MHSalud
ISSN: 1659-097X
revistamhsalud@una.cr
Universidad Nacional
Costa Rica

Coeficientes V de Aiken: diferencias en los juicios de validez de contenido

Merino-Soto, César

Coeficientes V de Aiken: diferencias en los juicios de validez de contenido
MHSalud, vol. 20, núm. 1, 2023
Universidad Nacional, Costa Rica
Disponible en: https://www.redalyc.org/articulo.oa?id=237072359003
DOI: https://doi.org/10.15359/mhs.20-1.3

Artículos

Coeficientes V de Aiken: diferencias en los juicios de validez de contenido

Aiken's V Coefficient: Differences in Content Validity Judgments

Coeficientes Aiken V: Diferenças nos julgamentos de validade do conteúdo

César Merino-Soto
Universidad de San Martín de Porres, Perú
sikayax@yahoo.com.ar
https://orcid.org/0000-0002-1407-8306

DOI: https://doi.org/10.15359/mhs.20-1.3
Redalyc: https://www.redalyc.org/articulo.oa?id=237072359003

Recepción: 15 Abril 2021
Aprobación: 30 Marzo 2022

Resumen:

Objetivo: Cuando es estudiada la validez de contenido en dos grupos independientes de jueces expertos, se requiere hacer una prueba formal de las diferencias entre sus juicios, dado que es posible obtener distintos juicios de validez de contenido. Pero, generalmente, la investigación de la validez de contenido no examina esta posible fuente de discrepancias. El presente reporte describe la implementación de un método para evaluar la diferencia de coeficientes V de Aiken aplicado al trabajo investigativo en ciencias del deporte.
Metodología: El procedimiento aplica una adaptación para construir el intervalo de confianza de la diferencia entre coeficientes V de Aiken y también implementa un estimador estandarizado del tamaño de la distinción entre los coeficientes V, específicamente, la transformación arcoseno de coeficientes V.
Resultados: Se desarrollan dos ejemplos, en un marco de análisis secundario de datos, y se demuestra la diferencia entre la conclusión con base impresionista y la conclusión con base empírica y evaluación formal. Se detectaron distinciones estadísticas no observadas previamente.
Conclusiones e implicaciones: El método que estima diferencias de coeficientes de validez de contenido V de Aiken para la investigación permite un avance en la metodología que valida instrumentos de medición. Se valora la aplicabilidad de este procedimiento en el contexto de ciencias del deporte y ciencias de la educación, así como en el diseño de la investigación.

Palabras clave: psicometría, test psicológico, análisis estadístico, estudio de validación, metodología.

Abstract:

Objective: When two independent groups of expert judges study content validity, a formal test of the differences between their judgments is required, since different content validity judgments can be obtained. But generally, content validity research does not examine this likely source of discrepancies. This report describes the implementation of a method to evaluate the difference in Aiken's V coefficients applied to research work in sports science.
Methodology: The procedure applies an adaptation to construct the confidence interval of the difference between Aiken's V coefficients and also implements a standardized estimator of the size of the difference between the V coefficients, specifically the arcsine transformation of V coefficients.
Results: In a secondary data analysis framework, two examples are developed, extracting data from both publications, and the difference between the impressionist-based conclusion and the empirical-based conclusion and formal evaluation is demonstrated. Statistical differences not previously observed were detected.
Conclusions and implications: The method to estimate differences in Aiken's content validity coefficients for research allows an advance in the methodology to validate measurement instruments. The applicability of this procedure in the context of sports sciences and education sciences, as well as in the research design involved, is assessed.

Keywords: psychometric, psychological test, statistical analysis, validation studies, methodology.

Resumo:

Objetivo: Quando a validade do conteúdo é estudada em dois grupos independentes de juízes especialistas, é necessário um teste formal das diferenças entre seus julgamentos, uma vez que é possível obter diferentes julgamentos de validade de conteúdo. Mas a pesquisa de validade do conteúdo geralmente não examina esta possível fonte de discrepâncias. Este relatório descreve a implementação de um método para avaliar a diferença dos coeficientes V de Aiken aplicada ao trabalho de investigação das ciências do desporto.

Metodologia: O procedimento aplica uma adaptação para construir o intervalo de confiança da diferença entre os coeficientes V de Aiken, e também implementa um estimador estandardizado do tamanho da diferença entre os coeficientes V, especificamente a transformação arco-seno dos coeficientes V.
Resultados: Dois exemplos são desenvolvidos em uma estrutura secundária de análise de dados, e a diferença entre a conclusão baseada no impressionismo e a conclusão baseada no empirismo com avaliação formal é demonstrada. Diferenças estatísticas não observadas anteriormente foram detectadas.
Conclusões e implicações: O método para estimar as diferenças nos coeficientes de validade do conteúdo V de Aiken para a pesquisa permite um avanço na metodologia de validação dos instrumentos de medição. A aplicabilidade deste procedimento no contexto da ciência do esporte e da ciência educacional, assim como no projeto de pesquisa, é avaliada.
Palavras-chave: Educação Física, direito humano, ensinar e aprender.

Introducción

Para obtener evidencias de validez una medida utilizable en ciencias del deporte y ciencias aliadas, posiblemente la evaluación de la validez de contenido condiciona la obtención de otras de evidencias de validez de un instrumento (American Educational Research Association et al., 2014; Koller et al., 2017). Esto es debido a que las características de contenido del constructo son creadas a priori, vinculando la experiencia profesional, la racionalidad del investigador en el constructo de interés y la literatura relevante de este. La validez de contenido es una fase en la que se elaboran y seleccionan los contenidos de los ítems; tiene como medio el juicio y la racionalidad del investigador. La metodología cuantitativa también está implicada en este proceso y, en consecuencia, una de las decisiones es elegir el estadístico o coeficiente que cuantifique el grado de validez.

En la literatura empírica o de revisión, se han identificado varios de estos coeficientes (Aiken, 1980; Fitch et al., 2001; Hambleton, 1984; Hernández-Nieto, 2002; Lawshe, 1975; Lynn, 1986; Penfield y Miller, 2004; Rovinelly y Hambleton, 1977) o procedimientos combinados con estimadores del acuerdo (Claeys et al., 2012; Moscoso y Merino-Soto, 2017; Rubio et al, 2003), de tal modo que el usuario tiene la oportunidad de elegir y comparar sus resultados. Tales coeficientes conducen a obtener un solo número sumario (i. e. estimación puntual), que sirve para interpretar la relación entre el contenido del ítem y el dominio de contenido al cual puede pertenecer. Una vez calculado y acompañado con otras referencias numéricas (e. g. intervalos de confianza), la estadística parece que culmina aquí su papel.

Sin embargo, el análisis puede expandirse hacia la comparación de estos resultados, como ocurriría habitualmente en diseños de investigación en los que se comparan estadísticos muestrales (e. g. medias o varianzas). Las revisiones temáticas hispanas sobre la validez de contenido (Cabero y Llórente, 2013; Escobar y Cuervo, 2008; Pedrosa et al., 2013; Robles y Rojas, 2015; Urrutia et al., 2014) describen con variado detalle los métodos para cuantificar los juicios de los jueces expertos, pero no alcanzan a orientar la investigación hacia nuevos diseños, como la comparación de grupos usando como insumos los coeficientes de validez de contenido calculados. Esta comparación de grupos ha sido demostrada como relevante para exponer las potenciales diferencias, aun entre jueces expertos y jueces experienciales (Merino-Soto, 2016) o jueces expertos de distinto origen (Moscoso y Merino-Soto, 2017).

Por otro lado, en la indagación psicométrica que incluye la evaluación de la validez de contenido en ciencias de la actividad física y del deporte (Burgueño et al., 2020; Calonge-Pascual et al., 2020; Collet et al., 2018; Gamonales et al., 2018; Moreno y Gómez, 2017; Ortega et al., 2018; Robles et al., 2016; Rodríguez et al., 2015), tampoco es usual identificar grupos que pueden añadir variabilidad a las estimaciones de validez de contenido (por ejemplo, con el coeficiente V, Aiken, 1980). De este modo, se desconoce si la percepción de validez de contenido producida por el distinto estatus, la experticia o los conocimientos de los jueces produce o variabilidad en las estimaciones del coeficiente V. En contraste, en un estudio se reportaron apropiadamente coeficientes V para dos grupos de jueces (Calonge-Pascual et al., 2020), pero no fueron cuantificadas, mediante un método formal, las eventuales diferencias entre ambos.

Para afrontar el vacío, el presente manuscrito metodológico tiene por objetivo mostrar dos desarrollos metodológicos, uno de ellos publicado recientemente (Merino, 2016)con el afán de comparar coeficientes V (Aiken, 1980 obtenidos de dos grupos independientes. Esta presentación intenta acercar la técnica a la población de usuarios e investigadores en ciencias del deporte, pero incorporando otra propuesta relevante a la comparación de coeficientes V. En este sentido, la segunda propuesta es derivar la racionalidad para usar un estimador puntual estandarizado de la diferencia entre coeficientes V. En Hispanoamérica, el coeficiente V parece ser habitualmente utilizado en ciencias de la conducta y educación para cuantificar la validez de contenido obtenida de expertos, y dada esta relevancia contextual la técnica sugerida por Merino (2018 se ejemplifica y se extiende hacia otra implementación, ambas de potencial utilidad dirigida al trabajo investigativo en ciencias del deporte.

Desarrollo

Marco general. La comparación de estimaciones de validez de contenido entre dos grupos es apropiada cuando el investigador, a priori, estima que la pertenencia a cualquiera de los grupos puede ser fuente de variabilidad en la percepción del contenido evaluado. Esta diferencia asociada a los dos grupos de jueces puede actuar como un moderador de sus juicios de validez de contenido (Merino-Soto, 2016; por ejemplo, entre dos agrupaciones de jueces expuestos a experiencias claramente distintas como estatus profesional (e. g. entrenador vs. atleta), oportunidades educativas, recursos sociales y económicos, entre otros.

Propuesta metodológica. El método propuesto para comparar coeficientes V entre dos grupos independientes (Merino-Soto, 2018) utiliza la construcción de intervalos de confianza (IC) de diferencia de proporciones. Este método es la generalización de un procedimiento para construir IC para las diferencias entre parámetros que pueden conceptualizarse como indicadores de la magnitud del efecto (Newcombe, 2012; Zou y Donner, 2008). Los detalles de la racionalidad de dicha técnica se encuentran en Merino-Soto (2016) y el lector puede consultarlos libremente.

La segunda propuesta inédita relativa a la primera expresa que, junto con el método de IC para las diferencias, es posible calcular un estimador puntual de esta diferencia, pero expresada en unidades estandarizadas. Para esta finalidad, y debido a que el coeficiente V puede ser tratado como una proporción (Aiken, 1980; Penfield y Giacobbi, 2004), es admisible utilizar un coeficiente que estandarice la distancia entre proporciones y evitar la diferencia cruda de V (también manifestada como un contraste de proporciones), la cual no tiene un escalamiento intervalar constante (Cohen, 2008). La racionalidad de esta elección proviene del tratamiento usual que se hace al coeficiente V, esto es, como una proporción (Aiken, 1980; Penfield y Giacobbi, 2004) con límites naturales de 0 y 1. Tal proporción se origina en la ecuación de V y, por lo tanto, las transformaciones para las proporciones también son viables para el coeficiente V.

Consecuentemente, para la estimación puntual de la diferencia entre coeficientes V, se propone la diferencia arcoseno (h), método asociado al trabajo de Cohen (2008), pero que tiene más tiempo de existencia (Ascombe, 1948; Freeman y Tukey, 1950; McCullagh y Nelder, 1989). Esta contraste entre coeficientes V requiere primero su transformación arcoseno (Cohen, 2008) y, en segundo lugar, calcular la diferencia entre ellas. Esto tiene varias ventajas: a) tiende a ser menos sesgada en distribuciones muy asimétricas (Lipsey y Wilson, 2001) y, por ende, puede ser la más apropiada para el tipo de distribuciones que ocurren en las calificaciones de los jueces, esto es, de tipo asimétricasasimétricas en las cuales las calificaciones son más densas en las colas de la distribución (Penfield y Giacobbi, 2004; Penfield, y Miller, 2004; b) esta transformación habitualmente arroja estimaciones conservadoras (Lipsey y Wilson, 2001), lo que conduce a disminuir el error tipo I; c) por otro lado, maneja proporciones con 0 y 1 sin producir resultados inestimables, como ocurre cuando se utiliza las transformaciones logit y probit (Rücker et al., 2008; Rücker et al., 2009). Por estos motivos, la transformación arcoseno puede proporcionar resultados creíbles e interpretables.

La diferencia entre dos coeficientes V con previa transformación arcoseno permite operacionalizar su tamaño o magnitud, en una métrica estandarizada. Sin embargo, existen varios modelos estadísticos para definir las unidades intervalares en este coeficiente h, que varían de acuerdo con la distribución poblacional del estadístico de interés (en esta situación, una proporción), como la distribución logística o la normal. Se pueden sugerir otros tipos de transformaciones (e. g. probit, Glass et al., 1981; logitCox, Cox, 1970), pero existe alguna inconsistencia entre ellas, debido a que estudios de simulación han encontrado que h basado en la trasformación arcoseno subestima el valor poblacional, mientras que otros muestran poco sesgo con estas transformaciones cuando la distribución es normal (Sánchez-Meca et al., 2003; Warton y Hui, 2011). No obstante, como se mencionó, una distribución normal de V y sus diferencias es poco probable que ocurra cuando se obtienen las calificaciones de estudios de validez de contenido y, por ello, no sería óptima esta generalización. Si hay alguna duda sobre la elección apropiada de una de estas transformaciones, posiblemente es mejor una aproximación razonable que la exactitud cuestionable (Agresti y Brent, 1998).

Resultados

En un marco de análisis secundario de datos y para ejemplificar tanto la aplicación como la relevancia de la presente metodología, se compararon los coeficientes V del estudio de validez de contenido de dos indagaciones publicadas. En el primero, se utilizó una medida de autoeficacia para la investigación (Domínguez-Lara, 2017), con una muestra de dos grupos de jueces que evaluaron la claridad y la relevancia del contenido del instrumento: 10 investigadores y 34 estudiantes (15 del posgrado y 19 de pregrado); con el propósito de demostrar el método, aquí elegimos el grupo de estudiantes. Este artículo fue seleccionado puesto que 1) es uno de los pocos en los que aparece la información mínima y necesaria para efectuar análisis secundarios; 2) el periodo de construcción y revisión actual del instrumento es relativamente reciente, así como se pueden aprovechar los resultados del reanálisis con esta nueva propuesta metodológica, y 3) es relevante para la aplicación del procedimiento.

De acuerdo con los resultados del autor, se concluye la buena claridad de los ítems, dado que los coeficientes V puntuales y sus intervalos de confianza fueron altos o moderadamente altos para cada uno, así como para los coeficientes promedio de cada dimensión y el contenido total. Con un criterio liberal de aceptación de la relevancia y claridad del ítem (> = .50; Domínguez-Lara, 2017), es posible aceptar una gran cantidad de ítems en las fases iniciales de construcción, con el riesgo de aumentar el error tipo II, esto es, aceptar ítems que deberían ser rechazados. Por otra parte, la claridad de los ítems entre los grupos (posgrado y pregrado) fue evaluada comparativamente mediante interpretaciones impresionistas sobre las similitudes o diferencias entre sus coeficientes V, expuestos en la sección izquierda de su tabla correspondiente. El análisis formal de estas comparaciones, por medio del intervalo de confianza de las diferencias de V y el tamaño del efecto (transformación arcoseno, h) presentados en la tabla 1 de este manuscrito, arrojó que hay nueve ítems en los que los juicios de claridad difieren. Algunas de estas diferencias pueden ser consideradas grandes o moderadas (respectivamente: h ≥ |.80| y h ≥ |.50|; Cohen, 2008). Esto implica que la observación impresionista y la prueba formal no necesariamente coinciden, aun en el ojo experto.

TABLA 1
Reanálisis de Domínguez-Lara (2017): intervalo de confianza para diferencia entre coeficientes V de validez de contenido

No. de ítem	IC de la diferencia		*h*	Conclusión
	Inferior	Superior		
1	-.242	.047	-.261	No diferente
2	-.093	.199	.163	No diferente
3	-.192	.144	-.048	No diferente
4	-.230	.089	-.165	No diferente
5	-.203	.084	-.152	No diferente
7	-.165	.121	-.048	No diferente
8	-.203	.084	-.152	No diferente
9	-.331	-.060	-.607	Diferente
10	-.331	-.060	-.607	Diferente
11	-.224	.086	-.166	No diferente
12	-.268	.028	-.317	No diferente
13	-.239	.066	-.217	No diferente
14	-.306	-.011	-.429	Diferente
15	-.217	.066	-.202	No diferente
16	-.285	.029	-.319	No diferente
17a	-.330	-.032	-.484	Diferente
17b	-.242	.047	-.261	No diferente
18	-.241	.017	-.350	No diferente
19	-.353	-.076	-.655	Diferente
20	-.199	.127	-.076	No diferente
21	-.172	.125	-.051	No diferente
22	-.306	-.011	-.429	Diferente
23	-.374	-.030	-.460	Diferente
24	-.426	-.135	-.811	Diferente
25	-.285	.005	-.381	No diferente
26	-.386	-.093	-.673	Diferente
27	-.239	.087	-.176	No diferente

Nota. h: diferencia estandarizada arcoseno. IC: intervalo de confianza

El otro ejemplo de aplicación se enfocó en el reporte bien documentado de validez de contenido de un instrumento de prescripción de la actividad física (Calonge-Pascual et al., 2020). En su tabla 6, se

presentan en detalle los resultados de dos grupos de jueces expertos, médicos del deporte y enfermeros. La aplicación del método de intervalos de confianza para la diferencia a los resultados de su tabla sugiere que, predominantemente, los ítems son percibidos de manera similar por ambos grupos de expertos, excepto en el 18, 22 y 23; particularmente, la diferencia estandarizada expresada por h puede considerarse grande (Cohen, 2008). Estas diferencias no parecen observables en el simple examen visual de la tabla 6 de Calonge-Pascual et al. (2020). Aunque todos los coeficientes V para el instrumento son altos, esta diferencia indica una discrepancia grande y focalizada, que merece más atención. Si el criterio de V para aceptar los ítems fuera más grande (V > .80), posiblemente la diferencia se percibiría con mayor claridad y llamaría a una revisión del ítem (tabla 2).

TABLA 2
Reanálisis de Calonge-Pascual et al. (2020): intervalo de confianza para diferencia entre coeficientes V de validez de contenido

No. de ítem	IC de la diferencia		*h*	Conclusión
	Inferior	Superior		
1	-.010	.193	.192	No diferente
2	-.020	.246	.205	No diferente
3	-.120	.125	-.132	No diferente
4	-.040	.182	.094	No diferente
5	-.040	.169	.070	No diferente
6	-.040	.207	.139	No diferente
7	-.050	.186	.058	No diferente
8	-.040	.169	.070	No diferente
9	-.190	.042	-.357	No diferente
10	-.110	.146	-.051	No diferente
11	-.200	.006	-.427	No diferente
12	-.090	.175	.000	No diferente
13	-.070	.114	-.070	No diferente
14	-.130	.039	-.344	No diferente
15	-.090	.137	-.058	No diferente
16	-.100	.089	-.164	No diferente
17	-.110	.059	-.287	No diferente
18	-.170	-.015	-.528	Diferente
19	-.060	.152	.000	No diferente
20	-.120	.022	-.390	No diferente
21	-.170	.030	-.354	No diferente
22	-.200	-.073	-.927	Diferente
23	-.190	-.035	-.579	Diferente
24	-.050	.225	.124	No diferente
25	-.160	.020	-.425	No diferente
26	-.070	.205	.073	No diferente
27	.000	.266	.263	No diferente
28	-.090	.166	.000	No diferente
29	-.170	.030	-.354	No diferente
30	-.120	.022	-.390	No diferente

Nota. h: diferencia estandarizada arcoseno. IC: intervalo de confianza

Discusión y conclusiones

El método propuesto para evaluar las diferencias entre coeficientes V introduce un enfoque inédito en el diseño de estudios de validez de contenido y, por lo tanto, facilita la implementación de un análisis en el que se puedan comparar dos grupos independientes de jueces. Como un medio objetivo y cuantificable para presentar resultados y tomar decisiones con base empírica y en un marco exploratorio o confirmatorio (Merino, 2018), tal método hace viable la selección, a priori, de grupos de evaluadores, con el fin de que puedan ser comparados posteriormente. Por ejemplo, en el diseño de un estudio de validez de contenido de un nuevo instrumento de acoso, la percepción de representatividad de ítems sobre hostigamiento en la práctica deportiva puede orientarse hacia evaluadores varones y mujeres o entre entrenadores y entrenados.

Respecto a los ejemplos reanalizados (Calonge-Pascual et al., 2020; Domínguez-Lara, 2017), la aplicación del método propuesto reveló que los instrumentos pueden requerir una exploración adicional de su contenido para verificar su claridad o relevancia y, consecuentemente, planificar ajustes en la fase de elaboración. Esto es más necesario cuando se emplea un criterio fuertemente liberal para la selección inicial de ítems, en la que no hay suficiente garantía para la calidad de la representación del constructo o cuando existe una clara distinción entre grupos de jueces que pueden añadir variabilidad a las estimaciones de validez de contenido. Finalmente, los lectores interesados pueden solicitar gratuitamente al autor del presente manuscrito el software elaborado en sintaxis SPSS, para implementar el procedimiento.

Referencias

Agresti, A. y Brent, C. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician, 52*(2), 119-126. https://doi.org/10.1080/00031305.1998.10480550

Aiken, L. (1980). Content validity and reliability of single items or questionnaire. *Educational and Psychological Measurement,40*(4), 955-959. https://doi.org/10.1177/001316448004000419

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika, 35*(3/4), 246-254. https://doi.org/10.1093/biomet/35.3-4.246

Burgueño, R., Macarro-Moreno, J. y Medina-Casaubón, J. (2020). Psychometry of the Multidimensional Perceived Autonomy Support Scale in Physical Education with Spanish secondary school students. *SAGE Open.* https://doi.org/10.1177/2158244019901253

Cabero, J. y Llorente, M. (2013). La aplicación del juicio de experto como técnica de evaluación de las tecnologías de la información (TIC). *Eduweb. Revista de Tecnología de Información y Comunicación en Educación, 7*(2), 11-22. http://servicio.bc.uc.edu.ve/educacion/eduweb/v7n2/art01.pdf

Calonge-Pascual, S., Fuentes-Jiménez, F., Casajús Mallén, J. A., y González-Gross, M. (2020). Design and validity of a choice-modeling questionnaire to analyze the feasibility of implementing physical activity on prescription at primary health-care settings. *International Journal of Environment Research and Public Health, 17*, 6627. https://doi.org/10.3390/ijerph17186627

Claeys, C., Nève, J., Tulkens, P. M. y Spinewine, A. (2012). Content validity and inter-rater reliability of an instrument to characterize unintentional medication discrepancies. *Drugs Aging, 29*, 577-591. https://doi.org/10.1007/bf03262275

Cohen, J. (2008). *Statistical power analysis for the behavioral sciences*. Second edition. Lawrence Erlbaum Associates, Inc.

Collet, C., Nascimento, J. V., Folle, A. e Ibáñez, S. J. (2018). Construcción y validación de un instrumento para el análisis de la formación deportiva en voleibol. *Cuadernos de Psicología del Deporte, 19*(1), 178-191. https://doi.org/10.6018/cpd.326361

Cox, D. R. (1970). *Analysis of binary data*. Chapman y Hall/CRC.

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. http://dx.doi.org/10.19083/ridu.11.514

Escobar, J. y Cuervo, Á. (2008). Validez de contenido y juicio de expertos: una aproximación a su utilización. *Avances en Medición, 6*(1), 27-36.

Fitch, K., Bernstein, S. J., Aguilar, M. D., Burnand, B., LaCalle, J. R., Lazaro, P., ... Kahan, J. P. (2001). *The RAND/UCLA Appropriateness Method User's Manual*. RAND corporation.

Freeman, M. F. y Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics, 21*, 607-611. https://doi.org/10.1214/aoms/1177729756

Gamonales, J., León, K., Muñoz, J., González-Espinosa, S. e Ibáñez, S. (2018). Validación del IOLF5C para la eficacia del lanzamiento en fútbol para ciegos. *Revista Internacional de Medicina y Ciencias de la Actividad Física y del Deporte, 18*(70). https://doi.org/10.15366/rimcafd2018.70.010

Glass, G. V., McGaw, B. y Smith, M. L. (1981). *Meta-analysis in social research.* Sage.

Hambleton, R. K. (1984). Validating the test score. En R. A. Berk (ed.), *A Guide to Criterion-Referenced Test Construction* (pp. 199-230). Johns Hopkins University Press.

Hernández-Nieto, R. A. (2002). *Contributions to Statistical Analysis*. Universidad de Los Andes.

Koller, I., Levenson, M. R. y Glück, J. (2017). What do you think you are measuring? A mixed-methods procedure for assessing the content validity of test items and theory-based scaling. *Frontiers in Psychology, 8*, 126. https://doi.org/10.3389/fpsyg.2017.00126

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Lipsey, M. y Wilson, D. (2001). *Practical meta-analysis*. Sage.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*, 382-385. https://doi.org/10.1097/00006199-198611000-00017

McCullagh, P. y Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall.

Merino-Soto, C. (2016). Percepción de la claridad de los ítems: Comparación del juicio de estudiantes y jueces-expertos. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 14*(2), 1469-1477. https://doi.org/10.11600/1692715x.14239120615

Merino-Soto, C. (2018). Confidence interval for difference between coefficients of content validity (Aiken's V): a SPSS syntax. *Anales de Psicología, 34*(3), 587-590. https://dx.doi.org/10.6018/analesps.34.3.283481

Moreno, E. y Gómez, M. (2017). Validación herramienta observacional para el análisis de rachas de lanzamiento en baloncesto. *Revista de Psicología del Deporte, 26*(1), 87-93.

Moscoso, M. S. y Merino-Soto, C. (2017). Construcción y validez de contenido del Inventario de Mindfulness y Ecuanimidad: una perspectiva iberoamericana. *Mindfulness & Compassion,2*(1), 9-16. https://doi.org/10.1016/j.mincom.2017.01.001

Newcombe, R. G. (2012). *Confidence Intervals for Proportions and Related Measures of Effect Size*. CRC Biostatistics Series.

Ortega, G., Abad, M., Giménez, F., Durán, L., Franco, J., Jiménez, A. y Robles, J. (2018). Design and validation of a satisfaction questionnaire with sports programmes in penitentiaries. *Apunts. Educación Física y Deportes, 131*(1), 21-33. http://dx.doi.org/10.5672/apunts.2014-0983.es.(2018/1).131.02

Pedrosa, I., Suárez-Álvarez, J. y García-Cueto, E. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica*, 10(2), 3-18. http://dx.doi.org/10.5944/ap.10.2.11820

Penfield, R. D. y Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education, 17*(4), 359-370. http://dx.doi.org/10.1207/s15324818ame1704_2

Penfield, R. y Giacobbi, P. (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

Robles, A., Robles, J., Giménez, F. y Abad, M. (2016). Validación de una entrevista para estudiar el proceso formativo de judokas de élite. *Revista Internacional de Medicina y Ciencias de la Actividad Física y del Deporte, 64*. https://doi.org/10.15366/rimcafd2016.64.007

Robles, P. y Rojas, M. (2015). La validación por juicio de expertos: dos investigaciones cualitativas en lingüística aplicada. *Revista Nebrija de Lingüística Aplicada*, (18). https://www.nebrija.com/revista-linguistica/files/articulosPDF/articulo_55002aca89c37.pdf

Rodríguez, P. L., Pérez, J. J., García, E. y Rosa, A. (2015). Adaptación transcultural de un cuestionario que evalúa la actividad física en niños de 10 y 11 años. *Archivos Argentinos de Pedriatría,113*(3), 198-204.

Rovinelli, R. J. y Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research, 2*, 49-60.

Rubio, D. M., Berg-Weber, M., Tebb, S. S., Lee, E. S. y Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research, 27*(2), 94-104. https://doi.org/10.1093/swr/27.2.94

Rücker, G., Schwarzer, G., Carpenter, J. y Olkin, I. (2009). Why add anything to nothing? The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine, 28*(5), 721-738. https://doi.org/10.1002/sim.3511

Rücker, G., Schwarzer, G. y Carpenter, J. (2008). Arcsine test for publication bias in meta-analyses with binary outcomes. *Statistics in Medicine*, 27(5), 746-763. https://doi.org/10.1002/sim.2971

Sánchez-Meca, J., Marín-Martínez, F. y Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods, 8*(4), 448-467. http://dx.doi.org/10.1111/j.1469-185X.2007.00027.x

Urrutia, M., Barrios, S., Gutiérrez, M. y Mayorga, M. (2014). Métodos óptimos para determinar validez de contenido. *Educación Médica Superior, 28*(3), 547-558.

Warton, D. I. y Hui, F. K. C. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology, 92*, 3-10. https://doi.org/10.1890/10-0340.1

Zou, G. Y. y Donner, A. (2008). Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine, 27*, 1693-1702. http://dx.doi.org/10.1002/sim.3095