

Estimación de la confiabilidad en mediciones de dos ítems: el coeficiente Angoff-Feldt

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

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

Recibido: 16/03/16
Aceptado: 25/05/16

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 congénicas, es decir, que no cumplan con las exigencias del coeficiente α (p.e., modelo tau-equivalente). Si bien el r_{AF} se pensó inicialmente para dos mitades congénicas, 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.

Palabras clave:
confiabilidad,
Angoff-Feldt,
medidas
congénicas

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

Keywords:
reliability,
Angoff-Feldt,
congeneric
measures

Citar como: Domínguez-Lara, S., Merino-Soto, C. & Navarro-Loli, J. (2016). Estimación de la Confiabilidad en Mediciones de Dos ítems: el Coeficiente Angoff-Feldt. *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: jnavarro11@usmp.pe

a 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.

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 véis. O objetivo deste artigo é apresentar o coeficiente Angoff-Feldt (r_{AF}) para a estimação da confiabilidade, que possa ser usado em medidas congénicas, ou seja, que não cumpram com as exigências do coeficiente α (p.e., modelo tau-equivalente). Enquanto o r_{AF} inicialmente pensado para duas metades congénicas, 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
congénicas

Aparentemente la tendencia actual en la evaluación psicológica es preferir medidas breves (Hain, Schermelleh-Engel, Freitag, Louwen, & Oddo, 2015), entre las que destacan aquellas que constan de dos ítems por factor, ya sea en instrumentos unidimensionales (Bennett et al., 2008; Brown, Leonard, Saunders, & Papanoulitis, 2001; Jensen, Keefe, Lefebvre, Romano, & Turner, 2003; Kroenke, Spitzer, & Williams, 2003; Kroenke, Spitzer, Williams, Monahan, & Lowe, 2007; Minoura, & Narita, 2013) o multidimensionales (Cakmak, & Cevik, 2010; Dominguez, & Merino, 2015; Garnefski, & Kraaij, 2006; Gosling, Rentfrow, & Swann, 2003). Dos fuertes motivaciones para esto son la rapidez en cuanto a su aplicación y calificación, y la evidencia satisfactoria sobre su equivalencia con las versiones extensas. Estas ventajas las hacen una alternativa viable cuando las condiciones para la evaluación no son las más óptimas en cuanto a tiempo o disposición del evaluado (Robins, Hendin, y Trzeaniewski, 2001), o cuando se pretende cubrir un amplio espectro de constructos mediante una batería de evaluación. No obstante, la practicidad que supone su brevedad presenta limitaciones metodológicas como la estimación de la confiabilidad. Para estimar la confiabilidad de los puntajes, el

coeficiente α (Cronbach, 1951) es el estimador más usado por los investigadores en psicología (Elosua & Zumbo, 2008; Ledesma, 2002; Zumbo, & Rupp, 2004). Aun con su uso extendido, muchas veces no se conoce si los datos cumplen algunas condiciones requeridas para su aplicación, como por ejemplo, el modelo de medición apropiado.

La teoría clásica de los tests indica que el puntaje observado (X) está compuesto por el puntaje verdadero (τ) y el error de medición (ϵ): $X = \tau + \epsilon$, donde $\tau = \lambda\tau + s$, siendo λ y s constantes multiplicativas y aditivas, respectivamente. Se asume además que los errores no se encuentran correlacionados entre sí, ni con el puntaje verdadero. Desde este marco, la estimación del error de medición plantea la definición del modelo estadístico apropiado para definirlo. Existen tres modelos, que se describen brevemente a continuación. Partiendo desde el modelo de medición más restrictivo, el modelo paralelo supone que los ítems miden el mismo constructo, presentan la misma varianza del error, puntaje verdadero (τ) y constante aditiva (s) (Eisinga, Te Grotenhuis, & Pelzer, 2012), pero además el mismo puntaje y varianzas observadas (Meyer, 2010).

Entonces, para dos ítems el esquema sería: $T1 = \lambda\tau + s1$; $T2 = \lambda\tau + s2$; donde $T1 = T2$ y $s1 = s2$. Por otro lado, el modelo de medición tau-equivalencia esencial (τ – equivalencia esencial) asume que todos los ítems miden el mismo constructo, en la misma escala, la misma varianza del error, y con todos los puntajes verdaderos similares (Graham, 2006), pero las varianzas y puntajes observados pueden variar debido a la constante aditiva (Eisinga et al., 2012; Meyer, 2010; Warrens, 2015). El Esquema es: $T1 = \lambda\tau + s1$; $T2 = \lambda\tau + s2$; donde $T1 = T2$ y $s1 \neq s2$. El modelo de medida de tau-equivalencia (τ – equivalencia) indica que los ítems presentan el mismo puntaje verdadero, pero pueden tener diferente varianza del error (Eisinga et al., 2012). Finalmente, el modelo de medida congénico asume que los ítems miden el mismo constructo, tienen distintas varianza de error, y el puntaje verdadero (τ) de cada uno puede variar a causa de las constantes multiplicativa y aditiva (λ y s) (Eisinga et al., 2012; Warrens, 2015). El esquema es: $T1 = \lambda1\tau + s1$; $T2 = \lambda2\tau + s2$; donde $T1 \neq T2$, $\lambda1 \neq \lambda2$, y $s1 \neq s2$. En la realidad, la distribución de los modelos de medición paralelo y tau equivalencia esencial presentan muchas restricciones (Sijtsma, 2012) y la tau-equivalencia es poco probable de obtener (Cortina, 1993).

Para estimar apropiadamente la confiabilidad de los puntajes mediante el coeficiente α , se espera que los ítems sean al menos tau-equivalentes (Graham, 2006); de lo contrario, existe la probabilidad que sea infraestimada (Eisinga et al., 2012; Graham, 2006). Si bien se considera al coeficiente α como el límite inferior de la confiabilidad, existe evidencia de que el incumplimiento de los requisitos para su estimación incrementa o reduce de forma espuria su magnitud (Rae, 2006). Por lo tanto, es necesario contar con alternativas de análisis para hacer frente a esta limitación y llegar a una estimación más precisa. Un procedimiento viable es calcular el coeficiente Spearman-Brown (SB), donde cada ítem se consideraría una mitad del test. Dicho coeficiente, aún en ausencia de la tau-equivalencia y cumpliendo solo con el modelo congénico, parece funcionar adecuadamente en el análisis de dos ítems (Eisinga et al., 2012). No obstante, un problema con el SB es que fue creado bajo el supuesto de medidas paralelas (Warrens, 2016), por lo que tiende a sobreestimar el coeficiente. Inclusive, a pesar de que existe evidencia acerca de la equivalencia

del α con el SB (Warrens, 2015), los resultados podrían resultar sesgados cuando los dos ítems no cumplen con los modelos de medición apropiados.

Entonces, se puede concluir que ambos coeficientes son usados asumiendo la universalidad de su aplicación sin considerar las características estadísticas de sus componentes y las condiciones en las cuales gozarían de mayor sustento, y si además se considera que las medidas obtenidas de la realidad no se aproximan al modelo de medición de tau equivalencia (Cortina, 1993), surge la necesidad de contar con alternativas de estimación de la confiabilidad como por ejemplo el coeficiente Angoff-Feldt (Feldt, & Charter, 2003).

EL COEFICIENTE ANGOFF-FELDT

Un estimador de la confiabilidad que se puede utilizar para variables continuas, y construido de forma específica para mitades congénicas, es decir, que no es afectado por las limitaciones de coeficientes que requieren el cumplimiento de tau-equivalencia o paralelas, es el coeficiente Angoff-Feldt (r_{AF} ; Angoff, 1953; Feldt, 1975; Feldt, & Brennan, 1989):

$$r_{AF} = \frac{4 (r_{12}) DE_{X_1} DE_{X_2}}{DE_{TOTAL}^2 - \frac{(DE_{X_1} DE_{X_2})^2}{DE_{TOTAL}^2}}$$

Donde r_{12} es la correlación entre las mitades (en este caso, los dos ítems); DE_{X_1} , DE_{X_2} , y DE_{TOTAL} son las desviaciones estándar (DE) de la primera mitad (ítem 1), de la segunda mitad (ítem 2) y del puntaje total, respectivamente; El cuadrado de cada uno es la respectiva varianza. Cuando la diferencia entre las DE de los ítems es muy marcada, incumpliendo la condición de que las medidas paralelas tienen la misma media aritmética y DE (Feldt, & Charter, 2003), el valor de r_{AF} se aleja notablemente de los otros coeficientes. Por tal motivo, es recomendable utilizar el r_{AF} dado que fue concebido para medidas *congénicas*.

Se ha propuesto un sistema para tomar decisiones respecto del uso de un coeficiente u otro: la razón de desviaciones estándar: DE_{mayor}/DE_{menor} (Feldt, & Charter, 2003; Warrens, 2015). Si las diferencias no son grandes

($DE_{mayor}/DE_{menor} \leq 1.15$), pueden usarse tanto el SB como el α ; si son moderadas ($1.15 < DE_{mayor}/DE_{menor} \leq 1.30$) es mejor usar el α o el r_{AF} . Pero si las diferencias son grandes ($1.30 < DE_{mayor}/DE_{menor}$), se asume que dichas diferencias son significativas, y se recomendaría el uso solo del r_{AF} .

Finalmente, la interpretación del r_{AF} puede realizarse bajo los mismos estándares que el coeficiente α , ya que pueden utilizarse los mismos procedimientos para obtener las pruebas de significancia e intervalos de confianza (Feldt, 2002).

COMPLEMENTO INFORMÁTICO

Se ofrece al lector una sintaxis de SPSS que puede solicitarla escribiendo a alguno de los autores. Al introducir los datos necesarios se pueden calcular fácilmente los coeficientes r_{AF} , SB y α .

APLICACIONES

Para ejemplificar el proceso, se considerará la estimación de la confiabilidad mediante tres métodos (α , SB, r_{AF}) en una muestra intencional de 230

estudiantes de psicología de una universidad privada (73% mujeres), con edades comprendidas entre 17 y 62 años ($M = 23.671$; $DE = 6.542$).

Los puntajes se obtuvieron del GAD-2 (Kroenke et al., 2003) y el PHQ-2 (Kroenke et al., 2007), cada uno de los cuales consta de dos ítems que evalúan ansiedad y depresión, respectivamente. Dado que las varianzas de sus ítems son equivalentes tanto en el PHQ-2 ($DE_{mayor}/DE_{menor} = 1.129$) como en el GAD-2 ($DE_{mayor}/DE_{menor} = .868$), no se observaron grandes diferencias en cuanto a los coeficientes calculados (situación 1).

Entonces, a fin de ejemplificar su impacto en la estimación de los coeficientes cuando las DE son equivalentes o distintas, se realizó un cambio en la DE de los ítems manteniendo constantes las correlaciones, motivo por el cual, la nueva razón entre las DE indican diferencias moderadas (Situación 2; $DE_{mayor}/DE_{menor} = 1.351$) y grandes (Situación 3; $DE_{mayor}/DE_{menor} = 1.875$). Los resultados muestran un decremento secuencial en la magnitud del coeficiente α a medida que la proporción entre DE_{mayor}/DE_{menor} es mayor (ver Tabla 1).

Tabla 1

Estimación de la confiabilidad de los puntajes con tres métodos en medidas de dos ítems

		<i>M</i>	<i>DE</i>	<i>DE</i> ²	<i>DE</i> ² _{tot}	<i>r</i> ₁₂	<i>a</i>	SB	<i>r</i> _{AF}
Situación 1 ($DE_{mayor}/DE_{menor} \leq 1.15$)									
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 ($1.15 < DE_{mayor}/DE_{menor} \leq 1.30$)									
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 ($1.30 < DE_{mayor}/DE_{menor}$)									
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					

Nota: $n = 230$; $M =$ Media; $DE:$ desviación estándar; $DE^2:$ varianza; $r_{12}:$ coeficiente de correlación de Pearson; $\alpha:$ coeficiente alfa; SB: Spearman-Brown; $r_{AF}:$ Angoff-Feldt; PHQ-2 = The Patient Health Questionnaire-2; GAD-2 = Generalized Anxiety Disorder Scale

COMENTARIOS FINALES

Para la implementación de los métodos propuestos, es *sine qua non* comprobar el cumplimiento de sus supuestos estadísticos y debe ser una práctica recomendada (Graham, 2006) y rutinaria en todo reporte de la confiabilidad. Por otro lado, debe anotarse que las estimaciones de confiabilidad presentadas se encuentran en el marco de los puntajes observados, pero existen otros marcos teóricos que pueden ser relevantes; por ejemplo, el que se puede aplicar desde el modelo de variables latentes o el modelamiento de ecuaciones estructurales (Structural Equation Modeling, SEM), el cual es especialmente útil para probar los supuestos psicométricos para cada modelo de confiabilidad. En este marco también los planteamientos acerca de la confiabilidad para medidas congénicas se han desarrollado satisfactoriamente (Fornell, & Laker, 1981; Raykov, 1997). A diferencia del r_{AF} esos procedimientos utilizan medidas basadas en variables latentes, por lo que su interpretación

se aproxima más a la confiabilidad del constructo (Hancock, & Mueller, 2001).

Finalmente, los resultados presentados responden a casos particulares, por lo cual no representan necesariamente todas las situaciones que podrían acontecer. Sin embargo, sirve para ejemplificar el uso del r_{AF} y recomendar su estimación cuando no se cumplen los supuestos del modelo tau equivalente. Si bien en ocasiones podrían coincidir los tres coeficientes en cuanto a su magnitud (Warrens, 2016), a medida que se pierden las condiciones necesarias para aplicar uno u otro, o cuando los estadísticos varían dependiendo de la muestra (Sánchez-Meca, & López-Pina, 2008), los valores obtenidos se van alejando entre sí tanto en el PHQ-2 (ver Figura 1) como en el GAD-2 (ver Figura 2), llegando a infraestimar el valor real de la confiabilidad de sus puntajes si solo se reporta el coeficiente α (Eisinga et al., 2012; Feldt, & Charter, 2003) para medidas de dos ítems.

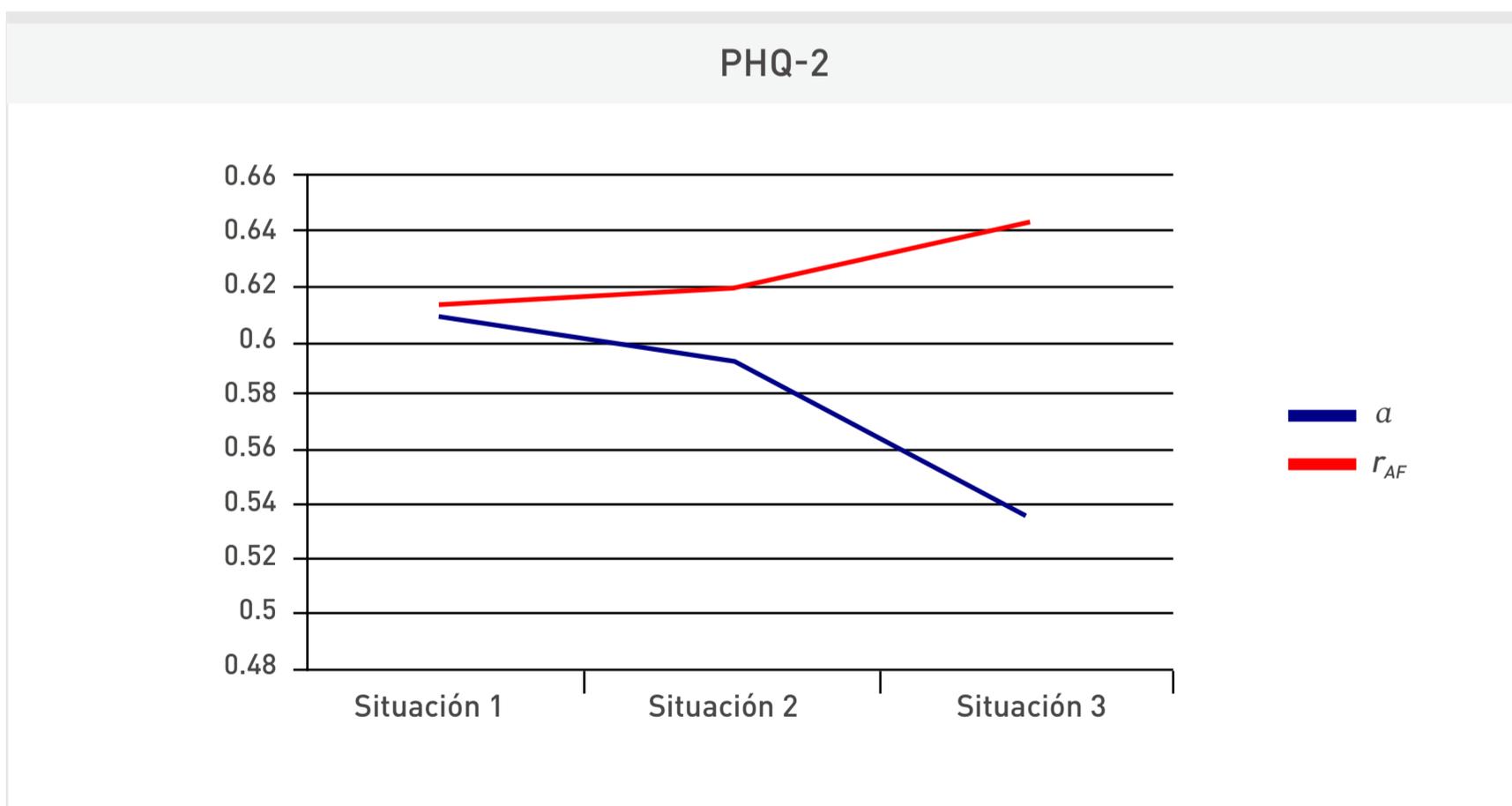


Figura 1. Variación de coeficientes de confiabilidad según situaciones: PHQ-2

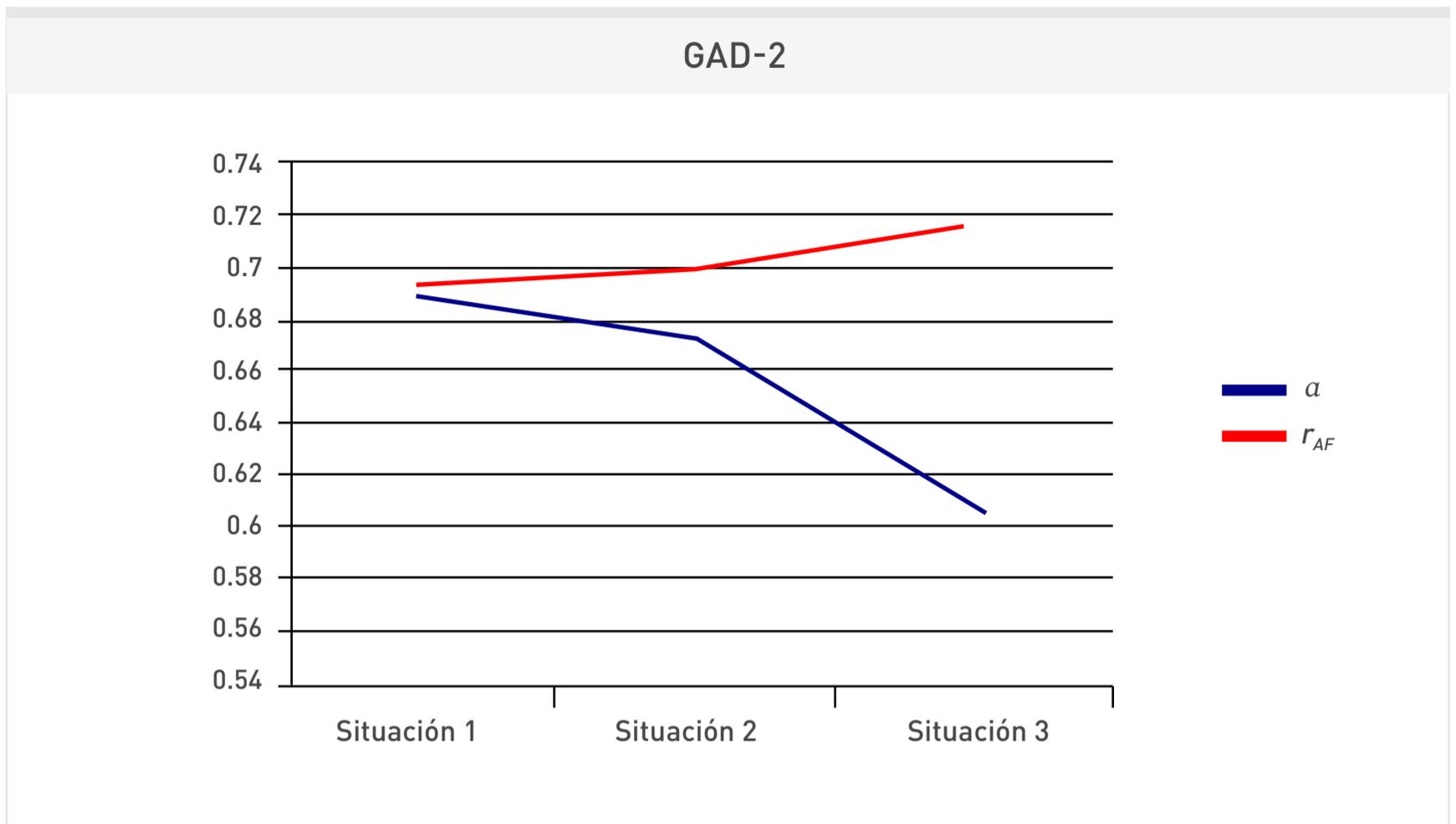


Figura 2. Variación de coeficientes de confiabilidad según situaciones: GAD-2

REFERENCIAS

- 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 questionnaire: 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 Questionnaire: 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. *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 questionnaire-development of a short 18-item version (CERQ-short). *Personality and Individual Differences*, 41(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. *Educational 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.), *Structural 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 Questionnaire-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. *Annals of International Medicine*, 146, 317-325.
- Ledesma, R. (2002). Análisis de consistencia interna mediante Alfa de Cronbach: un programa basado en gráficos dinámicos. *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 it 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. *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