Journal of Work and Organizational Psychology 32 (2016) 17–23

Journal of Work and Organizational Psychology

www.elsevier.es/rpto

Corrections for criterion reliability in validity generalization: The consistency of Hermes, the utility of Midas

Jesús F. Salgado[a,*], Silvia Moscoso[a], Neil Anderson[b]

[a] University of Santiago de Compostela, Spain
[b] Brunel University, U.K.

ARTICLE INFO

Article history:
Received 23 November 2015
Accepted 3 December 2015
Available online 4 February 2016

Keywords:
Interrater
Reliability
Validity generalization
Job performance
Ratings

ABSTRACT

There is criticism in the literature about the use of interrater coefficients to correct for criterion reliability in validity generalization (VG) studies and disputing whether .52 is an accurate and non-dubious estimate of interrater reliability of overall job performance (OJP) ratings. We present a second-order meta-analysis of three independent meta-analytic studies of the interrater reliability of job performance ratings and make a number of comments and reflections on LeBreton et al.'s paper. The results of our meta-analysis indicate that the interrater reliability for a single rater is .52 (k = 66, N = 18,582, SD = .105). Our main conclusions are: (a) the value of .52 is an accurate estimate of the interrater reliability of overall job performance for a single rater; (b) it is not reasonable to conclude that past VG studies that used .52 as the criterion reliability value have a less than secure statistical foundation; (c) based on interrater reliability, test-retest reliability, and coefficient alpha, supervisor ratings are a useful and appropriate measure of job performance and can be confidently used as a criterion; (d) validity correction for criterion unreliability has been unanimously recommended by "classical" psychometricians and I/O psychologists as the proper way to estimate predictor validity, and is still recommended at present; (e) the substantive contribution of VG procedures to inform HRM practices in organizations should not be lost in these technical points of debate.

Corrección por la fiabilidad del criterio en la generalization de la validez: la cohererencia de Hermes, la utilidad de Midas

Palabras clave:
Interjueces
Fiabilidad
Generalización de la validez
Desempeño en el trabajo
Valoraciones

RESUMEN

En la literature se critica el uso de los coeficientes interjueces para corregir por la fiabilidad del criterio en los estudios de generalización de la validez (GV) y cuestionan si .52 es un estimador preciso y no dudoso de la fiabilidad interjueces de las valoraciones del desempeño global en el trabajo. En este articulo, presentamos un meta-análisis de segundo orden de tres estudios meta-analíticos independientes sobre la fiabilidad interjueces de las valoraciones del desempeño en el trabajo y hacemos diversos comentarios y reflexiones sobre el artículo de LeBreton et al. Los resultados de nuestro meta-análisis indican que la fiabilidad interjueces es .52 (k = 66, N = 18.582, SD = .105) para un único supervisor. Nuestras principales conclusiones son: (a) el valor de .52 es un estimador preciso de la fiabilidad interjueces del desempeño global en el trabajo para un único valorador, (b) no es razonable concluir que los estudios de GV que han usado .52 como valor de la fiabilidad del criterio tengan una fundamentación estadística poco segura, (c) sobre la base de la fiabilidad interjueces, la fiabilidad test-retest y el coeficiente alfa, los juicios del supervisor son una medida

[*] Corresponding author. Department of Organizational Psychology. Faculty of Labor Relations. University of Santiago de Compostela. Campus Vida. 15782 Santiago de Compostela, A Coruña, Spain.

E-mail address: jesus.salgado@usc.es (J.F. Salgado).

http://dx.doi.org/10.1016/j.rpto.2015.12.001

útil y adecuada del desempeño en el trabajo y pueden ser usados con confianza como criterio, (d) la corrección de la validez por falta de fiabilidad del criterio ha sido unánimemente recomendada por los psicómetras y psicólogos industriales "clásicos" como el método correcto de estimar la validez del predictor y es todavía recomendada en la actualidad y (e) la contribución sustantiva de los procedimientos de GV para orientar las prácticas de recursos humanos en las organizaciones no debería perderse en estas cuestiones técnicas de debate.

LeBreton, Scherer, and James (2014) have written a challenging lead article in which they make a series of criticisms about the use of interrater coefficients to correct for criterion reliability in validity generalization (VG) studies and disputing whether .52 is an accurate and non-dubious estimate of interrater reliability of overall job performance (OJP) ratings. As researchers who have conducted several meta-analytical (MA) and VG studies in which the value of the interrater reliability was estimated, we here make a number of comments and reflections on LeBreton et al.'s paper. We organize our comments under six points: (1) whether .52 is in fact a dubious interrater reliability value of OJP, (2) their criticism that corrected coefficients were wrongly labelled as uncorrected coefficients, (3) to show that there are some labelling errors in LeBreton et al., (4) if it is appropriate to correct observed validity for criterion reliability, (5) whether interrater reliability is the appropriate coefficient to correct for criterion reliability in VG studies, and (6) wider issues over the value of VG studies for informing policies and practices in organizations.

In combination, we argue that these points indicate unequivocally that the case of LeBreton et al. (2014) is logically flawed, and indeed on closer inspection has been built up piecemeal on a number of outlier interpretations, *non-sequiters* of logical progression, and impractical calls for dataset treatment in VG studies. Following their recommendations risk "throwing the baby out with the bathwater" and reducing the likelihood that VG studies would continue to have important positive benefits for the practice in employee selection and other areas of I/O Psychology.

Is .52 a Dubious Interrater Reliability Value?

LeBreton et al. (2014) doubt whether .52 is a legitimate and accurate estimate of the interrater reliability. To quote, they argue that "the past VG studies which relied on this dubious criterion reliability value have a less than secure statistical foundation", and that they "suspect that researchers would conclude that .52 is not a credible estimate". The problem here is that these are simply opinions without empirical basis, or in fact any supporting rationale being proffered. LeBreton et al. do not provide any empirical support for rejecting .52 as a credible value beyond their suspicion. Should we accept this opinion to unilaterally jettison this well-established and widely used value without any supporting reasoning or empirical foundation? We believe absolutely not, especially when one considers the evidence upon which use of this interrater reliability value has been based.

Viswesvaran, Ones, and Schmidt (1996), for instance, found values of .52 (k = 40, N = 14,650) for interrater reliability, .81 for coefficients of stability (k = 12, N = 1,374) and .86 for coefficient alpha (k = 89, N = 17,899). These coefficients estimate three different sources of measurement error (Schmidt & Hunter, 1996; Viswesvaran, Schmidt, & Ones, 2002). Not all researchers agree that the interrater coefficient is the appropriate estimate of reliability. For instance, Murphy and De Shon (2000) suggested that it is the appropriate coefficient. However, one thing is to believe that another coefficient is the appropriate, as Murphy & De Shon have suggested, and another thing is to dispute that .52 is a credible and non-dubious estimate of interrater reliability, as LeBreton et al., 2014 have suggested. The only way to support this claim is to demonstrate beyond reasonable doubt that Viswesvaran et al. (1996) made errors when they calculated their estimates or, alternatively, to provide another estimate of the interrater correlation based on an independent database. In her large-sample study (N = 9,975) of the interrater reliability of overall performance ratings, Rothstein (1990) found the average interrater was .52. The meta-analysis by Salgado et al. (2003, Table 2) provided another estimate of interrater reliability of overall job performance with a European set of interrater coefficients. They found exactly the same value of .52 (k = 18, N = 1,936). In a third and more recent meta-analysis, Salgado and Tauriz (2014) found that the interrater reliability of overall performance ratings was .52 (k = 8, N = 1,996), using an independent data set. The difference between the estimates of Viswesvaran et al. (1996), Salgado, Anderson, and Tauriz (2015), and Salgado and Tauriz was that the standard deviation was .095, .19, and .05, respectively. That three MAs produced an identical interrater reliability estimate using entirely different samples of primary studies is more than just coincidental – it suggests that this estimate is reasonable and accurate. In a previous meta-analysis, Salgado and Moscoso (1996) estimated the interrater reliability for composite and single supervisory ratings criteria. They found mean interrater reliabilities of .618 and .402, respectively (average r_{yy} = .51). Table 1 reports the results of a second-order meta-analysis of the first three independent studies: Salgado and Moscoso's (1996) meta-analysis was not included because it does not include the sample sizes. As can be seen, the interrater reliability is .52 and the standard deviation combined is .105, which is very close to the figure found by Viswesvaran et al. (1996). In the present case, we used the formula given by McNemar (1962, p. 24) to determine the standard deviation for three distributions combined.

Table 1
Second-order Meta-analysis of the Interrater Reliability of Job Performance Ratings.

N	k	r_{yy}	SD	99% CI
18,582	66	.52	.1056	.518/.522

Note. N = total sample size; k = number of independent coefficients; r_{yy} = weighted-sample average interrater reliability; SD = standard deviation of r_{yy}; 99% CI = 99% confidence interval of interrater reliability.

Murphy and De Shon (2000, p. 896) suggested that the correlation of .52 can be a result of using contexts that encourage disagreement among raters and that encourage substantial rating inflation and, consequently, range restriction. Assuming than one rater uses the entire scale and the other only the top half of the scale, Murphy and De Shon estimated that the correlation among raters corrected for range restriction alone will be .68 and corrected for unreliability, using Viswesvaran et al.'s (1996) coefficient alpha estimate of .86, would be .79. Assuming that one rater uses the entire scale and another only the top third of the scale, their estimated values would be .91 and 1, respectively.

A problematic point in Murphy and De Shon's (2000) examples is that in addition to assuming that the interrater correlation is a