Industrial and Organizational Psychology
PERSPECTIVES ON SCIENCE AND PRACTICE

Industrial and Organizational Psychology, **7** (2014), 501–553.
Copyright © 2014 Society for Industrial and Organizational Psychology. 1754-9426/14

COMMENTARIES

When and Why Correcting Validity Coefficients for Interrater Reliability Makes Sense

PAUL R. SACKETT
University of Minnesota

The purpose of this short note is to attempt a clarification on a number of counts. Over the last decade, as an author, editor, and reviewer I have observed considerable continued disagreement about the use of corrections based on interrater correlations (e.g., reviewers instructing authors not to correct predictor–criterion correlations, citing Murphy and DeShon (2000); and other reviewers demanding that authors correct predictor–criterion correlations, citing Viswesvaran, Ones, and Schmidt (1996). In this note I hope to shed light on what correcting a predictor–criterion correlation based on an interrater correlation does and does not do. I would like to outline several points that I believe are useful in understanding when and why correction for attenuation makes sense.

Differentiating Level of Reliability Needed for Individual Decisions From Level Needed for Research Studies

LeBreton, Scherer, and James (2014) express amazement that anyone would use a measure with reliability as low as .50, offering quotations attesting to the need for highly reliable measures. What is critical to note is that both of the opening quotes offered deal with measures to be used operationally for decisions about individuals (e.g., "If important decisions are being made with respect to specific test scores" (Nunnally & Bernstein, 1994); "If a procedure is to be used to compare one individual with another" (Cascio & Aguinis, 2005). In such settings, high reliability is indeed very important. However, the focus of the LeBreton et al. article is not on measures for operational use but on measures used as criteria in research studies. Here I will argue that the use of measures with modest reliability can be useful and appropriate in research studies, as long as reliability is accurately estimated.

Let us use a concrete example: Assume an organization wants to examine whether a customer service inventory is indeed predictive of the quality of customer service provided. As a criterion measure of customer service quality, they propose to employ "secret shoppers": Trained individuals who pose as customers, have a service encounter with an employee, and rate the quality of service provided. McLellan, Schmidt, Amundson, and Blake (1998) report that ratings from seven secret shoppers are needed for a composite rating to reach the .50 level of reliability. From that value, I apply the Spearman-Brown formula and estimate that (a) the reliability

Correspondence concerning this article should be addressed to Paul R. Sackett.
E-mail: psackett@umn.edu
Address: Department of Psychology, University of Minnesota, Minneapolis, MN 55455

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press

of a single shopper is .14, and (b) it would take a composite of 56 shoppers' ratings to reach a reliability of .90.

For a moment let's assume we know something that is in reality not known, namely, the correlation of the customer service inventory score with a composite of ratings from an infinite number of secret shoppers, and that value is $r = .40$. From this we can apply the correction for attenuation in reverse to the seven-shopper setting (e.g., .40 times the square root of .5) and estimate that we could expect to obtain a value of .28. Similarly, in the 56-shopper setting with reliability .90, we would expect to obtain a validity estimate of .38. This is all straightforward: We see that we obtain a higher validity estimate with the highly reliability criterion than with the low reliability criterion.

Now here's the big question: Does the operational use of the customer service inventory to select employees really provide more value to the organization in terms of the level of customer service provided when it produces a correlation of .38 with the 56-shopper criterion than when it produces a correlation of .28 with the seven-shopper criterion? Of course not: The value of the customer service provided to customers is the same regardless of whether the measured indicator of service is highly reliable or much more modestly reliable. This is my central point: We generally are not *per se* interested in a measured fallible indicator of performance; we want to know how well we predict the underlying construct reflected by that fallible measure.

Armed with this knowledge about the reliability of secret shopper ratings, how would we advise an employer planning a validity study? A first instinct is to seek as reliable a criterion measure as possible. So we might encourage the organization to go for the criterion with reliability of .90. But there are other issues in play, such as costs. Hiring secret shoppers is expensive, and the organization may conclude having a large sample of employees each evaluated by 56 secret shoppers is cost prohibitive. Might we, they ask, get by with the much more affordable option of seven shoppers per employee? In fact, if we use the correction for attenuation we reach the same conclusion with 7 as with 56: An observed validity of .28 corrected for a criterion reliability of .50 gives an estimate of .4; so does an observed validity of .38 corrected for a criterion reliability of .90. The corrected validity of .40 is the best estimate of the operational value of the customer service inventory for predicting service as perceived by secret shoppers. Conceptually, the unattainable ideal of a rating by an infinite number of secret shoppers reflects the actual degree of customer service as perceived by secret shoppers. Thus, rather than it being something we are not interested in (because we never have an infinite number of raters), validity against this perfectly reliable criterion is, in fact, exactly what we wish to estimate.

A number of caveats: First, one gets nothing for free. Correlations corrected for attenuation have more sampling error than uncorrected correlations. So obtaining a highly reliable criterion measure and not using a correction produces a validity estimate with a smaller standard error than one obtained via a less reliable criterion and a correction for attenuation. It is a value judgment as to whether one is willing to accept the larger standard error that accompanies a less reliable criterion in return for the dramatically lower cost of obtaining the criterion measure.

Second, note that this whole discussion is based on the premise that customer service as rated by external observers is indeed what we want to measure. This is the basis for some confusion when the term "true validity" is used to refer to correlations corrected for attenuation. "True" in this context means "free from sources of variance not shared across raters;" it does not necessarily mean that the criterion is the "true" or proper one. For example, someone may argue that customer service as perceived directly by the customer is different from service as perceived by a secret shopper. They may argue for customer

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press

ratings rather than secret shopper ratings as the criterion of interest. (It is, of course, an empirical question as to whether the two rater perspectives actually differ.) But one might then follow a different strategy in creating a criterion measure, such as that used by Hardison, Sackett, and Mitchell (2002), who examined a setting in which follow-up phone calls were made to customers to obtain a rating following a service encounter. The issues parallel those above for secret shoppers. Hardison et al. report reliability of .03 for a single customer rating, with 35 customer ratings needed for a reliability of .50 and 309 ratings needed for a reliability of .90. As above, there is a cost/reliability tradeoff, and we can envision a conclusion that obtaining 309 customer ratings per employee is cost prohibitive, resulting in the choice the 35-rating criterion despite the modest reliability of .50. Correcting for attenuation estimates the operational value of the inventory for predicting customer ratings. Again, that is "true" validity is the sense of being free from variance due to rater differences rather than it being the "true" and proper criterion construct. My sense is that there is recognition of the confusion caused by the term "true" validity and that this usage is disappearing from the literature.

Third, it is important to be clear as to when a correction of attenuation is not conceptually appropriate. The key question is whether the measured criterion is viewed as a fallible indicator of a construct of interest or whether the indicator *per se* is of interest. To continue with our customer service example, imagine a setting in which ratings by seven secret shoppers are obtained for new employees at the end of a 3-month probationary period, with an organizational policy in place that employees with a mean rating below four on a seven-point scale are not retained. One might want to investigate how well a customer service inventory predicts whether new employees will reach this retention threshold. Here the measured performance indicator, fallible as it is, is what the organization cares about, and correction for attenuation is not appropriate. Thus one must be very conscious of whether the goal is to predict the specific criterion measure or to predict an underlying construct. I believe that it is more frequently the latter.

The above discussion has focused on the use of a measure with low reliability as a criterion in research studies. I briefly return to LeBreton's concern about using a measure with low reliability as the basis for individual decisions. First, the *Standards for Educational and Psychological Testing* (1999) offer a perspective on expectations for reliable and valid measurement:

> It is useful to distinguish between devices that lay claim to the concepts of and techniques of the field of educational and psychological testing from those which represent nonstandardized or less standardized aids to day-to-day evaluative decisions. Although the principles and concepts underlying the *Standards* can be fruitfully applied to day-to-day decisions, such as when a business owner interviews a job applicant, a manager evaluates the performance of subordinates, or a coach evaluates a prospective athlete, it would be overreaching to expect that the standards of the testing field be followed by those making such decisions (pp. 3–4).

Second, we do often bring our knowledge to the table in developing systems for making consequential decisions about employees (e.g., pay increases, promotions), with a wide variety of mechanisms brought into play (e.g., supervisors justifying compensation recommendations to a panel of managers, the systematic input of multiple rater perspectives when making promotion decisions). Thus it is not the case that the value of the reliability of a single supervisor's rating can be routinely equated with the reliability of decisions about individuals in organizations.

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press

Responding to Concerns That the Correction for Attenuation Is Inappropriate Because Ratings Cannot Be Simply Decomposed Into True and Error Scores

There is longstanding suspicion of using an interrater reliability coefficient as the basis for correction for attenuation based on concerns about the appropriateness of the classical test theory model of observed scores being decomposed into a true score component and a random error component. It is doubtless the case that the portion of rating variance that is not shared among raters can reflect systematic rather than random sources of error (e.g., systematic bias due to ratee attractiveness). Here I offer a take on this that bypasses classical test theory and, I hope, defuses the arguments

Although some may be skeptical of ratings as a criterion, many applied settings do use rating criteria, and what follows below is based on the notion that one is willing to use rating criteria. If one is willing to accept the rating by one supervisor as a criterion, then one logically should answer "yes" to the question "If a rating from a second comparable supervisor were available, would the use of a composite of the two supervisors' ratings be preferable to relying one only one supervisor?" (Note that "comparable supervisor" here assumes that there is a population of individuals who might serve as first line supervisors, and the raters to be combined into a composite are a random draw from that population. Thus, although any two supervisors might, in fact, rate very similarly or rate very differently, the expected value of the correlation between any two supervisors drawn from the population of potential raters is the same.)

If a composite if two raters is viewed as preferable to a single rater, then, by extension, the same logic would apply to a third, fourth, and fifth supervisor, up to an infinite number of supervisors. Logically, one must agree that the ideal situation would be to have an infinite number of supervisors even though this is, of course, never attainable. But one can ask the question: "What would the estimate of validity be if a composite of an infinite number of raters were used as the criterion?" or, equivalently, "What is the validity we'd expect if we could remove variance not shared across raters from the criterion?" Applying such a correction makes sense if one concludes that these are interesting questions. I suggest that they are and that in most cases are of more interest than the correlation between a predictor and a single rating of performance by a specific supervisor. We do not, for example, routinely require revalidation of a selection procedure after supervisory turnover.

The question can be answered using the theory of composites (Ghiselli, Campbell, & Zedeck, 1981). The theory of composites involves nothing more than some algebraic manipulation of the computational formula for the Pearson product–moment correlation (i.e., for the correlation between a predictor x and a criterion consisting of a composite of z raters one replaces "y" in the formula for r_{xy} with "$y_1 + y_2 + \ldots\ y_z$"). Note that the theory of composites does not make use of classical test theory and requires none of the classical test theory assumptions.

After substituting the composite for y in the formula for r_{xy}, it emerges that the correlation between a composite (i.e., multiple ratings) and another variable (i.e., the predictor) relies on three factors:

> A: the mean r between the predictor and the variables going into the composite. Each variable going into the composite is a rating by one supervisor, and the correlation between the predictor and performance as rated by one supervisor has the same expected value for each supervisor, as supervisors are assumed to be drawn from the same population. So the observed r between the predictor and the rating by one supervisor serves as the needed value here.
>
> B: the mean r between the variables going into the composite (in this case, the mean r between the rating by one supervisor and another). The mean

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press

interrater *r* from the literature serves as the needed value here.
C: the number of variables going into the composite (in this case, the number of supervisors providing ratings.

Using the above A, B, C notation, the correlation between an outside variable and a composite is:

$$A/(SQRT((1/C) + ((C-1)/C) * B))$$

(see formula 7-13 in Ghiselli et al., 1981).

Working with this formula reveals that as the number of raters contributing to the composite increases, the contribution of variance shared across raters to the total variance in the composite increases; at the same time the contribution of unshared variance decreases, and goes to zero as the number of raters goes to infinity. Thus *r* between a predictor and a composite of an infinite number of comparable raters is also *r* between the predictor and shared variance. When C is infinity, 1/C goes to zero, and (C-1)/C goes to one, and thus, as Ghiselli, Campbell, and Zedeck note, the formula reduces to:

$$A/SQRT\ B$$

(see formula 7-14 in Ghiselli et al., 1981).

This is the traditional correction for attenuation: Observed validity divided by the square root of the interrater *r*. The attenuation formula can thus be used to estimate *r* between the predictor and variance shared across raters, without any assumptions about the nature of either the shared variance or the unshared variance.

Responding to the Notion a Finding That Four Individuals Difference Predictors Explain 88% of Variance Is Implausible, and That Therefore the Reliability Value Used in the Correction for Attention Is Flawed

LeBreton et al. assemble a meta-analytic correlation matrix relating cognitive ability, integrity, structured interviews, and work sample tests to overall job performance, correcting all correlations for measurement error and range restriction. They then regress performance on the four predictors, producing an R^2 value of .88. They argue that this is implausible as it leaves very little room for the many known additional correlates of performance. They then argue that because the interrater reliability value of .52 is used in the generation of this R^2 value, it is the use of this value in the correction for attention that is the culprit.

My first observation is that the input to this meta-analysis is (a) 10 meta-analytic mean correlation estimates, (b) meta-analytic mean reliability estimates for all five variables, and (c) meta-analytic estimates of the degree of the mean degree of range restriction on each predictor. There is the potential for error in each of these values. For example, a recent meta-analysis of the integrity test literature by Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012) produces a markedly lower validity estimate than the value used by LeBreton; substituting that value into the matrix reduces the R^2 value from .88 to .74. Although I expect further developments regarding integrity test validity, the point is that there are a lot of moving parts in the estimation process.

Second, it is important to be precise in discussing "percent of variance accounted for." In research in the context of personnel selection, the prototypic study examines individuals within a single job within a single organization (e.g., what is the predictive power of this structured interview we've designed for selecting marketing analysts within our firm?). Other research may focus on individuals across jobs within a firm (e.g., a stress research may administer a burnout survey to all employees within an organization) and correlate burnout with performance. Yet other research may focus on factors that are often essentially constant within a jobs (e.g., task characteristics, such as skill variety and task identity) but that vary across jobs and thus are studied in samples made up of varied jobs. The emergence of multilevel perspectives in organizational

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press

research has helped highlight the importance of specifying just what variance one is explaining (e.g., individual variance within job in a single organization, individual level variance within job across organizations, individual variance across jobs, job-level variance, unit level variance). The key idea here is that if one variable or set of variables account for X% of individual variance within job and within organization and another set of variables account for Y% of variance across jobs and across organizations, the value of X puts no constraints on the value of Y (and vice versa). Thus the proposition, that a finding that a small set of individual difference variables accounts for a large proportion of individual variance within jobs is an implausible result because it leaves little room for other variables to operate, should be reconsidered in the context of what level of variance is at issue for the various individual and organizational variables under consideration.

In conclusion, I have offered (a) an alternate perspective on the argument that high reliability is needed for all measures at all times, (b) an argument for using interrater reliability as the basis for corrections that does not rely on classical test theory assumptions, and (c) some cautionary observations regarding the need to be clear as to what sources of variance are in play when one compares the % of variance explained by one set of variables versus another.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: Author.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences*. San Francisco, CA: W.H. Freeman and Co.

Hardison, C. M., Sackett, P. R., & Mitchell, T. W. (2002, April). *Customer ratings of performance: A useful criterion for validation research?* Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

LeBreton, J. M., Scherer, K. T., & James, L. R. (2014). Corrections for criterion reliability in validity generalization: A false prophet in a land of suspended judgment. *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *7*(4), 478–500.

McLellan, R. A., Schmidt, M., Amundson, M., & Blake, R. (1998, May). *Secret shopper as an individual-level criterion for validation studies*. Paper presented at the Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Murphy, K. R., & DeShon, R. (2000). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*(4), 873–900.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Van Iddekinge, C. H., Roth, P. L., Raymark, P. H., & Odle-Dusseau, H. N. (2012). The criterion-related validity of integrity tests: An updated meta-analysis. *Journal of Applied Psychology*, *97*(3), 499.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*(5), 557.

https://doi.org/10.1017/S1754942600006787 Published online by Cambridge University Press