RIDU REVISTA DIGITAL DE INVESTIGACIÓN EN DOCENCIA UNIVERSITARIA

METHODOLOGICAL ARTICLE

Reliability Estimate in Two-Item Measures: Angoff-Feldt Coefficient

Estimación de la Confiabilidad en Mediciones de Dos ítems: el Coeficiente Angoff-Feldt
Estimação da Confiabilidade em Medições de Dois itens: O Coeficiente Angoff-Feldt

Sergio Alexis Dominguez-Lara*, César Merino-Soto**, Jhonatan S. Navarro-Loli***
Instituto de Investigación de Psicología, Universidad de San Martín de Porres, Lima, Perú

Received: 3/16/16
Accepted: 5/25/2016

ABSTRACT. Nowadays, the use of brief measurements in psychological evaluations is extending rapidly, but many times the reliability analysis of the scores is limited by the conditions which must be overcome to use certain coefficients (e.g. α coefficient). The compliance of the tau equivalent measurement model is one of the most important the α coefficient demands, and as it is not achieved, the estimate may be biased. The aim of this paper is to present the Angoff-Feldt (r_{AF}) coefficient for reliability estimate, to be used in congeneric measures, that is, those that do not comply with the demands of the α coefficient (e.g. tau equivalent model). Although r_{AF} was initially thought for two congeneric halves, this paper presents the application for the measurement of two items, assuming each is a half. An example of its usefulness is provided and the possibility of implementing its use discussed.

Keywords: reliability, Angoff-Feldt, congeneric measures

RESUMEN. En la actualidad el uso de medidas breves en la evaluación psicológica se extiende con rapidez, aunque en muchas ocasiones el análisis de la confiabilidad de sus puntajes se ve limitado por algunas condiciones que deben superarse para utilizar determinados coeficientes (p.e., el coeficiente α). El cumplimiento del modelo tau-equivalente es uno de los requerimientos más importantes que exige el coeficiente α y al no alcanzarse, la estimación puede ser sesgada. El objetivo de este artículo es presentar el coeficiente Angoff-Feldt (r_{AF}) para la estimación de la confiabilidad, que puede ser usado en medidas congenéricas, es

Palabras clave: confiabilidad, Angoff-Feldt, medidas congenéricas

Cite as: Domínguez-Lara, S., Merino-Soto, C. & Navarro-Loli, J. (2016). Estimación de la Confiabilidad en Mediciones de Dos ítems: el Coeficiente Angoff-Feldt. [Reliability Estimate in Two-Item Measures: Angoff-Feldt Coefficient]. *Revista Digital de Investigación en Docencia Universitaria, 10*(1), 34-40. **doi:** http://dx.doi.org/10.19083/ridu.10.463
*** E-mail:** sdominguezl@usmp.pe, sdominguezmpcs@gmail.com
**** E-mail:** cmerinos@usmp.pe, sikayax@yahoo.com.ar
**** E-mail:** jnavarrol1@usmp.pe

decir, que no cumplan con las exigencias del coeficiente α (p.e., modelo tau-equivalente). Si bien el *rAF* se pensó inicialmente para dos mitades congenéricas, en el artículo se presenta la aplicación para medidas de dos ítems, asumiendo que cada uno es una mitad. Se brinda un ejemplo sobre su utilidad, y se discute la posibilidad de implementar su uso.

RESUMO. Hoje em dia o uso de medidas breves na avaliação psicológica se desenvolve cada vez mais rápido, mas em muitas ocasiões a análise de confiabilidade de suas pontuações se limita pelas condições que devem superar-se para utilizar determinados coeficientes (p.e., o coeficiente α). O cumprimento do modelo tau-equivalente é uma das mais importantes que exige o coeficiente α e ao não se alcançar, a estimação pode ser um víeis. O objetivo deste artigo é apresentar o coeficiente Angoff-Feldt (*rAF*) para a estimação da confiabilidade, que possa ser usado em medidas congenéricas, ou seja, que não cumpram com as exigências do coeficiente α (p.e., modelo tau-equivalente). Enquanto o *rAF* inicialmente pensado para duas metades congenéricas, no artigo relata a aplicação para medidas de dois itens, assumindo que cada um é uma metade. Exemplifica-se sobre sua utilidade, e menciona-se a possibilidade de implementar seu uso.

Palavras-chave: confiabilidade, Angoff-Feldt, medidas congenéricas

Apparently, the current trend in psychological assessment is two brief measures (Hain, Schermelleh-Engel, Freitag, Louwen, & Oddo, 2015). The ones that stand out are those that contain two items per factor, whether unidimensional (Bennett et al., 2008; Brown, Leonard, Saunders, & Papasouliotis, 2001; Jensen, Keefe, Lefebvre, Romano, & Turner, 2003; Kroenke, Spitzer, & Williams, 2003; Kroenke, Spitzer, Williams, Monahan, & Lowe, 2007; Minoura, & Narita, 2013) or multidimensional instruments (Cakmak & Cevik, 2010; Dominguez & Merino, 2015; Garnefski & Kraaij, 2006; Gosling, Rentfrow, & Swann, 2003). Two strong reasons for this are how quick they can be applied and qualified, and the satisfactory evidence on their equivalence with extensive versions. These advantages make it a viable alternative when conditions for assessment are not the most optimal in terms of time and subject's disposition (Robins, Hendin, & Trzeaniewski, 2001), or when a wide array of constructs is covered through an assessment battery. Nevertheless, the practicality its brevity entails presents methodological limitations such as the reliability estimate. To estimate the reliability of the scores, the coefficient α (Cronbach, 1951) is the most frequently used estimator by psychology researchers (Elosua & Zumbo, 2008; Ledesma, 2002; Zumbo & Rupp, 2004). Even with this extended use, many times it is not known if the data comply with some of the conditions required for application, such as the appropriate measurement model.

The Classical Test Theory (CTT) explains that the observed score (X) is composed of the true score ($\tau\alpha$) and the measurement error (ε): $X = \tau\alpha + \varepsilon$, where $\tau\alpha = \lambda\tau + s$, being λ and s multiplying and additive constants, respectively. It is also assumed that errors are not correlated, not even with the true scores. Based on this framework, measurement error estimation poses the definition of the appropriate statistical method for its definition. There are three models briefly describe below. Based on the most restrictive measurement model, the parallel model supposes that the items measure the same construct, present the same error variance, true score (τ) and additive constant (s) (Eisinga, Te Grotenhuis, & Pelzer, 2012), but also the same observed score and variances (Meyer, 2010). Thus, for two items the scheme would be: $\tau_1 = \lambda\tau + s_1$; $\tau_2 = \lambda\tau + s_2$; where $\tau_1 = \tau_2$ y $s_1 = s_2$. On the other hand, the *essentially tau-equivalent* measurement model

ISNN 2223-2516

(*essentially τ – equivalent*) assumes all items measure the same construct, on the same scale, the same error variance, and all with similar true scores (Graham, 2006), but the observed variances and scores may vary thanks to the additive constant (Eisinga et al., 2012; Meyer, 2010; Warrens, 2015). The scheme is: $\tau_1 = \lambda\tau + s_1$; $\tau_2 = \lambda\tau + s_2$; where $\tau_1 = \tau_2$ y $s_1 \neq s_2$. The *tau-equivalent* (*τ - equivalent*) measurement model indicates that the items present the same true score, but may have different error variance (Eisinga et al., 2012). Finally, the congeneric measurement model assumes that the items measure the same construct, have different error variances, and each one's true score (*τ*) may vary due to the multiplying and additive constants (λ and s) (Eisinga et al., 2012; Warrens, 2015). The scheme is: $\tau_1 = \lambda_1\tau + s_1$; $\tau_2 = \lambda_2\tau + s_2$; where $\tau_1 \neq \tau_2$, $\lambda_1 \neq \lambda_2$, y $s_1 \neq s_2$. In reality, the distribution of the *parallel and essentially tau-equivalent* measurement models present many restrictions (Sijtsma, 2012) and *tau-equivalent* is not likely to be obtained (Cortina, 1993).

To appropriately estimate the score reliability through coefficient α, we expect items to be at least *tau-equivalent* (Graham, 2006), otherwise there is the probability for it to be underestimated (Eisinga et al., 2012; Graham, 2006). Although coefficient α is considered the bottom confidence limit, there is evidence that failure to fulfill requirements for estimation increases or reduces magnitude spuriously. Hence, it is necessary to have analysis alternatives to face this limitation and reach a more precise estimate. A viable procedure is to calculate the Spearman-Brown (SB) coefficient, where each item would be considered a half of the test. Such coefficient, even in the absence of *tau-equivalent* and complying only with the *congeneric* model, seems to work adequately in the two-item analysis (Eisinga et al., 2012). However, a problem with SB is that it was created under the *parallel* measurement assumption (Warrens, 2016), and therefore tends to overestimate the coefficient. Additionally, although there is evidence about the equivalence of α with SB (Warrens, 2015), the results may be biased when the two items do not comply with the appropriate measurement models.

Consequently, it can be concluded that both coefficients are used assuming the universality of their application with no regards to the statistical characteristics of their elements and the conditions under which they would have larger support. And if we also consider that the measurements obtained from reality are not close to the *tau-equivalent* measurement model (Cortina, 1993), then reliability estimate alternatives, for instance, the Angoff-Feldt (Feldt, & Charter, 2003) coefficient are needed.

THE ANGOFF-FELDT COEFFICIENT

A reliable reliability estimator that can be used for continuous variables and constructed specifically for congeneric halves, i.e. not affected by the coefficient limitations that require the fulfillment of the *tau-equivalent* or *parallel* measurement, is the Angoff-Feldt coefficient (rAF; Angoff, 1953; Feldt, 1975; Feldt & Brennan, 1989):

$$r_{AF} = \frac{4\left(r_{12}\right) SD_{X_1} SD_{X_2}}{SD^{2}_{TOTAL} - \frac{\left(SD_{X_1} SD_{X_2}\right)^2}{SD^{2}_{TOTAL}}}$$

Where r12 is the correlation between the halves (in this case, both items); SD_{X_1}, SD_{X_2}, and SD_{TOTAL} are standard deviations (SD) of the first half (item 1), of the second half (item 2) and of the total score, respectively. The square of each one is the corresponding variance. When the difference between the SD of the items is noticeable, unfulfilling the condition that *parallel* measurements have the same arithmetic mean and SD (Feldt & Charter, 2003), the value rAF moves remarkably away from the other coefficients. Consequently, the use of rAF is recommendable, given that is was conceived for congeneric measurements.

A decision-making system with regards to the use of one coefficient or the other: the standard deviation ratio, has been proposed: $SD_{greater}/SD_{smaller}$ (Feldt & Charter, 2003; Warrens, 2015). If the differences are not large ($SD_{greater}/SD_{smaller} \leqslant 1.15$), both SB and α may be used; if they are moderate ($1.15 < SD_{greater}/SD_{smaller} \leqslant 1.30$), it is better to use α or rAF. But if the differences are large ($1.30 < SD_{greater}/SD_{smaller}$), it is assumed that those differences are significant, and only using rAF is recommended.

ISNN 2223-2516

Finally, the interpretation of r_{AF} may be made under the same standards as coefficient α, since the same procedures may be used to obtain the significance tests and confidence intervals (Feldt, 2002).

SOFTWARE COMPLEMENT

Readers are offered an SPSS syntax that may be requested by writing to any of the authors. Upon introducing the necessary data, the r_{AF}, SB and α are easily calculated.

APPLICATIONS

To exemplify the process, we will consider the reliability estimate through three methods (α, SB, r_{AF}) in an intentional sample of 230 psychology students from a private university (73% women), with ages between 17 and 62 (M = 23.671; SD = 6.542).

Scores were obtained from GAD-2 (Kroenke et al., 2003) and PHQ-2 (Kroenke et al., 2007), each one of them comprising two items that assess anxiety and depression, respectively. Since the variances of the items are equivalent for PHQ-2 ($SD_{greater}/SD_{smaller}$ = 1.129) and GAD-2 ($SD_{greater}/SD_{smaller}$ = .868), no great differences in terms of the coefficients calculated (situation 1) were observed.

Thus, in order to exemplify the impact on the coefficient estimate when SDs are equivalent or different, the SD of the items was subject to a change keeping correlations constant; therefore, the new ratio between the SDs indicate moderate differences (Situation 2; $SD_{greater}/SD_{smaller}$ = 1.351) and large differences (Situation 3; $SD_{greater}/SD_{smaller}$ = 1.875). Results show a sequential decrease in the α magnitude coefficient while the proportion between $SD_{greater}/SD_{smaller}$ is larger (see Table 1).

Table 1

Score Reliability Estimate with Three Measurement Methods for Two Items

		M	*SD*	SD^2	SD^2_{tot}	r_{12}	*α*	SB	r_{AF}
Situation 1 ($SD_{greater}/SD_{smaller} \leq 1.15$)									
PHQ-2	Ítem 1	0.50	0.710	0.504	1.294	.441	.609	.612	.613
	Ítem 2	0.39	0.629	0.396					
GAD-2	Ítem 1	0.42	0.627	0.693	1.390	.526	.685	.689	.691
	Ítem 2	0.52	0.722	0.521					
Situación 2 ($1.15 < SD_{greater}/SD_{smaller} \leq 1.30$)									
PHQ-2	Ítem 1	0.50	0.850	0.723	1.589	.441	.593	.612	.619
	Ítem 2	0.39	0.629	0.396					
GAD-2	Ítem 1	0.42	0.850	0.723	1.680	.526	.669	.689	.696
	Ítem 2	0.52	0.629	0.396					
Situación 3 ($1.30 < SD_{greater}/SD_{smaller}$)									
PHQ-2	Ítem 1	0.50	0.75	0.563	0.987	.441	.536	.612	.643
	Ítem 2	0.39	0.40	0.160					
GAD-2	Ítem 1	0.42	0.75	0.563	1.038	.526	.608	.689	.716
	Ítem 2	0.52	0.40	0.160					

Note: *n = 230; M = median; SD: standard deviation; SD^2: variance; r_{12}: Pearson correlation coefficient; α: alfa coefficient; SB: Spearman-Brown; r_{AF}: Angoff-Feldt; PHQ-2 = The Patient Health Questionnaire-2; GAD-2 = Generalized Anxiety Disorder Scale*

ISNN 2223-2516

FINAL COMMENTS

For the implementation of the methods proposed, it is *sine qua non* to confirm the fulfillment of the statistical assumptions and must be a recommended (Graham, 2006) and routine practice in all reliability reports. On the other hand, it should be noted that the reliability estimates presented are within the framework of the scores observed, but there are other theoretical frameworks that may be relevant, for example, the one applicable based on the model of latent variables or structural equation modeling (SEM), which is especially useful to prove the psychometric assumptions for each reliability model. Within this framework, the approaches for congeneric measurements reliability have been developed satisfactorily (Fornell & Laker, 1981; Raykov, 1997). In contrast to rAF, these procedures use measurements based on latent variables, so their interpretation is closer to construct reliability (Hancock & Mueller, 2001).

Finally, the results presented respond to particular cases, so they do not necessarily represent all of the possible situations. However, it is useful to exemplify the use of rAF and recommend its estimate when the tau-equivalent model assumptions are not fulfilled. Although, on occasions, the three coefficients may coincide in terms of magnitude (Warrens, 2016), when the necessary conditions to apply one or the other are lost, or when the statistics vary depending on the sample (Sánchez-Meca & López-Pina, 2008), the values obtained move farther away from each other in the PHQ-2 (see Figure 1) and in the GAD-2 (see Figure 2), even underestimating the real reliability value of their scores if only coefficient α is reported (Eisinga et al., 2012; Feldt & Charter, 2003) for two-item measurements.

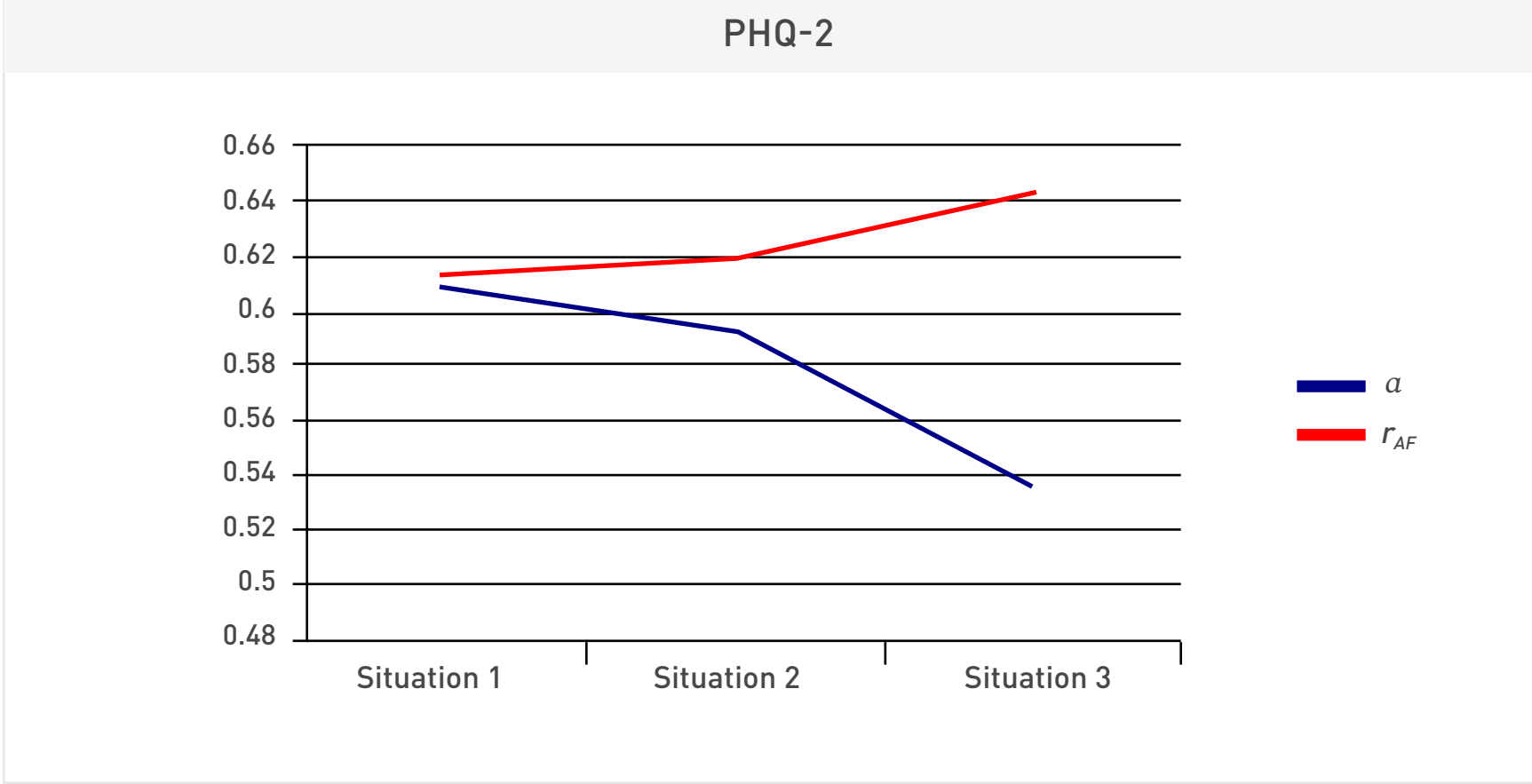

Figure 1. *Variation of confidence coefficients according to situations PHQ-2*

ISNN 2223-2516

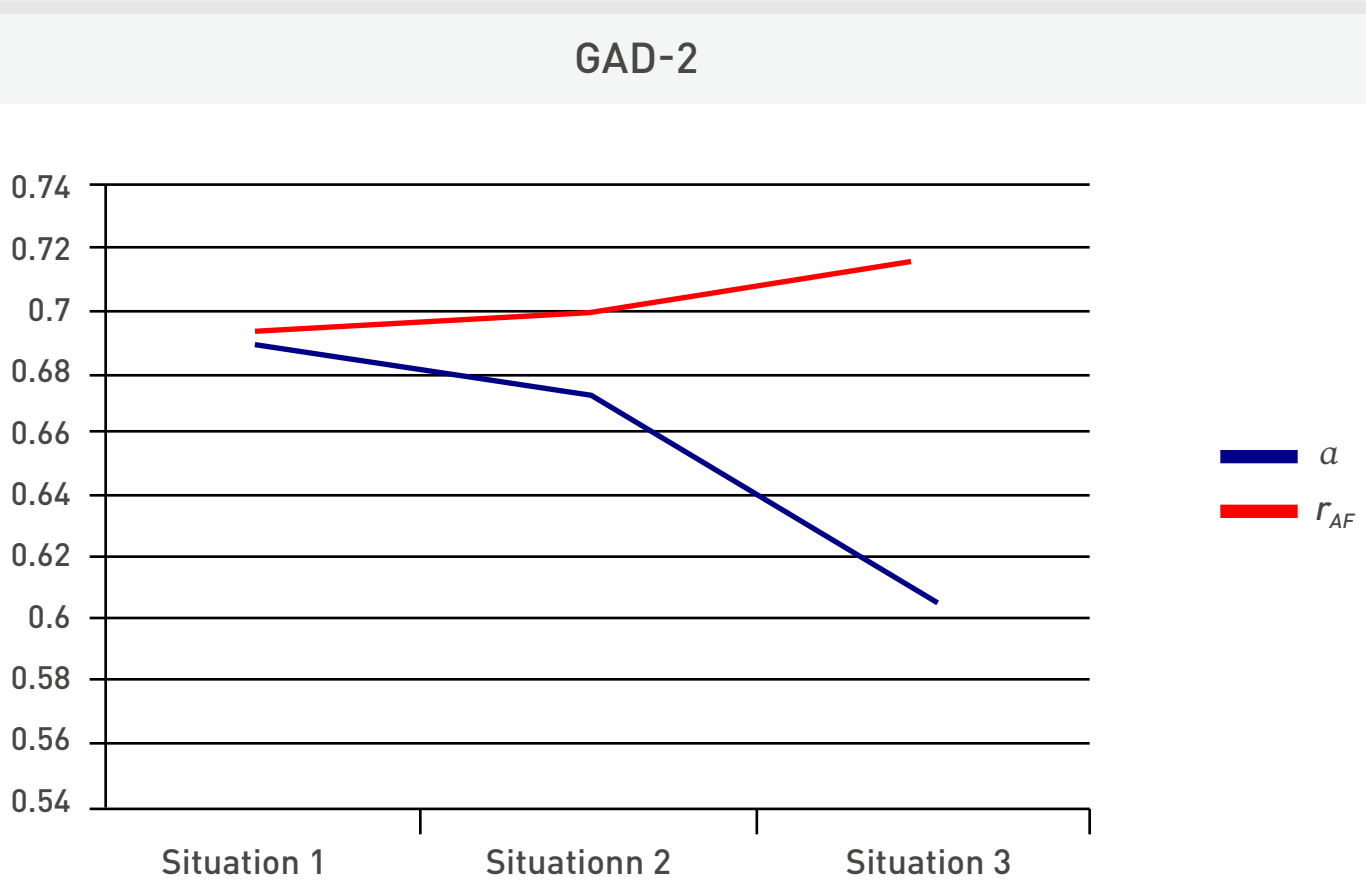

Figure 2. *Variation of confidence coefficients according to situations GAD-2*

REFERENCES

Angoff, W. H. (1953). Test reliability and effective test length. *Psychometrika, 18*(1), 1-14. doi:10.1007/BF02289023

Bennett, I. M., Coco, A., Coyne, J. C., Mitchell, A. J., Nicholson, J., Johnson, E., Horst, M., & Ratcliffe, S. (2008). Efficiency of a Two-Item Pre-Screen to Reduce the Burden of Depression Screening in Pregnancy and Postpartum: An IMPLICIT Network Study. *The American Board of Family Medicine, 21*(4), 317-325. doi:10.3122/jabfm.2008.04.080048

Brown, R. L., Leonard, T., Saunders, L. A., & Papasouliotis, O. (2001). A two-item conjoint screen for alcohol and other drug problems. *The American Board of Family Medicine, 14*(2), 95-106.

Cakmak, A., & Cevik, E. (2010). Cognitive emotion regulation questionnarie: Development of Turkish version of 18-item short form. *African Journal of Business Management, 4*(10), 2097-2102.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. doi:10.1007/BF02310555

Dominguez, S., & Merino, C. (2015). Una versión breve del Cognitive Emotional Regulation Questionnarie: Análisis estructural del CERQ-18 en estudiantes universitarios limeños. *Revista Peruana de Psicología y Trabajo Social, 4*(1), 25-36.

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2012). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown. *International Journal of Public Health, 58*(4), 637–642. doi:10.1007/s00038-012-0416-3

Elosua, P. & Zumbo, B. (2008). Coeficiente alfa para escalas de respuesta categórica ordenada. [Reliability Coefficients for Ordinal Response Scales]. *Psicothema, 20*(4), 896-901.

Feldt, L.S. (1975). Estimation of reliability of a test divided into two parts of unequal length. *Psychometrika, 40*(4), 557–561. doi:10.1007/BF02291556

Feldt L.S. (2002). Reliability Estimation When a Test Is Split Into Two Parts of Unknown Effective Length. *Measurement in Education, 15*(3), 295-308. doi:10.1207/S15324818AME1503_4

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test Split into two parts of equal or unequal length. *Psychological*

Methods, 8(1), 102-109. doi:10.1037/1082-989X.8.1.102

Feldt, L. S., & Brennan, R. L. (1989). Reliability. En Linn, R. L. (Ed), *Educational Measurement (*pp. 105 – 146). New York: Macmillan.

Fornell, C., & Larcker, D.F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. doi:10.2307/3151312

Garnefski, N., & Kraaij, V. (2006). Cognitive emotion regulation questionnarie-development of a short 18-item version (CERQ-short). P*ersonality and Individual Differences, 4*1(6), 1045-1053. doi:10.1016/j.paid.2006.04.010

Gosling, S., Rentfrow, P., & Swann, W., (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37*, 504-528. doi:10.1016/S0092-6566(03)00046-1

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability: what they are and how to use them. *Educacional and Psychological Measurement, 66*(6), 930 – 944. doi:10.1177/0013164406288165

Hain, S., Schermelleh-Engel, K., Freitag, C., Louwen, F., & Oddo, S. (2015). Development of a Short Form of the Personality Styles and Disorder Inventory (PSDI-6). *European Journal of Psychological Assessment*. Recuperado de: http://econtent.hogrefe.com/doi/full/10.1027/1015-5759/a000260.

Hancock, G. R., & Mueller, R. O. (2001). Rethinking construct reliability within latent variable systems. En R. Cudeck, S. H. C. du Toit & D. Sörbom (Eds.), S*tructural equation modeling: Past and present. A Festschrift in honor of Karl G. Jöreskog* (pp. 195-261). Chicago: Scientific Software International.

Jensen, M. P., Keefe, F. J., Lefebvre, J. C., Romano, J. M., & Turner, J. A. (2003). One- and two- item measures of pain beliefs and coping strategies. *Pain, 104*(3), 453-469. doi:10.1016/S0304-3959(03)00076-9

Kroenke, K., Spitzer, R., & Williams, J. (2003). The Patient Health Questionnarie-2: Validity of Two-item Depression Screener. *Medical Care, 41*(11), 1284-1292.

Kroenke, K., Spitzer, R., Williams, J., Monahan, P. & Lowe, B. (2007). Anxiety disorders in primary care: prevalence, impairment, comorbidity, and detection. A*nnals of International Medicine, 146*, 317-325.

Ledesma, R. (2002). Análisis de consistencia interna mediante Alfa de Cronbach: un programa basado en gráficos dinámicos. [Internal Consistency Analysis by Means of Cronbach's Alpha: a Computer Program Based on Dynamic Graphics]. *Psico-USF, 7*(2), 143-152. doi:10.1590/S1413-82712002000200003

Meyer, J. P. (2010). *Reliability*. New York: Oxford University Press.

Minoura, Y., & Narita, K. (2013). The development of the Two-Item Self-Esteem scale (TISE): Reliability and validity. *Japanese Journal of Research on Emotions, 21*(1), 37-45. doi:10.4092/jsre.21.37

Rae, G. (2006). Correcting coefficient alpha for correlated errors: Is α_k a lower Bound to reliability? *Applied Psychological Measurement, 30*(1), 56-59. doi:10.1177/0146621605280355

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 22*(2), 173-184. doi:10.1177/01466216970212006

Robins, R., Hendin, H., & Trzeaniewski, K. (2001). Measuring Global Self-Esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27*(2), 151-161. doi:10.1177/0146167201272002

Sánchez-Meca, J., & López-Pina, J. (2008). El enfoque meta-analítico de generalización de la fiabilidad. [The Meta-analytic Approach of Reliability Generalization]. *Acción Psicológica, 5*(2), 37-64.

Sijtsma, K. (2011). Future of psychometrics: Ask what psychometrics can do for Psychology. *Psychometrika, 77*(1), 4-20. doi:10.1007/s11336-011-9242-4

Warrens, M. J. (2015). Some relationships between Cronbach´s alpha and the Spearman-Brown formula. *Journal of Classification, 32*(1), 127-137. doi:10.1007/s00357-015-9168-0

Warrens, M. J. (2016). A comparison of reliability coefficients for psychometric tests that consist of two parts. *Advances in Data Analysis and Classification, 10*(1), 71–84. doi:10.1007/s11634-015-0198-6

Zumbo, B. D., & Rupp, A. A. (2004). Responsible modelling of measurement data for appropriate inferences: Important advances in reliability and validity theory. En D. Kaplan (Ed.): *The SAGE Handbook of Quantitative Methodology for the Social Sciences* (pp. 73-92). Thousand Oaks, CA: Sage Press