiente de W de Kendall con baja concordancia (W=0,18; p=0,228).

Juez/a	Suficiencia	Claridad	Coherencia	Relevancia
A	3	1	2	2
B	3	2	3	1
C	2	3	1	3
D	3	3	3	4
E	4	1	2	2
F	4	2	1	2
G	4	3	2	1
H	2	1	3	3

Okwesili, Mazzuchi y Sarkani proponen eliminar aquellos datos en los cuales no se tiene consenso por parte de quienes lo evaluaron (Okwesili, Mazzuchi & Sarkani, 2016). En otro enfoque, Chaves Torres (Chaves Torres & cols., 2016) plantea reenviar todas las respuestas obtenidas al mismo grupo, para que en una segunda ronda de evaluación expresen si consideran mantener la puntuación otorgada o lo modifican, al conocer las respuestas de las demás personas evaluadoras. Por último, Blanco-Sánchez (2014) y Dorantes-Nova (2016), sugieren revisar las preguntas en las cuales no se registra acuerdo e intentar construir nuevas preguntas para modificarlas y mejorarlas. En nuestra opinión, la persona que investiga debe conocer en profundidad el tema y los aspectos que se pretende que el instrumento mida. Por tal motivo, considerando su

experiencia previa, tiene la capacidad de determinar si, ante una falta de acuerdo, es posible eliminar la pregunta que se encuentra bajo discusión. En caso de que sea importante mantener esta pregunta en el cuestionario, debe considerar la aplicación de modificaciones sugeridas por el grupo que lo evaluó para mejorar el ítem.

3. Caso con alta concordancia y coeficiente W no significativo

En ocasiones, para ciertas preguntas, el análisis preliminar de los puntajes otorgados permite suponer que existe una elevada coincidencia entre las opiniones de las personas evaluadoras. Sin embargo, al efectuar el cálculo del coeficiente W, los resultados que generalmente se observan muestran que la significancia estadística es baja (Tabla 4). Como se observa en el ejemplo, quienes han analizado la pregunta, han otorgado valoraciones similares. Por tal motivo, mediante una observación simple de los resultados, es posible afirmar que existe un gran consenso. Sin embargo, el cálculo del coeficiente W, mediante el programa PSPP, devuelve W = 0,17 y un *p*-valor de 0,261. Los mismos valores se obtienen con el empleo de otros softwares.

Tabla 4. Caso de aplicación del coeficiente de W de Kendall con alta concordancia no significativa. (W = 0,17; p= 0,261.)

Juez/a	Suficiencia	Claridad	Coherencia	Relevancia
A	4	4	2	4
B	3	4	4	4
C	4	4	4	4
D	3	4	4	4
E	4	4	4	4
F	4	4	4	4
G	4	4	4	4
H	4	4	3	4

Estos resultados indican que no existe acuerdo, lo que constituye una contradicción con la observación de los datos. Se observa alta concordancia, pero el coeficiente arroja resultados negativos.

Esta situación se conoce como la paradoja de Kappa, dado que se estudió originalmente con el coeficiente Kappa de Cohen (Gwet, 2008). Cuando un rasgo es muy raro u omnipresente (como la gran cantidad de "4" en la Tabla 4), el coeficiente W refleja un bajo grado de acuerdo, sin significación estadística. En otras palabras, si la prevalencia de un rasgo es muy baja o muy alta, este coeficiente no se puede aplicar para determinar el nivel de acuerdo.

En la bibliografía no hay una propuesta clara para resolver esta situación. En dos ejemplos en nuestro equipo, se obtuvo esta paradoja en un estudio sobre la adaptación de una encuesta sobre percepción del cuidado enfermero en Ecuador (Franco Poveda & Calzolari, datos no publicados) y en un estudio sobre la adaptación de un cuestionario sobre escritura académica para diversas carreras, incluida Enfermería (Vanegas Quizhpy, Maskavizan & Calzolari, enviado a publicar). Ante esta situación, en ambos casos se ha optado por complementar el análisis del coeficiente W con otros coeficientes de concordancia, como el coeficiente AC_1 y AC_2 de Gwet (2008) o el coeficiente AD (Burke & Dunlap, 2002).

En otro caso real, en este caso de aplicación del coeficiente W de Kendall en un estudio de desarrollo de un test para predicción temprana de riesgo de dislexia (Arista, Fonseca & Calzolari, 2023), 6 de las 8 preguntas tuvieron coeficientes superiores a 0,667 y p<0,05. Dos de las preguntas, con altísima concordancia en las opiniones de jueces/juezas, mostraron W de 0,333 y 0,417, con valores no significativos de p (0,416 y 0,283 respectivamente). Nuestra decisión fue aplicar el sentido común y dar a esas preguntas como concordantes por parte de quienes hicieron el jueceo. El análisis por el coeficiente AC_1 de Gwet para estas 2 preguntas mostró que eran concordantes.

4. Caso con concordancia máxima

La concordancia máxima se puede dar en dos casos: (a) cuando quienes evalúan la pregunta coinciden en otorgarle la misma puntuación a todas las categorías o (b) coindicen en las puntuaciones otorgadas para cada criterio analizado, es decir, todos los valores de cada columna son iguales entre sí.

En el primero de los casos, (a) el software no brinda un valor del coeficiente W, dado que la fórmula para su cálculo resulta en una indeterminación. Sin embargo, esto puede salvarse de una manera sencilla, dado que se puede observar que el acuerdo es absoluto. Se apela aquí al sentido común: el coeficiente W no puede aplicarse, pero hay máximo acuerdo en las puntuaciones.

En el caso del segundo ejemplo, (b) el resultado obtenido corresponde a W = 1,000 y p = 0,000. Este resultado es el máximo valor del coeficiente que puede obtenerse, con lo cual se concluye que el grado

Tabla 5. Características de software en relación al cálculo de coeficiente de concordancia de Kendall.

Software	Licencia paga	Solución	Observaciones
SPSS® v25	Sí	Completa	Lectura de los resultados muy sencilla. Permite el cálculo del coeficiente con y sin empates. Permite importar datos desde una hoja de cálculo.
PSPP	No	Completa	El procedimiento es el mismo que para SPSS® v25.
Openstat®	No	Parcial	No permite hallar el coeficiente para datos con empates en las puntuaciones otorgadas.
Kendall's W	No	Completa	Es complejo de utilizar ya que su interfaz es anticuada. Está disponible en inglés o francés.
Stata®	Sí	Parcial	Requiere instalar el comando snp2. No permite hallar el coeficiente para datos con empates en las puntuaciones otorgadas.
Complemento para Excel® RealStats®	No	Completa	Sencillo de incorporar a Excel®. Permite el cálculo del coeficiente con o sin empates. No ordena los datos en forma de rangos de manera automática y hay que cargarlos de modo manual.
Epidat®	No lo calculan		
InfoStat®			
InStat+®			
Medcalc®			
Minitab®			

de acuerdo es máximo.

5. Análisis de software

Si bien para el análisis de los casos presentados se utilizaron los software PSPP e SPSS®, se describen a continuación un conjunto de programas que fueron evaluados para su utilidad en el cálculo del coeficiente W (Tabla 5).

De este análisis se desprende que el uso de SPSS® o PSPP resultan los más sencillos, puesto que calculan el coeficiente W de manera automática y completa. Entre las ventajas de estos programas se pueden citar: la introducción de los valores de las características evaluadas de modo directo, el cálculo automático de los rangos y la posibilidad de importar valores de una hoja de cálculo como Excel®. Asimismo, permite resolver casos de evaluaciones con y sin empates o valores repetidos en los rangos, y no es necesario identificar procesos de cálculo diferentes. También indica directamente el valor de la variable χ^2 de prueba según los grados de libertad de cada experiencia y presenta el *p*-valor para mostrar el grado de significación de la prueba.

6. Cálculo del coeficiente W mediante el complemento para Excel® RealStats®

Para el cálculo del coeficiente W, los datos deben ordenarse en rangos. Esto es, hay que identificar la posición que ocupa cada valor asignado a los atributos para calcular los rangos. En el caso que existan empates en alguno de ellos, se deberán promediar las posiciones que ocupan los valores empatados, y reemplazar los valores de las puntuaciones empatadas por dicho promedio.

En la Tabla 6 se muestran los mismos datos de la Tabla 1 ordenados en rangos. Para el caso del juez B, por ejemplo, los valores de la Tabla 1 fueron 3, 2, 3 y 4 para las 4 columnas, respectivamente. El 2 es el valor más bajo y le corresponde el 1; el 4 el valor más alto y le corresponde el 4; y hay dos valores están empatados en 3; como a estos les corresponde la 2da y la 3ra posición, el promedio de ambas es 2,5. De modo similar, para los datos d la columna H, los datos de la Tabla 1 fueron 2, 1, 3 y 3 respectivamente. Los valores 1 y 2 ocupan las posiciones 1 y 2; y hay 2 valores empatados en 3. Como a estos dos les corresponde ocupar las posiciones 3ra y 4ta, el promedio es 3,5.

Para calcular el coeficiente W se ingresa al menú del complemento RealStats® (mediante la combinación Ctrl+m), luego a la pestaña *Correlation and Reliability, Interrater Reliability*, y, finalmente, se selecciona la opción *Kendall's W with ties*, seleccionando los datos de las puntuaciones ordenadas en rangos. Los resultados que se obtienen son los mismos que los mostrados para la Tabla 1 con el software PSPP.

Tabla 6. Caso de aplicación del coeficiente de W de Kendall con datos ordenados en rangos para análisis mediante RealStats® en Excel®.

Juez/a	Suficiencia	Claridad	Coherencia	Relevancia
A	3,5	1	2	3,5
B	2,5	1	2,5	4
C	4	2,5	1	2,5
D	2,5	1	2,5	4
E	4	1	2,5	2,5
F	3,5	1	2	3,5
G	3,5	1	2	3,5
H	2	1	3,5	3,5

DISCUSIÓN

En este trabajo, se analizó el uso del coeficiente de concordancia W de Kendall para medir el acuerdo entre un grupo de jueces/juezas que evalúan un instrumento como un cuestionario o test. Se detalló el procedimiento de cálculo mediante los softwares PSPP, SPSS® y el complemento RealStat® de Excel® y se describieron distintos casos que pueden ocurrir en la práctica. Además, se analizaron los posibles resultados de la prueba y se detalló la metodología para extraer conclusiones de ellos, de modo que no se requiera un profundo manejo estadístico para conceptualizar los resultados del cálculo de este coeficiente.

Se encontró que existen distintos softwares disponibles para el cálculo de este coeficiente (Tabla 5) pero que no todos contemplan el caso con existencia de empates entre las puntuaciones otorgadas, siendo PSPP y SPSS® los más adecuados y completos. El complemento RealStats® (Zaiontz, 2020) presenta la ventaja de permitir su cálculo desde la planilla de cálculo de Excel® en la cual se cargaron los datos. Sin embargo, es necesario determinar los rangos a partir de las valoraciones otorgadas de forma manual, como se señala en la Tabla 6. El software gratuito Kendall's W (Legendre, 2005) también permite el cálculo del coeficiente W en forma completa, aunque su interfaz es menos amigable que PSPP o SPSS®.

Los ejemplos mostrados pusieron de manifiesto que pueden presentarse diversas situaciones en la práctica: acuerdo bajo, alto o máximo (Aiken, 1985; Arízaga Piedra, Monge Loja & Muñoz Pauta, 2019; Escobar-Pérez & Cuervo-Martínez 2008; Siegel & Castellán, 1995).

Este coeficiente presenta fortalezas y debilidades. Entre las primeras, se destaca la valiosa información que proporciona, ya que señala el grado de acuerdo entre los puntajes asignados a las características evaluadas de un cuestionario. De este modo, brinda un fundamento para estimar la validez de un instrumento en un trabajo de investigación en Enfermería.

Asimismo, es posible efectuar un análisis estadístico asociado a este coeficiente a través del cálculo de *p*-valor y su comparación con un nivel de significación preestablecido. Esto le confiere mayor precisión y confiabilidad que otros, como, por ejemplo, el Índice de Acuerdo, que no tiene una prueba de significación asociada (Escurra, 1988). Es también, más sencillo de calcular que el coeficiente AC_1 o AC_2 de Gwet (Gwet, 2008).

Este coeficiente es ampliamente utilizado en Enfermería (Antomás, Huarte & Gárriz, 2011; Goyzueta & Consuelo Cervantes, 2020; Padilla Reyes & Guerra Cari, 2016) y otras áreas del conocimiento para estimar validez de instrumentos.

Como se mostró anteriormente al analizar los casos que pueden ocurrir en la práctica (Tablas 1, 3 y 4) es posible obtener diferentes valores del coeficiente W y su correspondiente *p*-valor. A partir de los resultados presentados (Tabla 2), se concluye que el análisis del *p*-valor es muy importante para considerar si existe o no acuerdo.

En contraposición, existen situaciones en las que se identifica la llamada paradoja de Kappa (Gwet, 2008). En estos casos, en los cuales el acuerdo es muy elevado o máximo, como se describió en la Tabla 4 y en la sección 4, el coeficiente W no es de utilidad. El sentido común señala que, al existir una elevada similitud en los valores asignados por quienes la evaluaron, la pregunta o criterio de análisis se acepta. No obstante, en caso de necesitar disponer de una estimación numérica del grado de acuerdo, es necesario recurrir a otros coeficientes como el AD (Burke & Dunlap, 2002) o el AC_1 y AC_2 de Gwet (2008). Para la comprensión del proceso matemático involucrado para el cálculo del coeficiente W, sugerimos leer a Siegel y Castellán (1995), Pérez Tejada (2008) y Sprent y Smeeton (2001).

El coeficiente W solamente permite concluir sobre el grado de acuerdo existente entre los rangos. En el caso de la validación de un instrumento, si esta prueba no indica un nivel de concordancia aceptable, el grupo de investigación debe tomar una decisión al

respecto y luego de un exhaustivo análisis, definir si mantiene, modifica o elimina la pregunta.

En los casos en los cuales la concordancia es aceptable, se debe completar el estudio de validez de contenido de los ítems del cuestionario o encuesta a través de la aplicación de otras pruebas. Por ejemplo, se puede utilizar el coeficiente V de Aiken (1985) para determinar si la pregunta es clara, pertinente, relevante y suficiente. En una publicación posterior se discutirá el empleo de este coeficiente.

CONCLUSIONES

En este artículo se muestra la utilidad del coeficiente de concordancia W de Kendall para estudios en Enfermería y otras áreas disciplinares, métodos para su cálculo de manera sencilla mediante los softwares PSPP, SPSS® y el complemento RealStats® para Excel®. Asimismo, se proponen metodologías para resolver diversas situaciones que ocurren en la práctica y se exponen las limitaciones que posee este coeficiente.

CONTRIBUCIÓN AUTORAL

A.J.M. participó en Metodología, Ejecución de la investigación, Validación, Análisis de los datos, Escritura manuscrito inicial. A.N.P. participó en análisis de los datos, análisis de softwares. A.C. participó en Conceptualización, Escritura manuscrito inicial. Mejora y finalización del manuscrito. Todas/os aprobaron la versión final.

AGRADECIMIENTOS

Agradecemos al Dr. Guillermo MacBeth por su aporte en el software Kendall W. No hay trabajo con personas humanas ni instituciones, no hay conflicto de intereses y no se ha recibido ningún financiamiento para el desarrollo del trabajo. A.J.M. tiene una beca doctoral del Consejo de Investigaciones Científicas y Técnicas (CONICET).

REFERENCIAS BIBLIOGRÁFICAS

Abdel Tawab, R., James, D. H., Fichtinger, A., Clatworthy, J., Horne, R. & Davies, G. (2011). Development and validation of the Medication-Related Consultation Framework (MRCF). Patient Education and Counseling; 83(3):451-457. doi: 10.1016/j.pec.2011.05.005

Aiken, R. L. (1985). Three coefficients for analyzing the reliability and validity of ratings. Educational and Psychological Measurement; 45(1):131-142. DOI: 10.1177/0013164485451012

Álvarez Cáceres, R. (1995). Estadística multivariante y no paramétrica con SPSS: Aplicación a las ciencias de la salud. Madrid: Ediciones Díaz de Santos

Álvarez Gómez, G., Romero Fernández, A. J., Castro Sánche,z F. & Comas Rodríguez, R. (2019). Valoración de expertos sobre la viabilidad de un modelo didáctico para la formación de modos de actuación interdisciplinarios. Revista Espacios; 40(44):24-32. https://www.revistaespacios.com

Antomás, J., Huarte, S. & Gárriz, C. (2011). Registros de enfermería sobre el sueño y percepción de los pacientes en una unidad psiquiátrica. Estudio comparativo. An. Sist. Sanit. Navar; 34 (3):373-381.

Arista, G. A., Fonseca, L. E. & Calzolari, A. (2023). Rdislex: test rápido para riesgo de dislexia en niños prelectores y lectores iniciales. Prueba piloto. Journal of Applied Cognitive Neuroscience, 4(1), e00354813. https://doi.org/10.17981/JACN.4.1.2023.02

Arízaga Piedra, T. K., Monge Loja, D. A. & Muñoz Pauta, M. A. (2019). Adaptación lingüística de la Escala de Desarrollo Armónico 0 - 5 años. Revista Cubana de Educación Superior; 38(4):32-53. http://www.rces.uh.cu/index.php/RCES

Bailón García, T. I., Parroquín Amaya, P., Romero López, R., Alvarado Iniesta, A., Canales Valdiviezo, I. & García Heredia, F. J. (2015). Diseño y validación de un instrumento para medir el impacto de los factores logísticos en la competitividad de las PyMEs. Cultura Científica y Tecnológica; 12(57):344-351. https://erevistas.uacj.mx

Barrientos Sánchez, J., Hernández Cantoral, A. & Hernández-Zavala, M. (2013). Adaptación y validación de un instrumento de valoración de riesgo de caída en pacientes pediátricos hospitalizados. Enfermería universitaria, 10(4), 114-119.

Beatty, P. C., Collins, D., Kaye, L., Padilla, J. L., Willis, G. B. & Wilmot A. (2020) Advances in questionnaire design, development, evaluation, and testing. Hoboken: JohnWiley y Sons, Inc.

Blanco Sánchez, J. P. (2014). Validación de una escala para medir la habilidad de cuidado de cuidadores. Aquichan; 14(3):351-363. DOI: 10.5294/aqui.2014.14.3.7.

Bracke, M. B. M., Edwards, S. A., Engel, B., Buist, W. G. & Algers, B. (2008). Expert opinion as 'validation' of risk assessment applied to calf welfare. Acta Veterinaria Scandinavica; 50:29. doi:10.1186/1751-0147-50-29

Burke, M. J. & Dunlap, W. P. (2002). Estimating interrater agreement with the Average Deviation index: a user's guide. Organizational Research Methods; 5(2):159-172. doi: 10.1177/1094428102005002002

Castañeda Hidalgo, H., Vega Infante, E., González, F., Rangel, S., Flores, F. & Peñarrieta de Cordova, I. (2018). Validación del instrumento: Competencias del personal de enfermería en Salud Pública, en el contexto mexicano. Enfermería Universitaria; 15(4): 361-369. http://dx.doi.org/10.22201/eneo.23958421e.2018.4.535

Chaves Torres, N. M., Echeverri Sarmiento, J. E., Ballesteros, D. A., Quijano Rodríguez, J. & Camacho, D. (2016). Validación de la escala de Morisky de 8 ítems en pacientes con enfermedad renal

crónica. Revista Med; 24(2):23-32. DOI: 10.18359/rmed.2627

De León García, F. E., Cabello Guerrero, A., Armas González, V. I. & Molina Morejón, V. M. (2018). Gestión de la cadena de suministro en empresa de manufactura avanzada. VinculaTegica Efan; 3(3):41-47. http://www.web.facpya.uanl.mx/vinculategica/

Domínguez Martínez, V. & Cortés Escárcega, I. (2019). Adaptación cultural y validación de Parental Stressor Scale: Neonatal Intensive Care Unit en padres mexicanos. Enfermería Universitaria; 16(4): 339-350. https://doi.org/10.22201/eneo.23958421e.2019.4.641

Dorantes Nova, J. A., Hernández Mosqueda, J. S. & Tobón Tobón, S. (2016). Juicio de expertos para la validación de un instrumento de medición del sindrome de burnout en la docencia. Ra Ximhai; 12(6):327-346. DOI: 10.35197/rx.12.01.e3.2016.22.jd

Escobar Pérez, J. & Cuervo Martínez, A. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. Avances en Medición; 6:27–36. http://www.humanas.unal.edu.co/psicometria/revista

Escurra, L. M. (1988). Cuantificación de la validez de contenido por criterio de jueces. Revista de Psicología; 6:103-111.

Goyzueta, A. P. & Consuelo Cervantes, S. (2020). Nivel de autocuidado de los pies en pacientes diabéticos de un hospital de Lima Norte. Health Care & Global Health; 4(2):56-64. DOI: 10.22258/hgh.2020.42.81

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology; 61:29-48. DOI: 10.1348/000711006X126600

Howell, D. C. (2010). Statistical methods for Psychology, 7ma ed. Belmont, Estados Unidos: Wadsworth: Cengage Learning.

Kendall, M. G. (1938). A new measure of rank correlation. Biometrika; 30(1-2):81-93. DOI: 10.1093/biomet/30.1-2.81

Legendre, P. (2005). Species associations: the Kendall coefficient of concordance revisited. Journal of Agricultural, Biological, and Environmental Statistics; 10(2):226-245. DOI: 10.1198/108571105X46642

Lissitz, R. W. (2009: The concept of validity: revisions, new directions, and applications. Information Age Publishing Inc.: Charlotte.

Matzumura Kasano, J. P., Gutiérrez Crespo, H., Zamudio Eslava, L. A. & Zavala Gonzales, J.C. (2018). Aprendizaje invertido para la mejora y logro de metas de aprendizaje en el Curso de Metodología de la Investigación en estudiantes de universidad. Educare; 22(3):1-21. DOI: 10.15359/ree.22-3.9

Morales Castillo, F. A., Hernández Cruz, M. C., Rodríguez, M. & Landeros Olvera E. A. (2016). Validación y estandarización del instrumento: Evaluación de los comportamientos de cuidado otorgado en enfermeras mexicanas. Enfermería Universitaria; 13(1): 3-11. http://dx.doi.org/10.1016/j.reu.2015.11.005

Muñiz, J., Elosúa, P. & Hambleton, R. K. (2013). Directrices para la traducción y adaptación de los tests: segunda edición. Psicothema; 25(2):151-157. DOI: 10.7334/psicothema2013.24

Okwesili, P., Mazzuchi, T. & Sarkani, S. (2016). Risk assessment using paired comparison expert judgment for ranking of compounding outsourcing facilities. IEEE Engineering Management Review; 44(1):47-56. DOI: 10.1109/EMR.2016.2530646

Padilla Reyes, J. M. & Guerra Cari, C. L. (2016). Patrones de comportamiento y su relación con el síndrome de burnout en los profesionales de enfermería del STAE – EsSalud Lima – Perú. Tesis. Universidad de San Pedro, Chimbote; http://200.48.38.121/handle/USANPEDRO/12365

Parejo Jiménez, N., Expósito López, J., García Garnica, M., Parra González, M. E. & Olmedo Moreno E. M. (2020). Construcción y validación de un instrumento para la evaluación de la competencia docente en la atención educativa del alumnado menor extranjero no acompañado. Revista Espacios; 41(20):115-129.

Pérez Tejada, H. E. (2008). Estadística para las ciencias sociales, del comportamiento y de la salud, 3ra ed. México D.F.: Cengage Learning.

Santos Díaz, M. L., Pérez Calderón, D. M., Lozada Pérez, M., Ramírez Girón, N. & Landeros Olvera P. (2021). Validez y confiabilidad de la Escala de Adopción al Rol Materno en madres adolescentes mexicanas. Enfermería universitaria; 18(1): 69-80. https://doi.org/10.22201/eneo.23958421e.2021.1.880

Schmidt, R. C. (1997). Managing Delphi surveys using nonparametric statistical techniques. Decision Sciences; 28(3):763-774. DOI: 10.1111/j.1540-5915.1997.tb01330.x

Siegel, S. & Castellán, N. J. (1995). Estadística no paramétrica: aplicada a las ciencias de la conducta, 4ta ed. México: Trillas. Cap. 7.

Sprent, P. & Smeeton, N. C. (2001). Applied nonparametric statistical methods, 3ra ed. Boca Ratón:CRC Press.

Uribe Romero, L. R. & Rivas Espinosa, J. G. (2010). Percepción de los Alumnos de Enfermería sobre el Desempeño docente en la Práctica Clínica. Enfermería Universitaria; 7(4): 29-34.

Valdiviezo Rodríguez, D. C. & Larzabal Fernández, A. (2017). Elaboración de un manual mediante el método Delphi para la enseñanza de patronaje. Ecos De La Academia; 5:2015-214. http://revistasojs.utn.edu.ec

Velázquez Moreno, E., González Velázquez, M. S. & Peña León B. (2020). Diseño y validación de material didáctico para una intervención educativa de enfermería en cuidadores primarios. Enfermería Universitaria; 17(4): 390-402.

Yáñez Corral, A. M. & Zárate Grajales, R. A. (2016). Diseño de un instrumento para evaluar el proceso de enlace de turno de enfermería. Enfermería Universitaria; 13(2): 99-106. http://dx.doi.org/10.1016/j.reu.2016.03.002

Zaiontz, C. (2020). Real Statistics using Excel. http://www.real-statistics.com/

Zapata Zabala, M., Cárdenas Niño, L. & Cuartas Arias, J. M. (2019). INECO Frontal Screening (IFS): una herramienta psicométrica para evaluar la disfunción ejecutiva en policonsumidores. Suma Psicológica; 26(2):119-126. DOI: 10.14349/sumapsi.2019.v26.n2.8