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.

© 2015 Colegio Oficial de Psicólogos de Madrid. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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
1576-5962/© 2015 Colegio Oficial de Psicólogos de Madrid. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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.

© 2015 Colegio Oficial de Psicólogos de Madrid. Publicado por Elsevier España, S.L.U. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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

validity coefficient, they applied the Throndike's formula for Case II (Thorndike, 1949, p. 173) for correcting for range restriction. However, in their examples, the proper formula to correct for range restriction would be the Thorndike's formula for Case I, because the restriction is in the criterion (see the formula in the Appendix). Applying this formula, the correlation corrected for range restriction alone would be .82, and corrected for unreliability in Y_1 and Y_2 would be .95 (using alpha = .86) in the first example of Murphy and De Shon. In the second example, the correlation corrected for range restriction would be .96 and corrected for unreliability would be 1.12 (using alpha = .86), which is an impossible value. Moreover, it should be taken into account that if the ratings are restricted in range Murphy and De Shon should have attenuated the reliability proportionally in order to correct for unreliability. This can be done using the formula developed independently each other by Otis (1922) and Kelley (1921) and reproduced in many books and articles (see the Otis-Kelley formula in the Appendix). The application of this formula would result in an alpha coefficient of .68 in the first case and of -.26 in the second. Repeating the calculations with the attenuated reliability value for the first example, this would be .82 divided by the square root of .86 multiplied by .68 equals to 1.08. In other words, the corrections carried out for the two examples give two impossible values, which cast doubt both on Murphy and De Shon's rationale and the realism of the assumed values of range restriction.

If one accepts Murphy and De Shon's (2000) rationale, then it is surely unrealistic to think that if the context produces range restriction one rater uses the entire scale and the second rater only a fraction of the scale. It would be more realistic to think that the two raters were affected by range restriction, and consequently both would use a fraction of the scale, for example, one using the top ¾ of the scale and the second the top half of the scale. This appears to us a more realistic case. However, this case would need a correction for double range restriction. To this regard, six years after developing the formulas popularized by Thorndike (1949), Pearson (1908) developed the formula for correcting for double range-restriction, which is the formula to be applied in this case (see Formula 3 in the Appendix). Applying this formula, the corrected interrater correlation would be .88.

However, if we accept that the interrater correlation is a reliability estimate (as LeBreton et al., 2014, do), and we apply the Otis-Kelley's formula, this gives values of .79 and .95 as interrater reliability coefficients for the first and second cases of Murphy and De Shon (2000), respectively. In order to estimate the predictive validity of a test, and accepting that the criterion distribution is restricted in range, then the correction should be done on both the criterion reliability and the predictor restricted validity. For example, if the observed correlation between a predictor (e.g., GMA) and overall job performance ratings is .25, and the value of range restriction is U = 1.5, as in the Murphy and De Shon's first example, the full corrected validity would be .77 (rounded), using Case I formula (because the restriction is in the criterion). This requires three steps: (a) to correct the interrater reliability of .52 for range restriction using Otis-Kelley's formula, which results in .79; (b) to correct the validity of .25 by the square root of .79, which gives .28; and (c) to correct .28 for range restriction using a U value of 1.5, which produces a corrected validity of .77. If we use the formula for disattenuation only, with the criterion reliability value of .52, the corrected validity would be .35 (rounded). In other words, the correction for range restriction of the interrater reliability implies the correction for range restriction of the validity coefficient using Case I formula, and the consequence is that a larger validity (and unrealistic) value is obtained. Moreover, it would still be lacking in properly correcting for range restriction in the predictor.

With regard to criterion reliability, sixty-five years ago, Thorndike (1949, pp. 106-107) wrote that "it is not of critical importance that the reliability of a criterion be high as long as it is established as definitely greater than zero. Even when the reliability of a criterion is quite low, given that it is definitely greater than zero, it is still possible to obtain fairly substantial correlations between that criterion and reliable tests and to carry out useful statistical analyses in connection with the prediction of that criterion. Given a test or composite of tests with a reliability of .90 and a criterion with reliability of .40, it is theoretically possible to obtain a correlation of .60 between the two. . . It is more important that the reliability of a criterion measure be known than that it be high."

According to Thorndike (1949) and many classical psychometricians and I/O psychologists (e.g., Ghiselli, Campbell, & Zedeck, 1981; Guilford, 1954; Guion, 1965, 1998; Gulliksen, 1950; Nunnally, 1978, among others), when the criterion measure is unreliable, what it is of critical importance is that the sample size be increased in order to allow for sampling fluctuations and to get stability in the relative size of the validity coefficients.

In summary, while no other more accurate estimate of the interrater reliability is available, researchers can be confident that .52 is currently a robust, accurate, and useful estimate of interrater reliability of overall job performance for a single rater.

Were Corrected Coefficients Labelled as Uncorrected Coefficients?

Le Breton et al. (2014, p. 492) write that "coefficients that have been corrected should be so denoted ($\hat{\rho}$) rather than simply labelled as observed correlation coefficients (r) or referred to as "validities" without clearly articulating the various corrections made to these correlations (cf. Hunter & Hunter, 1984, Table 10; Schmidt & Hunter, 1998, Table 1). Labelling corrected coefficients as uncorrected coefficients could lead some psychologists (or HR managers) to draw improper inferences from meta-analyses." We are not aware that this constitutes an endemic or even frequent problem in VG studies and many published papers can be cited to demonstrate this. In addition, we strongly disagree with the use that LeBreton has made of the Table 1 of Schmidt and Hunter (1998) and Table 10 of Hunter and Hunter (1984). In the footnote of Table 1, Schmidt and Hunter wrote the following (the same text is repeated in Table 2):

All of the validities in this table are for the criterion of overall job performance. Unless otherwise noted, *all validity estimates are corrected for the downward bias due to measurement error in the measure of job performance* (emphasis added) and range restriction on the predictor in incumbent samples relative to applicant populations. The correlations between GMA and other predictors are corrected for range restriction but not for measurement error in either measure (thus they are smaller than fully corrected mean values in the literature). These correlations represent observed score correlations between selection methods in applicant populations.

With regard to Hunter and Hunter's (1984) Table 10, once again, LeBreton et al. (2014) are not fair, and instead adopt an extreme interpretation and position. Hunter and Hunter wrote:

If the predictors are to be compared, the criterion for job performance must be the same for all. This necessity inexorably leads to the choice of supervisor ratings (*with correction for measurement error*) as the criterion because they are prediction studies for supervisory ratings for all predictors (p. 89).

Therefore, Hunter and Hunter (1984) explained that they made corrections for criterion unreliability. Consequently, Hunter and Hunter and Schmidt and Hunter (1998) properly labelled the corrected coefficients and they have not lead psychologists (or HR managers) to draw improper inferences from meta-analyses. In other words, if a reader draws improper inferences is because he/she has not properly read the footnote and the explanations. The responsibility is that of the reader not of the writers and it is, ironically, LeBreton et al. (2014) who may have misguided readers in the

present article by not mentioning Schmidt and Hunter's footnote and by omitting to have noted Hunter and Hunter's unambiguous explanation of the corrections done.

We agree that improper inferences could have serious consequences for I/O Psychology not only in terms of future theory construction but also in terms of practical policy decisions affecting I/O psychologists working in organizations. However, we also suggest that improper attributions could have the same serious consequences for the credibility and respectability of the I/O psychology as a science.

Labelling Errors in LeBreton et al. (2014)

LeBreton et al (2014, p. 4) write that corrections for range restriction were based on U values, which they define as the ratio of the restricted standard deviation to unrestricted standard deviation. The first part of the sentence is true. The second part is unfortunately wrong. VG studies have indeed used U values to correct for range restriction. However, U values are the ratio of the unrestricted standard deviation to the restricted standard deviation. Second, it is incorrect to assert that U values of 0.66, 0.81, 0.65, and 1.00 were used in the meta-analyses they mention. These last values are estimates of u values, which are just the inverse of U values. If the meta-analyses mentioned by LeBreton et al. used those values as U values in the formula for range restriction correction, the corrected correlations would be smaller than the observed correlations. In fact, the meta-analyses they mention used U values of 1.51, 1.23, 1.54, and 1, respectively. LeBreton et al. have confused the coefficient of heterogeneity (U values) with the coefficient of homogeneity (u values). We suspect that this is simply a slip, but an important one potentially calling for a correction note to be published.

Is it Appropriate to Correct for Criterion Unreliability only?

LeBreton et al. (2014) consider that to correct for criterion reliability alone is not appropriate and that labelling this coefficient as "operational validity" is erroneous. There are two points in this assertion. The first one is about the appropriateness of the correction for measurement error in the criterion. The second point is about the appropriateness of labelling this coefficient as "operational validity". With regard to the first point, let see what relevant psychometricians and I/O psychologists have written. For example, Guilford (1954), in his famous book titled *Psychometric Methods*, stated the following:

In predicting criterion measures from test scores, one should not make a complete correction for attenuation. Correction should be made in the criterion only. On the one hand it is not a fallible criterion that we should aim to predict, including all of its errors; it is a "true" criterion or the true component of the obtained criterion.

On the other hand, we should not correct for errors in the test, because it is the fallible scores from which we must make predictions. We never know the true scores from which to predict. We should obtain a very erroneous idea of how well we are doing with a selection test or composite score if the reliability of criterion measure were only .30, which can very well happen, and if no correction for attenuation were made. . . This type of correction should be used much more than it is, even though it requires an estimate of criterion reliability. (Guilford, 1954, p. 401)

Another frequently cited and respected psychometrician, Nunally (1978, p. 238), shared this view when he wrote:

Another important use of the correction for attenuation is in applied settings where a test is used to forecast a criterion. If, as it often happens, the criterion is not highly reliable, correcting for unreliability of the criterion will estimate the real validity of the test. Here, however, it would be wrong to make the double correction for attenuation, since the issue is how well a test actually works rather than how well it would work if it were perfectly reliable. In predictions problems, the reliability of the predictor instrument places a limit on its ability to forecast a criterion, and the corrections for attenuation cannot make a test more predictive than it actually is. The only use for this double correction would be in estimating the limit of predictive validity of the test as both test and criterion are made more and more reliable.

Three well-known I/O psychologists and past presidents of SIOP, Ghiselli et al. (1981, p. 291), in their book on *Measurement Theory for the Behavioral Sciences* wrote that:

When considering criterion-related validity, it is generally not appropriate to correct for unreliability in the predictor. In applied psychology we must always deal with fallible predictor scores. That is, the decision must be made on the basis of the predictor data in hand it is the validity of the fallible predictor information that is at issue.

More recently, two additional past-presidents of SIOP, Schmitt and Klimoski (1991, p. 99) appear to have a similar view when they wrote that:

A frequent application of this correlation in personnel selection research involves the correction for attenuation for unreliability in the criterion (a job performance measure which we are trying to predict). Corrections to the observed correlation for lack of criterion reliability are made to estimate the true validity of a potential predictor. Similar corrections for predictor unreliability are not made because the use of the predictor in other situations will always involve a similar lack of reliability.

In order to not to tire out the readers, we finally mention what Guion and Highouse, (2006, p. 163) who argue that "coefficients should therefore be corrected only for criterion unreliability."

We could mention many other authoritative writers supporting the same view. While we would be the first to concede that it is possible that all of these eminent scholars were wrong and LeBreton et al. (2014) are right, this is unlikely in our view. Rather corrections for criterion unreliability comply with the recommended practice in personnel selection for over 65 years. Again, to overturn this convention requires both compelling logic and empirical evidence - LeBreton et al. have provided neither.

With regard to the labelling of the corrected coefficient, this is more a question of preference and consensus among researchers than of substantance. In the typical use of this term in VG studies, it refers to the capacity of a procedure for predicting a specific criterion, assuming that the criterion measure is free of error. In our view, after an excess of 30 years of using this term in I/O Psychology in hundreds of published papers, books, conference presentations, and doctoral dissertations, the use and meaning of "operational validity" is well established and does not lead to confusion, especially when the term is explained in the paper.

Is Interrater Reliability the Appropriate Coefficient to Correct for Criterion Reliability?

Quite frequent debates have occurred in I/O Psychology over what is the appropriate coefficient to correct for attenuation due to measurement error (see, for instance, the compilation of articles by Ronan & Prien, 1971). The classical view (and many times forgotten) on this issue is to use the type of reliability estimate that treats as error those things one decides should be treated as errors (Ghiselli et al., 1981; Guilford, 1954; Schneider & Schmitt, 1986). For example, Ghiselli et al. (1981, p. 290) suggested that a test-retest coefficient would not be appropriate if true scores change over time because this coefficient would count such changes as error and a similar reasoning would apply if an internal consistency coefficient (e.g., Cronbach's alpha) were used to estimate the reliability of a

multidimensional measure. At present, there is general agreement among researchers that job performance is a multi-dimensional construct and that job performance can change over time. Therefore, test-retest and internal consistency coefficients would not be optimal estimates of job performance reliability. Cronbach (1990, p. 587), in addition to accept that interrater correlation is a reliability coefficient, affirmed that "averaging across judges allows many of their errors to balance out. . . the bias of one judge tends to cancel the bias of another, and each adds information the other had no opportunity to observe" and point out that the Spearman-Brown's formula is the way to obtain the reliability estimate of the average rating. This concurs with Schmidt and Hunter's (1996) suggestion that interrater reliability should be used to correct for attenuation when ratings are used as criterion.

According to Guion (1965, p. 45):

With ratings, conspect reliability is extremely important. The 'score' on a rating form depends not only on the behavior of the person observed or rated, it also depends upon the traits and behaviors of the rater. These characteristics may produce distortions or error; to the extent that they differ from one rater to another, they are variable or random errors in a set of ratings obtained with multiple judges. The reliability of ratings may be defined as the degree to which the ratings are free from error variance arising from the behavior of either the ratee of the rater.

Schmitt and Klimoski (1991, p. 99; see also Schneider & Schmitt, 1986, p. 211) pointed out that, in personnel selection, "if we are using ratings of job performance as criteria, then rate-rerate estimates with an appropriate time interval and by different raters would be the appropriate estimate of reliability." In other words, Schmitt and Klimoski suggest that the Coefficient of Equivalence and Stability (CES) is the most appropriate reliability coefficient. Schmidt, Le, and Remus (2003) agree with this view, and developed a formula for estimating CES. However, it is very difficult in practice to achieve the required data for estimating CES. For this reason, the interrater correlation is typically used. Additionally, CES is always a smaller reliability estimate than interrater reliability. Very recently, Salgado (2015) presented the estimates of CES for a time interval of 1, 2, and 3 years. The figures obtained for a single rater were .51, .48, and .44, respectively, and for two raters were .59, .55, and .51. He found that the interrater coefficient overestimated the reliability of job performance ratings, as it does not control for transient error. Consequently, the magnitude of the operational validity has been slightly underestimated in validity generalization studies.

Since the mid-1970s, the majority of the published VG studies have used a random effect method and many of them have used the software developed by Schmidt and his colleagues. It is in this context where we should debate if the interrater coefficient is the appropriate reliability to correct for criterion measurement error. Synthetically, Hunter, Schmidt, and their associates posited that the interrater reliability is the coefficient of interest because it corrects most of the unsystematic errors in supervisor ratings (see, Schmidt & Hunter, 1996; Viswesvaran et al., 1996), although other researchers (e.g., Murphy & De Shon, 2000) disagree with this point. To this regard, Sackett (2003) has suggested that the interrater reliability is the coefficient of interest when a meta-analysis of random effects is carried out. Therefore, the appropriate coefficient was used in VG studies under the Hunter and Schmidt's meta-analysis model (considering that the CES was not available for the researchers).

Wider Issues over the Value of VG Studies for Informing Organizational Policies and Practices

Taken together, the cumulative effects of LeBreton et al.'s (2014) rather technical criticisms, outlier interpretations, and penchant for iconoclastic dismissal of generally accepted analytical norms, run the risk of undermining the vitally important and far-reaching benefits that MA and VG evidence undoubtedly have for personnel practices in organizations (Anderson, 2005). Indeed, our final point is that these benefits have been substantial and the work of leading I/O psychologists in this field should be properly acknowledged, a point given scant regard in the lead article. The previous points we make in this rejoinder centre upon the consistency of VG as a messenger statistical technique to convey important empirical findings (Hermes, as it were); our last point concerns the utility value of MA and VG procedures to inform evidence-based practices in organizations (Midas, as it were). We should keep in mind this wider perspective, we would argue, since over the years the findings from VG studies have made huge contributions to the integrating science and practice in personnel selection specifically, but also in several other areas in I/O Psychology and management generally.

It should not go unaccredited that I/O psychologists were at the forefront of these contributions (e.g. Jack Hunter, Frank Schmidt, among others), and that their procedures have been widely adopted in other areas of the social and health sciences. It is of course desirable that we actively debate various technical issues of VG procedures and interpretive norms within our field; however, it would indeed be regrettable if such debates spilled-over to undermine the notable value that these contributions have made taking a more general and balanced overview. The LeBreton et al.'s (2014) criticisms, whilst well-intentioned, are ill-founded as our previous points have indicated. Given the huge contributions that MA and VG procedures have made, and VG proponents in particular, it would be unwise for our field to allow these criticisms to run wild and to detract from these wider and most valuable contributions. We think this is unlikely to be the case, especially given the flaws in their arguments, but would highlight this final point out of concern for balance and the need to keep such debate points in a more general and constructive perspective.

Conclusion

As a summary, the following are our main conclusions. First, the value of .52 is an accurate estimate of the interrater reliability of overall job performance for a single rater. At present, no meta-analyses of interrater reliability have demonstrated that this value is erroneous, dubious, or inaccurate, and neither does the case put forward in the lead article. Second, 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. On the contrary, they are solid. The interpretations of the corrected correlations based on this estimate are appropriate. Third, based on interrater reliability, test-retest reliability, and coefficient alpha, supervisor ratings are a useful and appropriate measure of job performance and they can be confidently used as a criterion. Fourth, 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 it is still recommended at present (e.g., Guion & Highhouse, 2006; Sackett, Putka, & McCloy, 2012; Schmitt & Klimoski, 1991; Schneider & Schmitt, 1986). The case advocated in LeBreton et al.'s (2014) article fails to override this generally-accepted convention. Fifth, and finally, the substantive contribution of VG procedures to inform HRM practices in organizations should not be lost in these technical points of debate.

We have also three suggestions for research into validity procedures and organizational practices in personnel selection. First, we agree with LeBreton et al. (2014) that more reliable measures should be used and that more reliable are preferred to less reliable measures. However, this does not mean that supervisory

ratings should be dismissed as appropriate criteria in validation research. Quite the opposite. For example, if the criterion measure is based on two raters, the interrater reliability of the summed scores would be approximately .68; if the raters were three, the reliability would be .76 (this last value is very common in many personality questionnaires, for instance). Therefore, it would be enough to change the standards marginally and to require that for research purposes the criterion scores should be based on two or three raters. Recently, Salgado, Bastida, Vázquez, and Moscoso (manuscript under review) have found that the inter-rater reliability for the sum of two independent raters was .77, .76, .73, and, .71, for four consecutive years and that the test-retest coefficients ranged from .62 to .83, with and average of .71 across the four years. In other words, both the interrater reliability and the coefficients of stability were very close. However, this requirement of additional raters can create other practical problems as it is difficult to find companies that can have two or more raters for the same employee and that they can provide ratings for several consecutive years of the same employee made by the same supervisor.

Our second suggestion is to distinguish between the reliability requirements of ratings when they are used as a criterion and when they are used as predictors or for making personnel decisions (e.g., promotions, tenure, and compensation). Used as predictors and to make personnel decisions, the procedure should be as higher reliable as possible. Over 65 years ago, Ghiselli and Brown (1948) suggested that a level of .85 would be necessary in order to make personnel decisions acceptable. This rule-of-thumb is both appropriate for tests and ratings. However, in personnel selection, the decision is based on the predictor score (e.g., a test) not on the criterion score. For this reason, reliability requirements of ratings when they are used as a criterion in validity studies are less problematic than if the ratings are used to make promotion decisions, for instance, in which case, those ratings would be the predictor and the measure of performance on the new job (e.g., new ratings) would be the criterion. In this second case, the reliability requirements would be similar to the other selection procedures. For this reason, promotion decisions based on ratings should not be based on a single rater. However, if the decision is based on the ratings of four or five raters, the reliability would be between .81 and 85. A different issue is whether organizations have four or five independent raters for making the decision. The most probably answer is no and, consequently, the recommendation is that organizations should not base their personnel decisions on ratings alone (Rothstein, 1990).

Third, and finally, our suggestion is to retain the advice and recommendations made over many years by leading scholars. It is incumbent upon any researcher to argue their point for overturning these conventions beyond reasonable doubt and for their replacement procedures to be robust, practical, and generally applicable. VG procedures used in published papers based upon these conventions are the bedrock of our scientific findings and these remain robust if we simply "follow the advice of the classics".

Conflict of Interest

The authors of this article declare no conflict of interest.

Financial Support

The research reported was partially supported by Grant PSI2014-56615-P from the Ministry of Economy and Competitiveness to Jesús F. Salgado and Silvia Moscoso and by a Leverhulme Trust grant number IN-2012-095 to Neil Anderson.

Appendix.

Formulas used

1) Formula to attenuate the reliability coefficient due to range restriction (Formula of Otis-Kelley):
$r_{yy} = 1 - \left[U^2\left(1 - R_{yy}\right)\right]$; where $U = SD/sd$ and R_{yy} = unrestricted reliability coefficient.
2) Formula to correct validity for range restriction (Thorndike's Case I; Pearson, 1902; Thondike,1949):
$R_{12} = \sqrt{1 - u_2^2\left(1 - r_{12}^2\right)}$; where $u = sd/SD$; r_{12} = observed validity; R_{12} = corrected validity.
3) Formula to correct validity for double range restriction (Pearson, 1908):
$R_{12} = \frac{r_{12}u_1u_2}{\sqrt{1-r_{12}^2\left(1-u_1^2\right)}\sqrt{1-r_{12}^2\left(1-u_2^2\right)}}$; where $u_1 = sd/SD$ in variable 1, and u_2 = sd/SD in variable 2; r_{12} = observed validity, and R_{12} = validity corrected for double range restriction.

References

Anderson, N. R. (2005). Relationships between practice and research in personnel selection: Does the left hand know what the right is doing? In A. Evers, N. Anderson, & O. Smit-Voskuyl (Eds.), *Handbook of personnel selection* (pp. 1–24). Oxford, UK: Blackwell.
Cronbach, L. J. (1990). *Essentials of psychological testing* (Fifth edition). New York, NY: Harper & Row.
Ghiselli, E., & Brown, C. (1948). *Personnel and industrial psychology*. New York, NY: McGraw-Hill.
Ghiselli, E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: Freeman.
Guilford, J. P. (1954). *Psychometric methods*. New York, NY: McGraw-Hill.
Guion, R. M. (1965). *Personnel testing*. New York, NY: McGraw-Hill.
Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NJ: Lawrence Erlbaum Associates.
Guion, R. M., & Highouse, S. (2006). *Essentials of personnel assessment and selection*. Mahwah, NJ: Erlbaum.
Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin*, *96*, 72–98.
Kelley, T. L. (1921). The reliability of test scores. *Journal of Educational Research*, *3*, 370–379.
LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Correction for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology*, *7*, 478–500.
McNemar, Q. (1962). *Psychological statistics* (3rd edition). New York, NY: Wiley.
Murphy, K. R., & De Shon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873–900.
Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.
Otis, A. S. (1922). A method for inferring the change in a coefficient of correlation resulting from a change in the heterogeneity of the group. *Journal of Educational Psychology*, *13*, 293–294.
Pearson, K. (1908). On the influence of double selection on the variation and correlation of two characters. *Biometrika*, *6*, 111–112.
Ronan, W.W., & Prien, E. (Eds.) (1971). *Perspectives of the measurement of human performance*. New York, NY: Appleton-Century Crofts.
Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, *75*, 322–327.
Sackett, P. R. (2003). The status of validity generalization research: Key issues in drawing inferences from cumulative research studies. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 91–114). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Sackett, P. R., Putka, D. J., & McCloy, R. A. (2012). The concept of validity and the process of validation. In Neal Schmitt (Ed.), *The Oxford handbook of personnel assessment and selection*. Oxford, UK: Oxford University Press.
Salgado, J. F. (2015). Estimating coefficients of equivalence and stability for job performance ratings: the importance of controlling for transient error on criterion measurement. *International Journal of Selection and Assessment*, *23*, 37–44.
Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., De Fruyt, F., & Rolland, J. P. (2003). A meta-analytic study of general mental ability validity for different occupations in the European Community. *Journal of Applied Psychology*, *88*, 1068–1081.
Salgado, J. F., Anderson, N., & Tauriz, G. (2015). The validity of ipsative and quasi-ipsative forced-choice personality inventories for different occupational groups: A comprehensive meta-analysis. *Journal of Occupational and Organizational Psychology*, *88*, 797–834.
Salgado, J. F., Bastida, M., Vázquez, S., & Moscoso, S. (manuscript under review). *Happiness, positive emotions and job performance: a four-year longitudinal study*.

Salgado, J. F., & Moscoso, S. (1996). Meta-analysis of interrater reliability of job performance ratings in validity studies of personnel selection. *Perceptual and Motor Skills*, *83*, 1195–1201.
Salgado, J. F., & Tauriz, G. (2014). The five-factor model, forced-choice personality inventories and performance: A comprehensive meta-analysis of academic and occupational validity studies. *European Journal of Work and Organizational Psychology*, *23*, 3–30.
Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, *1*, 199–223.
Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274.
Schmidt, F. L., Le, H., & Remus, I. (2003). Beyond Alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, *8*, 206–224.
Schmitt, N., & Klimoski, R. (1991). *Research methods in human resources management*. Cincinnati, OH: South-Western Publishing Co.
Schneider, B., & Schmitt, N. (1986). *Staffing organizations*. Glenview, IL: Scott, Foresman.
Thorndike, R. L. (1949). *Personnel selection. Test and measurement techniques*. New York, NY: Wiley.
Viswesvaran, C., Ones, D., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557–574.
Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisor and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, *87*, 345–354.