anales de psicología / annals of psychology
2018, vol. 34, nº 3 (october), 587-590
http://dx.doi.org/10.6018/analesps.34.3.326801

ISSN print edition: 0212-9728. ISSN on line edition (http://revistas.um.es/analesps): 1695-2294.

Confidence interval for difference between coefficients of content validity (Aiken's V): A SPSS syntax

César Merino-Soto*

Universidad de San Martín de Porres (Perú).

Título: Intervalos de confianza para la diferencia entre coeficientes de validez de contenido (V Aiken): Una sintaxis SPSS.
Resumen: El análisis de la validez de contenido usualmente incluye medidas para cuantificar sumariamente las calificaciones de jueces expertos. Sin embargo, las diferencias entre grupos de jueces no han sido previamente planteadas, y es plausible que estas diferencias ocurran. El presente manuscrito plantea la comparación entre-grupos del coeficiente V (Aiken, 1980), como una estrategia que debe ser implementada en el análisis de la validez de contenido. Se presenta un método basado en la construcción de intervalos de confianza para la diferencia entre coeficientes V, para dos grupos independientes. Finalmente, también se incluye un programa en sintaxis SPSS de libre distribución para su implementación.
Palabras clave: Validez de contenido, validez, software, psicometría.

Abstract: Analysis of content validity usually includes measures to summarize the ratings of expert judges. However, differences between groups of judges have not been previously raised, and it is plausible that these differences occur. The present manuscript raises the comparison between-groups of the coefficient V (Aiken, 1980), as a strategy that must be implemented in the analysis of the content validity. We present a method based on the construction of confidence intervals for the difference between coefficients V, for two independent groups. Finally, a free SPSS syntax is also included.
Keywords: Content validity, validity, software, psychometry.

Introduction

Among the methods to quantify the validity of content, obtained through the participation of expert judges, the use of the coefficient V (Aiken, 1980, 1985), also known as V of Aiken, seems to have increased in psychological research or in fields that involve psychological constructs applied in various disciplines. The introduction of this method in Spanish language seems to be located in Escurra (1988), and later, Merino and Livia (2009) disseminated the asymmetric confidence interval (CI) approach, developed by Penfield and Giacobbi (2004). This method seems to fit very well the estimates of content validity by means of coefficients, since it is reasonable to find negative asymmetric distributions in validity judgments, especially when the instruments have rigorously sampled the content domain of the construct; in this situation, validity judgments will present denser distributions at the higher extremes of the rating. This procedure is an important advance to generate intervals in the level 100 (1 - α) % to conclude in the statistical and practical significance of the coefficient V.

In the current content validity reports that used the V coefficient (Domínguez & Villegas, 2012; Domínguez, Villegas, Yauri, Mattos, & Ramírez, 2012; Freiberg, de la Iglesia, Stover, & Fernández, 2014; Gómez, Sainz de Baranda, Ortega, Contreras & Olmedilla, 2014; García & García, 2013; García, Merino & Valero, 2015; Medrano, Liporace & Pérez, 2014; Merino & Valero, 2014; Palao, Manzanares & Ortega, 2015; Sánchez-Alcaraz & Parra-Meroño, 2013; Vallejo-Medina, Granados & Sierra, 2014), to date, the variability of expert judgment related to a fixed factor that differentiates two (or more) groups of experts has not been raised. The content validity estimated in the previous studies rests on the assumption that the results do not interact with differences between groups of experts; however, it is plausible to suppose the opposite if a rigorous analysis of the characteristics of the subjects that served as expert judges is made.

Indeed, the review of previous research suggests that the situation least raised in obtaining quantitative evidence of content validity is the comparison of groups of judges with respect to this type of evidence. This comparison occurs when there is the necessary justification to suppose that the evaluation of the content of the items is associated with the variations related to specific groups. For example, between men and women, between subjects with differentiated years of formal education (technical education vs. university education), between culturally or socioeconomically distinct groups (e.g., Peruvians and South Africans), or according to the degree of thematic or experiential expertise derived from different professional practices (eg, educational psychologists vs. forensic, or doctors and nurses). Therefore, the aim of the present manuscript is to propose a method and an SPSS syntax, for the comparison of quantitative evidences of content validity, by means of confidence intervals for the difference between coefficients V. To continue with this approach, the present work will only formulate the situation of comparing between two independent groups, since the most basic comparison between groups will be considered in the analyzes, and it will be the context of the presentation of the confidence interval method for the difference between coefficients V in two independent groups.

Confidence intervals for the difference between V coefficients

Aiken (1985) defined his coefficient V as a proportion, and used the binomial distribution to create a hypothesis test of the population value centered on .50. Subsequently, Pen-

*** Correspondence address [Dirección para correspondencia]:**
César Merino-Soto. Instituto de Investigación de Psicología, Universidad de San Martín de Porres, Av. Tomás Marsano 242 (5to piso), Lima 34 – (Perú).
E-mail: cmerinos@usmp.pe; sikayax@yahoo.com.ar

field and Giacobbi (2004) derived asymmetric confidence intervals (CI) for V based on the score method (Wilson, 1927), a procedure that also serves to generate CI for other parameters (for example, dependent and independent proportions, Newcombe & Merino, 2006). To compare coefficients V from two independent groups of judges, the present manuscript adapts a procedure based on confidence intervals for the difference of two quantities, which represent indicators of effect size of the phenomenon evaluated. This approach was derived from the general method proposed by Zou and Donner (2008), to obtain intervals of difference for estimates of effect magnitude; these intervals are a modification of the traditional approach for the same purpose (Smithson, 2003). In this context, the coefficient V is conceptualized as a measure of the magnitude or intensity of judges' judgments regarding their validity qualifications (Merino, 2013), making it possible to quantify the degree to which the items adequately represent the content domain, or the clarity of it. This method is robust to the asymmetry of the sampling distribution between two parameters and does not require assuming any known distribution (Zou, 2007; Zou & Donner, 2008). The implementation requires the sample estimates of the investigated parameter ($\hat{\theta}$), and its limits of the lower (*i*) and superior (*s*) intervals in the level 100 (1 - *a*)% previously obtained.

For its implementation in the present context, $\hat{\theta}=V$; therefore, parameter $\theta_1-\theta_2$, is $V_1 - V_2$ for each item and calculated between the two groups of judges compared. With this information, the formulation of the confidence interval for the difference, according to the described method (Zou, 2007; Zou & Donner, 2008) is calculated for the lower limit as $V_1-V_2-\sqrt{(V_1-i_1)^2+(s_2-V_2)^2}$; and the upper limit as $V_1-V_2+\sqrt{(s_1-V_1)^2+(V_2-i_2)^2}$.

The method has also been adapted to compare parameters in various analysis contexts, such as the comparison of kappa agreement coefficients and intraclass correlation for dependent samples (Donner & Zou, 2002; Ramasundarahettige, Donner & Zou, 2009), product correlations of moments and R^2 (Zou, 2007), means of normal distribution (Wang & Chow, 2002) and of lognormal distribution (Zou, Taleban & Huo, 2009a, 2009b), relative risk (Rotondi, 2014), and linear combinations of parameters (Newcombe, 2011), between others. This method is also called MOVER, *Method Of Variance Estimates Recovery* (Donner & Zou, 2002; Newcombe, 2012; Zou & Donnner, 2008).

Computer program

To implement this method, a friendly syntax has been created in the corresponding window within the SPSS program. This platform was chosen because it can be considered as a software its high frequency of use among professionals and researchers at undergraduate and graduate levels. The user must enter the information corresponding to the V coefficients and their CI 100 (1 - *a*) % obtained in each group, or the quantitative inputs to calculate it according to Merino & Livia (2009). The syntax is freely available and can be requested from the author's contact address.

Example

Recently (Merino-Soto, 2016), the clarity of the items of the Eysenck Personality Questionnaire, revised version (EPQR, Eysenck & Eysenck, 2001), was evaluated within a framework of research of content validity. In accordance with the traditional strategy, a group of judges was asked to assess the clarity of the items, but the participation of the examinees was also introduced. In this way, there were two groups: university students (n = 36) and psychology teachers (n = 7); the latter had professional and teaching experience in the teaching of psychological tests. The final objective was to reveal the plausible differences between both groups of judges, and the valuation of the subjects as legitimate evaluators of the clarity of the content of the items. The coefficient V and its asymmetric confidence intervals were applied. However, no direct quantitative comparisons were made between the coefficients V of each item of both groups. Applying the procedure presented in this manuscript, to the results exposed for the Extraversion subscale, the confidence interval for the difference for each item (95% confidence level) is shown in Table 1. The only difference that can be considered statistically significant occurred for item 69, indicating that students perceive the item more clearly compared to expert judges. The researcher should consider this discrepancy as post hoc information to evaluate it in the context of the study.

Table 1. Confidence intervals for difference between coefficients V.

EPQ-R Items	Confidence intervals for differences (90%)		Conclusion
	Lower	Upper	
3	-.105	.069	No different
6	-.061	.283	No different
12	-.270	.063	No different
16	-.111	.098	No different
22	-.051	.205	No different
25	-.096	.109	No different
27	-.055	.175	No different
28	-.079	.088	No different
31	-.143	.127	No different
39	-.037	.240	No different
46	-.198	-.004	No different
47	-.111	.098	No different
49	-.120	.091	No different
53	-.096	.109	No different
57	-.096	.109	No different
58	-.096	.076	No different
69	.019	.377	Different
70	-.151	.065	No different
77	-.106	.180	No different

Note. Author.

Final comments

The method depends on the validity of the calculated CIs for the coefficients V obtained in each group. Fortunately, an appropriate method has been proposed for small samples and insensitive to asymmetric distributions, such as those usually found in content validity judgments (Merino, 2013, Penfield & Giacobbi, 2004). This method has been implemented in a Visual Basic program (Merino & Livia, 2009).

The identification of comparison groups can follow two strategies: an a priori and a posteriori. In the a priori way, the researcher must directly support the possible discrepancy between groups regarding the judgment of validity of each one, and this framework may correspond to an exploratory direction (in which there is no previous evidence or rationality is not considered to evaluate the difference between groups) or confirmatory. In both situations, the objective of the research can be directly proposed as the evaluation of the differences in the content validity, quantified by the confidence interval for differences in coefficients V. On the other hand, in a posteriori form, the researcher evaluates the existence of possible discrepancies between groups after obtaining the coefficients V and their CIs. In any situation, it is suggested to choose the same confidence level to calculate the intervals in each group.

Finally, it is useful to observe the empirical distribution of each item evaluated by the judges, to detect bimodal distributions, heterogeneous dispersion of items between different groups or judges with extreme ratings (outlier), and quantify the consensus among the judges. Consensus or agreement among judges is a prerequisite because the coefficient V uses the mean in its definition, and the mean is interpretable when the dispersion of the data is not large and the consensus is reasonably concentrated. The user must calculate some appropriate statistic for it; however a simple and "crude" approach is to compare the observed range (maximum grade obtained - minimum grade obtained) against the theoretical range (maximum score possible - minimum possible rating); values close to 1 indicate maximum variation of the judges' qualifications.

References

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

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

Domínguez, S., & Villegas, G. (2012). Estimación de la validez de contenido de una escala de calidad de vida para personas adultas con discapacidad intelectual. *Revista de Psicología (Arequipa), 2*(2), 207-219.

Domínguez, S., Villegas, G., Yauri, C., Mattos, E., & Ramírez, F. (2012). Propiedades psicométricas de una escala de autoeficacia para situaciones académicas en estudiantes universitarios peruanos. *Revista de Psicología (Universidad Católica San Pablo)*, *2*(1), 29-39.

Donner, A. & Zou, G. Y. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics, 58*, 209–214. doi:10.1111/j.0006-341x.2002.00209.x

Donner, A., & Zou, G. Y. (2002). Interval estimation for a difference between intraclass kappa statistics. *Biometrics, 58*, 209–215. doi:10.1111/j.0006-341x.2002.00209.x

Escurra, L. (1988). Cuantificación de la validez de contenido por el criterio de jueces. *Revista de Psicología, 6*(1/2), 103-111.

Eysenck, H. J. & Eysenck, S. G. (2001). *Cuestionario de Personalidad de Eysenck, versión revisada (EPQ-R)*. Madrid: TEA.

Freiberg, A., de la Iglesia, G., Beatriz Stover, J., & Fernández, M. (2014). Paradoxical Personality Scale: Its development and construct validity analysis. *International Journal of Psychological Research*, *7*(1), 49-72. doi:10.21500/20112084.667

García, M., & García, M. (2013). Estimación de la validez de contenido en una escala de valoración de grado de violencia de género soportado en adolescentes. *Acción psicológica, 10*(2), 41-58. doi:10.5944/ap.10.2.11823

García, S., Merino, J., & Valero, A. (2015). Análisis de la opinión de los alumnos sobre la calidad de las clases de educación física impartidas por los docentes de secundaria. *Journal of Sport and Health Research, 7*(3), 181-192.

Gómez, P., Sainz de Baranda., P., Ortega, E., Contreras, O., & Olmedilla, A. (2014). Diseño y validación de un cuestionario sobre la percepción del deportista respecto a su reincorporación al entrenamiento tras una lesión. *Revista de Psicología del Deporte, 23*(2), 479-487.

Medrano, L. A., Liporace, M. F., & Pérez, E. (2014). Sistema de evaluación informatizado de la satisfacción académica para estudiantes universitarios de primer año. *Electronic Journal of Research in Educational Psychology, 12*(33), 541-562.

Merino, C. (2013, Septiembre). *Coeficiente V de Aiken recargado: Avances metodológicos para su uso*. Ponencia presentada en el Primer Congreso Regional de Psicología, del 9 al 11 de septiembre, Lima, Perú.

Merino, C., & Livia, J. (2009). Intervalos de confianza asimétricos para el índice de validez de contenido: un programa Visual Basic para la V de Aiken. *Anales de Psicología, 25*(1), 169-171.

Merino, J., & Valero, A., (2014). Validación de contenido de la escala de los estilos de enseñanza recogida en el cuestionario DEMEVI. *Habilidad Motriz, 43*, 12-24.

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. doi:10.11600/1692715x.14239120615

Newcombe, R. G. & Merino, C. (2006) Intervalos de confianza para las estimaciones de proporciones y sus diferencias entre ellas. *Interdisciplinaria, 23*(2), 141-154

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

Newcombe, R.G. (2011). Propagating imprecision: combining confidence intervals from independent sources. *Communications in Statistics—Theory and Methods, 40*, 3154–3180. doi:10.1080/03610921003764225

Palao, J., Manzanares, P., & Ortega, E. (2015). Diseño y validación de un instrumento de observación para las acciones técnico-tácticas en voleyplaya. *Motriz: Revista de Educação Física*, *21*(2), 137-147. doi:10.1590/S1980-65742015000200004

Penfield, R. D. & Giacobbi, P. R., Jr. (2004) Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in Physical Education and Exercise Science, 8*(4), 213-225. doi:10.1207/s15327841mpee0804_3

Ramasundarahettige, C. F., Donner, A. & Zou, G. Y. (2009). Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine, 28*, 1041–1053. doi:10.1002/sim.3523

Rotondi, M. A. (2014). Towards the estimation of effect measures in studies using respondent-driven sampling. *Journal of Urban Healt : Bulletin of the New York Academy of Medicine*, *91*(3), 592–597. doi:10.1007/s11524-013-9836-5

Sánchez-Alcaraz, B., & Parra-Meroño, M. (2013). Diseño y validación de un cuestionario de satisfacción laboral para técnicos deportivos (CSLTD). *Cultura, Ciencia y Deporte, 9*(8), 119-127. doi:10.12800/ccd.v8i23.296

Smithson, M. (2003). *Confidence Intervals*. (Sage University Papers Series on Quantitative Applications in the Social Sciences), 07-140, Thousand Oaks, CA: Sage. doi:10.4135/9781412983761

Vallejo-Medina, P., Granados, M. R., & Sierra, J. C. (2014). Propuesta y validación de una versión breve del Sexual Opinion Survey en población española. *Revista Internacional de Andrología*, *12*(2), 47-54. doi:10.1016/j.androl.2013.04.004

Wang, H. & Chow, S. C. (2002). A practical approach for comparing means of two groups without equal variance assumption. *Statistics in Medicine, 21*, 3137–3151. doi:10.1002/sim.1238

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, *22*, 209-212. doi:10.2307/2276774

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods, 12*, 399–413. doi:10.1037/1082-989x.12.4.399.supp

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

Zou, G. Y., Huo, C. Y., & Taleban, J. (2009a). Simple confidence intervals for lognormal means and their differences with environmental applications. *Environmetrics, 20*, 172–180. doi:10.1002/env.919

Zou, G. Y., Taleban, J. & Huo, C. Y. (2009b). Confidence interval estimation for lognormal data with application to health economics. *Computational Statistics & Data Analysis, 53*, 3755–3764. doi:10.1016/j.csda.2009.03.016

(Article received: 15-02-2017; revised: 01-07-2017; accepted: 26-08-2017)