anales de psicología / annals of psychology
2022, vol. 38, nº 2 (may), 395-398
https://doi.org/10.6018/analesps.489431

Rating mean of expert judges and asymmetric confidence intervals in content validity: An SPSS syntax

César Merino-Soto[1*], and José Livia-Segovia[2]

1 Psychology Research Institute, University of San Martín de Porres (Perú)
2 Faculty of Psychology, Federico Villarreal National University (Perú)

Título: Calificación promedio de jueces expertos e intervalos de confianza asimétricos en la validez de contenido: Una sintaxis SPSS.
Resumen: La estimación de la validez de contenido, obtenida mediante el análisis racional de jueces expertos, habitualmente se hace con coeficientes que estandarizan entre 0.0 y 1.0 el juicio de los jueces. Sin embargo, esta estimación también puede expresarse en la métrica de las respuestas de los jueces, en la forma de la media de respuesta, y con intervalos de confianza asimétricos alrededor de esta media. El objetivo del presente manuscrito es implementar un procedimiento para estas estimaciones (media de respuesta e intervalos de confianza asimétricos) en un programa escrito en sintaxis SPSS. Se explica la racionalidad del procedimiento, y se desarrolla un ejemplo aplicado del cálculo. El programa es de distribución libre, solicitándolo a los autores.
Palabras clave: Validez de contenido. Software. Jueces expertos de contenido. Estadística. Validez.

Abstract: The estimation of content validity, obtained by rational analysis of expert judges, is usually done with coefficients that standardize between 0.0 and 1.0 the judges' judgment. However, this estimate can also be expressed in the metric of the judges' responses, in the form of the response mean, and with asymmetric confidence intervals around this mean. The aim of the present manuscript is to implement a procedure for these estimates (response mean and asymmetric confidence intervals) in a program written in SPSS syntax. The rationale of the procedure is explained, and an applied example of the calculation is developed. The program is freely distributed upon request to the authors.
Keywords: Content validity. Software. Subject matter experts. Statistics. Validity.

Context

In the research of content validity by means of expert judges or study participants, one of the usual steps is to quantify the results with various methods (e.g., for a summary of the methods, see Pedrosa, Suarez-Alvarez & Garcia-Cueto, 2013). Quantifying the content validity is usually calculated in one coefficient, ranging from 0.0 to 1.0 (e.g., Aiken, 1980; Hernandez-Nieto, 2002; Lynn, 1986; Osterlind, 1992), or from -1.0 to 1.0 (e.g., Lawshe, 1975; Rovinelli & Hambleton, 1977), and generally values close to 1.0 are interpreted as evidence of the strength of the validity of the measured attribute. This standardization applies regardless of the measure of the responses; for example, a scaling from 1 to 3, or from 0 to 7. This type of transformation is very common, because its interpretation framework eliminates infinite values at the extremes, it is comparable to a transformation based on percentages (from 0.0% to 100%), and you can ensure that the final consumers of the results will quickly understand the information (Bonett & Price, 2020). Also, this transformation makes the interpretation independent of the scaling frame in which the data were collected. For example, a coefficient V (Aiken, 1980) of .70 can be obtained from an average of 3.8 on a scaling 1 to 5, an average of 2.8 on a scaling 0 to 4, or an average of 7.3 within a scale of 1 to 10. However, it is not the only possible transformation, nor the only framework to understand its results.

In a study of content validity, where the data collected are formulated in a particular measure, for example, from 1 to 5, the results also may be in the same metric, and does not require any other processing to communicate the results. The expression of results in the metric of the rating scale is useful in preventing another metric for its interpretation, and it has the advantage of contextualizing in the same units that the individual scores were produced. For example, the evidence of content representativeness can be scaled to 1 (completely unrepresentative), 2 (unrepresentative), 3 (moderately representative), 4 (acceptably representative), and 5 (fully representative); if a group of judges produces an average scoring of 4.1, and this average suggests an acceptable representation of the construct, but it also can be said that the trend of the perceived validity is just at this level 4. The accuracy of this kind of information requires other interpretable quantities, such as confidence intervals (CI).

Due to a point estimate does not guarantee accuracy itself, a confidence interval indicates its accuracy through a range of variability of the estimated value. For the user, this can be interpreted as the degree of certainty of finding the estimated value in the reference population, which will allow a better decision-making in the research works, and added information in the interpretation of the results (Escrig-Sos et al., 2007; Tellez et al., 2015). The CIs for the content validity coefficients seem to have barely been developed (Penfield & Giacobbi, 2004). Also, a casual review of the literature repeatedly performed by the authors of this manuscript, found that, beyond Penfield's work (Miller & Penfield, 2005; Penfield & Giacobbi, 2004), no CI procedures for these types of coefficients have been widespread or derived to date.

*** Correspondence address [Dirección para correspondencia]:**
César Merino-Soto. Universidad de San Martín de Porres, Lima (Perú).
E-mail: sikayax@yahoo.com.ar
(Article received: 14-08-2021; reviewed: 06-11-2021; accepted: 15-11-2021)

Method

To add information about the statistics accuracy of the coefficients of content validity, confidence intervals must be built around the estimated coefficient. However, the traditional method for creating confidence intervals (i.e., the Wald method) requires the assumption of normal distribution of calculated average (Penfield, 2003). Because the data obtained in content validity studies are usually discrete, scaled down to five or fewer options, show skewed distribution, and the usual size of the judges' sample is generally small, another method is required to estimate appropriate intervals. (Miller & Penfield, 2005; Penfield & Miller, 2004).

Based on the work of Penfield (2003), who adapted the Wilson (1927) *score* procedure to generate asymmetric confidence intervals, Penfield and Miller (2004) used this adaptation for data produced in content validity studies; specifically, scores based on ordinal metrics, small judges' samples (less than 10), reduced number of response options, and skewed distribution. There are other methods to generate asymmetric confidence intervals (e.g., Willink, 2005), but Wilson's method appears to be efficient for the conditions in which content validity studies are conducted (Penfield, 2003; Penfield & Miller, 2004). Because the manual calculation is prone to error at computing, a computer code is proposed for this purpose, using the Statistics Package for the Social Sciences (abbreviated as SPSS). In recent years, R has been a free powerful platform for software development, but the SPSS is still used in different fields of the methodological research (e.g., Duricki, Soleman, & Moon, 2016; Vanus, Kubicek, Gorjani, & Koziorek, 2019; Valeri, & Vanderweele, 2013), and that over time its ease of handling and immediate understanding especially in the field of science social was noted (Gogoi, 2020; Rivadeneira et al., 2020).

The description of procedure and its rationality can be seen directly in Penfield and Miller (2004), so readers can inspect the formulas. The literature related to this method, based on the score procedure (Wilson, 1924), can be found in Penfield (2003), Penfield and Giacobbi (2004), and Miller and Penfield (2005).

Program

To calculate the average of the content validation and the asymmetric confidence intervals, an SPSS syntax is written which is adapted to the SAS syntax by Miller and Penfield (2005). Even with the rise and popularity of the free software (Haine, 2019; Muenchen, 2017), the SPSS is of persistent use and it still holds the attention of researchers in different disciplines (Haine, 2019; Masuadi et al., 2021; Muenchen, 2017; Shaikh, 2016, 2017). Similar to a few ad hoc programs to obtain evidence of the content validity (e.g., Merino-Soto, 2018; Merino-Soto & Livia-Segovia, 2009), the program presented here implements the method of Penfield and Miller (2004) for constructing confidence intervals around the mean value of the scores, and reports the mean of the expert judges' scores, its equivalent expression in proportion to the scaling range, the asymmetric CIs around this mean, and the width of the interval. The program is freely accessible and it must be requested from the corresponding author.

Application example

To illustrate the procedure, the results for the evidence of clarity of the items of the dimension *Atención Sostenida*, made by Moscoso and Merino-Soto (2017; see Table 2) were used, in which the Inventario de Ecuanimidad y Mindfulness was validated. In the input specifications for the program, you have to report the mean (*M*), the sum (*S*) of the judges' ratings for each item, the number of judges (17), the number of scaling responses (*k*; in this example, 6 ordinal points, from no validity to full validity), the error rate (*alpha*; among the options: .10, .05, .01), and the value of the number (*start*) with which *k* starts, that is, 0 or 1 (see Table 1, Input heading). In the Output heading of Table 1, the results after applying the program for the calculations are shown. It is observed that, in all the items, except item 1, the lower limit of the interval exceeds the value 4. If the researcher establishes a priori that the items with lower limit of CI must exceed is 4, then this item does not seem to fit this criterion. However, given the difference of .02, the researcher must make a decision whether to apply the criteria strictly or to be flexible.

Table 1
Input and output information for the program

	Input						Output				
	M	S	*N*	K	alpha	start	*M*	P	Low	Upp	Range
Item 1	4.53	77	17	6	.05	1	4.53	.75	3.98	5.08	1.10
Item 2	5.18	88	17	6	.05	1	5.18	.86	4.7	5.66	.96
Item 3	5.47	93	17	6	.05	1	5.47	.91	5.04	5.89	.85
Item 4	5.06	86	17	6	.05	1	5.06	.84	4.56	5.55	.99
Item 5	4.65	79	17	6	.05	1	4.65	.77	4.10	5.19	1.08

Note. M: Mean. S: Sum of the qualifications. N: Number of judges. K: Number of the response options (i.e., scaling) used by the judges. Alpha: Error rate (it can be .01, .05, or .10). Start: Start of the scaling used by the judges (0 or 1; see manuscript). P: Ratio equivalent to the mean, of M with respect to the scale used by the judges. Low: Lower limit of the confidence interval. Upp: Upper limit of the confidence interval. Range: Width of the estimated confidence interval.

Final comments

This manuscript presents a computer program, built in the SPSS syntax, to quantify the degree to which the judges agree on the content validity, and their asymmetric confidence intervals. The fundamental difference of this method with others (eg., Merino & Livia, 2009) is that the results are presented in the same metric of the judges' responses. Because this crude coefficient is equivalent to some transformation between 0.0 and 1.0, the expected linear association between this method and the methods based on standardized coefficients for content validity is perfect or very high. Therefore, the user should focus on: a) how the evidences of the content validity is expressed (i.e., in the metric of the judges' responses, or in a range between 0.0 and 1.0), instead of the validity or accuracy of the procedures; b) the confidence level of the interval , which can be generally 90% when the number of judges is small (Merino & Livia, 2009), but the 95% or 99% levels can also be chosen); and c) in the existing computer programs to calculate them complementary results (e.g., Merino, 2018; Merino-Soto & Livia-Segovia, 2009.

Conflict of interest.- The authors of this article declare no conflict of interest.

Financial support.- No funding.

References

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

Bonett, D. G., & Price, R. M. (2020). Confidence intervals for ratios of means and *medians. Journal of Educational and Behavioral Statistics*, *20*(10), 1–21. https://doi.org/10.3102/1076998620934125

Elosua, P. (2009). ¿Existe vida más allá del SPSS? Descubre R. *Psicothema, 21*(4), 652-655.

Escrig, J., Miralles, J. M., Martinez, D., & Rivadulla, I. (2007). Intervalos de confianza: por qué usarlos. *Cirugía Española*, *81*(3), 121-125. https://doi.org/10.1016/S0009-739X(07)71281-2

Duricki, D. A., Soleman, S., & Moon, L. D. (2016). Analysis of longitudinal data from animals with missing values using SPSS. *Nature protocols, 11*(6), 1112–1129. https://doi.org/10.1038/nprot.2016.048

Gogoi, P. (2020). Application of SPSS programme in the field of social science research. *International Journal of Recent Technology and Engineering*, *8*(5), 2424-2426. https://doi.org/10.35940/ijrte.D9260.018520

Haine, D. (2019). Popularity of statistical softwares in epidemiology. Available in: https://www.denishaine.ca/blog/popepi-rmd/

Hernandez-Nieto, R. A. (2002), *Contributions to Statistical Analysis*. Merida, Venezuela: Universidad de Los Andes.

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

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382-385. https://doi.org/10.1097/00006199-198611000-00017

Masuadi, E., Mohamud, M., Almutairi, M., Alsunaidi, A., Alswayed, A. K., & Aldhafeeri, O. F. (2021). Trends in the usage of statistical software and their associated study designs in health sciences research: A bibliometric analysis. *Cureus*, *13*(1), e12639. https://doi.org/10.7759/cureus.12639

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

Merino-Soto, C., & Livia-Segovia, J. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken [Asymmetric confidence intervals for index content validity: A Visual Basic program for Aiken's V]. *Anales de Psicología, 25*(1),169-171.

Miller, J. M., & Penfield, R. D. (2005). Using the score method to construct asymmetric confidence intervals: An SAS program for content validation in scale development. *Behavior Research Methods, 37***,** 450-452. https://doi.org/10.3758/BF03192713

Moscoso, M. S., & Merino-Soto, C. (2017). Construcción y validez de contenido del Inventario de Mindfulness y Ecuanimidad: una perspectiva iberoamericana [Construction and content validity of the Mindfulness and Equanimity Inventory: an Ibero-American perspective]. *Mindfulness & Compassion,* *2*(1), 9-16. https://doi.org/10.1016/j.mincom.2017.01.001

Muenchen, R. A. (2017). The popularity of data science software. Acceso: 22/04/2021. Available in: http://r4stats.com/articles/popularity/

Osterlind, S. J. (1992). *Constructing test items: multiple-choice, constructed-response, performance, and other formats*. Boston: Kluwer Academic Publishers.

Pedrosa, I., Suárez-Álvarez, J., & García-Cueto, E. (2013). Evidencias sobre la validez de contenido: avances teóricos y métodos para su estimación. *Acción Psicológica, 10*(2), 3-18.

Penfield, R. D. (2003). A score method of constructing asymmetric confidence intervals for the mean of a rating scale item. *Psychological methods*, *8*(2), 149–163. https://doi.org/10.1037/1082-989x.8.2.149

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

Penfield, R. D., & Giacobbi, P. R. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science,* *8*, 213-225. https://doi.org/10.1207/s15327841mpee0804_

Rivadeneira Pacheco, J. L., Barrera Argüello, M. V., & De La Hoz Suárez, A. I. (2020). Análisis general del SPSS y su utilidad en la estadística [General analysis of SPSS and its usefulness in statistics]. *Journal of Business Sciences*, *2*(4), 17-25. https://revista.estudioidea.org/ojs/index.php/eidea/article/view/19

Rovinelli, R. J. & 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.

Shaikh, M. A. (2016). Use of statistical tests and statistical software choice in 2014: tale from three Medline indexed Pakistani journals. *The Journal of the Pakistan Medical Association*, *66*(4), 464-466.

Shaikh, M. A. (2017). Study designs, use of statistical tests, and statistical analysis software choice in 2015: Results from two Pakistani monthly Medline indexed journals. *The Journal of the Pakistan Medical Association*, *67*(9), 1428-1431.

Tellez, A., Garcia, C., & Corral-Verdugo, V. (2015). Effect size, confidence intervals and statistical power in psychological research**.** *Psychology in Russia: State of the Art,* *8*(3), 27-46. https://doi.org/10.11621/pir.2015.0303

Valeri, L., & Vanderweele, T. J. (2013). Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods, 18*(2), 137–150. https://doi.org/10.1037/a0031034

Vanus, J., Kubicek, J., Gorjani, O. M., & Koziorek, J. (2019). Using the IBM SPSS SW Tool with wavelet transformation for CO₂ prediction within iot in smart home care. *Sensors, 19*(6), 1407. https://doi.org/10.3390/s19061407

Willink, R. (2005). A confidence interval and test for the mean of an asymmetric distribution. *Communications in Statistics - Theory and Methods, 34*(4), 753-766. https://doi.org/10.1081/STA-200054419

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212. https://doi.org/10.1080/01621459.1927.10502953