Oswald, F. L., & McCloy, R. A. (2003). Meta-analysis and the art of the average. In K. Murphy (Ed.), *Validity generalization: A critical review* (pp. 311–338). Mahwah, NJ: Erlbaum.

Schmidt, F. L., Hunter, J. E., Pearlman, K., Hirsh, H. R., Sackett, p. R., Schmitt, N., … Sedeck, S. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*(4), 697–798.

Tett, R. P., Hundley, N., & Christiansen, N. D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10*(3), 421–456.

A Failed Challenge to Validity Generalization: Addressing a Fundamental Misunderstanding of the Nature of VG

Frank L. Schmidt
University of Iowa

Chockalingam Viswesvaran
Florida International University

Deniz S. Ones
University of Minnesota

Huy Le
University of Texas at San Antonio

The lengthy and complex focal article by Tett, Hundley, and Christiansen (2017) is based on a fundamental misunderstanding of the nature of validity generalization (VG): It is based on the assumption that what is generalized in VG is the estimated value of mean rho ($\bar{\rho}$). This erroneous assumption is stated repeatedly throughout the article. A conclusion of validity generalization does not imply that $\bar{\rho}$ is identical across all situations. If VG is present, most, if not all, validities in the validity distribution are positive and useful even if there is some variation in that distribution. What is generalized is the entire distribution of rho ($\bar{\rho}$), not just the estimated $\bar{\rho}$ or any other specific value of validity included in the distribution. This distribution is described by its mean ($\bar{\rho}$) and standard deviation (SD_ρ). A helpful concept based

Frank L. Schmidt, University of Iowa; Chockalingam Viswesvaran, Florida International University; Deniz S. Ones, University of Minnesota; Huy Le, University of Texas at San Antonio.

The order of authorship following the lead author is by seniority. We thank Philip Roth and In-Sue Oh for suggestions and comments on an earlier draft of this manuscript.

Correspondence concerning this article should be addressed to Chockalingam Viswesvaran, Florida International University, Department of Psychology, University Park, Miami, FL 33199. E-mail: Vish@fiu.edu; or Deniz S. Ones, University of Minnesota, Department of Psychology, 75 East River Road, Minneapolis, MN 55455. E-mail: Deniz.S.Ones-1@tc.umn.edu

on these parameters (assuming ρ is normally distributed) is the credibility interval, which reflects the range where most of the values of ρ can be found. The lower end of the 80% credibility interval (the 90% credibility value, CV $= \bar{\rho} - 1.28 \times SD\rho$) is used to facilitate understanding of this distribution by indicating the statistical "worst case" for validity, for practitioners using VG. Validity has an estimated 90% chance of lying above this value. This concept has long been recognized in the literature (see Hunter & Hunter, 1984, for an example; see also Schmidt, Law, Hunter, Rothstein, Pearlman, & McDaniel, 1993, and hundreds of VG articles that have appeared in the literature over the past 40 years since the invention of psychometric meta-analysis as a means of examining VG [Schmidt & Hunter, 1977]). The $\bar{\rho}$ is the value in the distribution with the highest likelihood of occurring (although often by only a small amount), but it is the whole distribution that is generalized. Tett et al. (2017) state that some meta-analysis articles claim that they are generalizing only $\bar{\rho}$. If true, this is inappropriate. Because $\bar{\rho}$ has the highest likelihood in the ρ distribution, discussion often focuses on that value as a matter of convenience, but $\bar{\rho}$ is not what is generalized in VG. What is generalized is the conclusion that there is validity throughout the credibility interval. The false assumption that it is $\bar{\rho}$ and not the ρ distribution as a whole that is generalized in VG is the basis for the Tett et al. article and is its Achilles heel. In this commentary, we examine the target article's basic arguments and point out errors and omissions that led Tett et al. to falsely conclude that VG is a "myth."

Validity Distributions Are Generalized in VG, Not Mean Corrected Validity

Tett et al. (2017) argue that if the overall ρ distribution has a $SD\rho$ that they regard as *too large for generalizing the mean rho*, and if subgrouping moderator analyses indicate that some or all of the subgroup $\bar{\rho}$s are different from the overall $\bar{\rho}$ and the corresponding subgroup $SD\rho$ values are smaller than the overall $SD\rho$, then the overall ρ distribution (in Tett et al.'s terms, the overall $\bar{\rho}$) cannot be the basis for VG because there are moderators within the overall ρ distribution (in Tett et al.'s terms, subgroup $\bar{\rho}$s different from the overall $\bar{\rho}$ and each other). This conclusion is false.

The initial distribution of ρ still allows VG, because what is generalized is the ρ distribution, not just $\bar{\rho}$, as we have clarified above. The user can generalize across the moderators; they fall within the credibility interval because they are included in the validity distribution, and the overall distribution indicates that validity (and validity generalization) is present regardless of whether moderators exist or not. Although it is true that the larger $SD\rho$ values often found in the overall ρ distribution mean less certainty of the actual value of ρ in any user application, the distribution still indicates the

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press

presence of practical validity. This is what is important in determining whether a procedure can be used with positive utility. This is what VG is.

It is important to note that the same principle applies to the subgroup moderator analyses: Although the uncertainty may be reduced somewhat due to a smaller $SD\rho$, there is still considerable uncertainty (as discussed later in our section below on second order sampling error). So there is no qualitative difference between the overall distribution and the moderator-level distributions.

Standards for Generalization and Indices for Precision

Tett et al. (2017) state that VG in the sense described here sets a standard for generalization that is too low. This is an arbitrary judgment on their part. Further, their judgment is based on the false assumption that what is generalized in VG is $\bar{\rho}$, as noted above. Throughout the target article, the authors state that $SD\rho$ is the measure of the precision of the estimate of $\bar{\rho}$. This is not true. The measure of the precision of the estimate of $\bar{\rho}$ is the standard error of $\bar{\rho}$, ($SE\bar{\rho}$), which they do not calculate or present anywhere in their article. It is also missing from their figure 1.[1]

Although Tett et al. (2017) mistakenly use $SD\rho$ as the measure of the precision of the $\bar{\rho}$ estimate, they also include a measure of what they call uncertainty in the estimate of the mean observed r ($\bar{r}$) (p. 12); this is the *SE* of the $\bar{r}$, symbolized by them as SEr_{xy} (their symbol omits the usual bar [representing the mean] over r that indicates the *SE* applies to the $\bar{r}$). What should be used here is again the *SE* of $\bar{\rho}$. They justify use of the *SE* of $\bar{r}$ in footnote 15 (p. 449), which states "Practical applications call for generalizability of mean *r*, not mean rho." This is incorrect for two reasons. First, what is relevant in VG is ρ**,** operational validity in selection research**;** given the focus on mean rho here, it is the mean observed validity corrected for measurement error in the criterion measure and applicable range restriction). It is not r or $\bar{r}$. Second, it is the distribution of ρ that is generalized not the single point in the distribution that is the $\bar{\rho}$ estimate.

Based on their assumption that it is $\bar{\rho}$ that is generalized in VG, the article by Tett et al. (2017) sets up arbitrary standards for generalizability; that is, their standards *for generalizing* $\bar{\rho}$. Using a select set of 24 industrial and organizational (I-O) psychology meta-analyses from the literature, they show that when the overall ρ distribution is broken down into subgroup analyses based on potential moderators, the subgroup $\bar{\rho}$ estimates are often different from the $\bar{\rho}$ estimate in the overall ρ distribution, and the SD_ρ

[1] The equations for *SE* of $\bar{\rho}$ are given in Schmidt and Hunter (2015, p. 364). They are also given in the Hunter and Schmidt (2004) book cited by Tett et al. (2017). See Burke and Landis (2003) for further elaboration on these equations.

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press

estimates are often smaller than the overall SD_ρ estimate. Further, the percent variance accounted for is often larger. These findings are not surprising and are expected. Tett et al.'s article also shows that the *SE* of the $\bar{r}$ is larger in the subgroup analyses. This means that the estimated $\bar{r}$ is less precise in the subgroup analyses. (Again, *SE* of $\bar{r}$ is not the relevant statistic; *SE* of $\bar{\rho}$ should have been used.) This is also expected, because each subgroup analysis is based on fewer studies (*k*) and a smaller total *N*. Their "finding" that estimates of $\bar{r}$ have a larger *SE* when the number of studies (*k*) is reduced in subgroup analysis has been known for years; it is not new. (But again, the relevant statistic is the *SE* of $\bar{\rho}$ [which will also be larger in the subgroups], not the *SE* of $\bar{r}$.) When they evaluate the results of all meta-analyses (overall analyses and moderator/subgroup analyses) against their arbitrary standards for generalizability of $\bar{\rho}$, they conclude that these standards are met only in about 4% of the analyses. However, not only are their generalizability standards arbitrary, but the basis for these standards is the false assumption that it is the estimate of $\bar{\rho}$, and not the distribution of ρ, that is generalized. Hence, Tett et al.'s conclusion that generalizability is rare in the meta-analytic literature is not correct. VG should be interpreted as a matter of degree, not mechanically as a matter of dichotomy (VG or not).

The article by Tett et al. (2017) does acknowledge that some artifacts that cause variance in SD_ρ estimates in meta-analytic study results are typically not corrected for, meaning that estimates of SD_ρ are almost always overestimates and thus exaggerate the amount of uncertainty in the ρ distribution. However, they do not make any allowance for this fact when they set up their arbitrary standards for generalizability. This is an important omission, because such considerations are important. In those cases in which it is possible to correct for the effects of artifacts that are not corrected for in other meta-analyses, the results show that such corrections greatly reduce the size of the SD_ρ estimates. For example, in the meta-analytic study by Schmidt et al. (1993), simply removing non-Pearson *r*s (which have larger sampling error variances than Pearson *r*s) led to a large increase in percent variance accounted for and to smaller SD_ρ values**.** This article lists six other artifacts that the authors were unable to correct for, suggesting that actual SD_ρ values were even smaller than they reported.

Some of the "moderators" listed in Tett et al.'s (2017) table 1 are actually artifacts whose effects should be corrected for, not real moderators. Examples include measurement length—multi-item vs. single item, and independent variable and dependent variable share common bias. Also, it is possible that in creating their overall VG analyses, some of the 24 studies included may have violated the requirements of meta-analysis (Schmidt & Hunter, 2015) by including completely different independent and/or dependent

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press

variable constructs.[2] An example would be a meta-analysis that included as dependent variable measures of both job performance and counterproductive work behaviors. Another example would be a personality meta-analysis that included several personality constructs (e.g., Conscientiousness, Ambition, and Traditionalism) in the same overall VG analysis. Although it has long been known that only the same or very similar constructs should be included in a meta-analysis, it is not uncommon to see this requirement violated. In such cases, the subsequent subgroup analyses will appear to indicate moderator effects (perhaps *large* moderator effects) that are not real moderator effects but rather construct effects. This could account for some of the reductions in SD_ρ and increases in percent variance accounted for observed by Tett et al. in moving from the overall VG distribution to moderator subgroupings (i.e., mixing apples and oranges first in the overall meta-analysis and then separating apples from oranges in the subsequent subgroup analyses; Cortina, 2003).

On the Importance of Considering Second-Order Sampling Error in VG Research

Tett et al.'s (2017) article also discusses second-order sampling error and publication bias, but does so only in passing. When the goal is to understand the actual distribution of validities, second-order sampling error cannot be ignored. When the number of independent samples contributing to a meta-analysis (k) is small for each subgroup, seemingly different subgroup $\bar{\rho}$s with associated SD_ρ values (often smaller than the SD_ρ from the overall analysis) are likely to represent an artifactual finding that is due to second-order sampling error, not real moderating effects (see examples in Schmidt & Oh, 2013). To the extent that the number of independent samples included in a meta-analysis are small, there is a random chance element involved in uncovering the real distribution of validities (i.e., second order sampling error; Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper & Koenka, 2012; Hunter & Schmidt, 2004; Schmidt & Hunter, 2015). Further, publication (availability) bias traditionally conceived, if present, tends to overestimate $\bar{\rho}$ and to underestimate SD_ρ (assuming $\bar{\rho}$ is positive in the population).[3] Thus,

[2] This may or may not have happened in this particular sample of meta-analyses includes in Tett et al.'s (2017) article, but it has happened with some frequency in the general literature, and so it is a point worth mentioning. Our ability to check was limited by the one-month time limit given to comment on target articles in this journal.

[3] Although publication bias typically and traditionally has been used to refer to nonrepresentative availability of significant positive effect sizes in the published literature, there are research areas where the published literature appears to be distorted by favoring nonsignificant/nil effects (e.g., meta-analyses of published articles on the validities of standardized admissions tests yields somewhat lower validities than those available in unpublished sources [e.g., dissertations, technical reports; Hezlett et al., 2001]).

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press

second-order sampling error and publication bias have important implications for VG. Methods for second-order sampling error are covered in detail in Schmidt and Oh (2013).[4] Publication bias has been thoroughly discussed in many places in the literature; for an overview, see Schmidt and Hunter (2015, chapter 13). Ones, Viswesvaran, and Schmidt (2017) provide concrete, valuable recommendations about ensuring *comprehensive* databases in meta-analytic research. We refer the reader to these sources. Tett et al.'s minimal attendance to potential second-order sampling error in their subgroup analyses is at best naïve and could lead to erroneous conclusions about the actual distributions of validities when comparing results across different subgroupings.

Calling Into Question Previous Research Findings in I-O Psychology

Tett et al.'s (2017) article is an example of a recent pattern in the I-O psychology literature, namely attempts to reopen a settled question in I-O psychology with the goal of overturning well established research findings. We have no qualms about overturning established theories and empirical findings based on accumulating data and new evidence. However, we find such attempts to be hollow when (a) there is little or no new empirical evidence that is brought to bear by the newer publications, or when (b) techniques and statistical corrections used are inappropriate and inadequate. An early example is Tett, Jackson, and Rothstein's (1991) personality–job performance meta-analysis (see Ones, Mount, Barrick, and Hunter [1994] for a methodological critique of this study). Recent examples include the challenges to established findings regarding the lack of differential validity by race of cognitive ability tests (original research by Hunter, Schmidt, and Hunter [1979]), validity of integrity tests (original research by Ones, Viswesvaran, and Schmidt [1993]), and measurement error in job performance ratings (original research by Viswesvaran, Ones, and Schmidt [1996]). In all these cases, statistical and psychometric misunderstandings led to misinformed analyses and/or inaccurate results. Significant resources and subsequent articles had to be devoted to correcting the scientific record (e.g., Harris et al., 2012; Ones, Viswesvaran, & Schmidt, 2012; Roth et al., 2014; Roth, Le, Oh, Van Iddekinge, & Robbins, 2017; Viswesvaran, Ones, & Schmidt, 2016; and Viswesvaran, Ones, Schmidt, Le, & Oh, 2014).

The focal article by Tett et al. (2017) attempts to challenge the established principles of meta-analytic VG, contending that VG is "a myth." Our commentary explains that this attempt is based on a central underlying

[4] See Ones et al. (2012) for how second-order sampling error can create the illusion of cross-cultural variability and how meta-analyses in general and consideration of second-order sampling error can disentangle sampling error from true cross-cultural differences.

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press

assumption that is false. The entire validity distribution is generalized in VG, not just mean corrected validity. Tett et al. describe arbitrary and erroneous standards for VG and rely on incorrect indices for precision. Their inadequate attention to second-order sampling error ignores an important potential threat to the veracity of meta-analytic inferences. As a result of these errors, their attempt to challenge established VG principles and findings fails.

References

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Burke, M. J., & Landis, R. (2003). Methodological and conceptual issues in applications of meta-analysis. In K. Murphy (Ed.), *Validity generalization: A critical review* (pp. 287–310). Hillsdale, NJ: Lawrence Erlbaum.

Cooper, H., & Koenka, A. C. (2012). The overview of reviews: Unique challenges and opportunities when research syntheses are the principal elements of new integrative scholarship. *American Psychologist, 67*, 446–462.

Cortina, J. M. (2003). Apples and oranges (and pears, oh my!): The search for moderators in meta-analysis. *Organizational Research Methods, 6*, 415–439.

Harris, W. G., Jones, J. W., Klion, R., Arnold, D, Camara, W., & Cunningham, M. R. (2012). Test publishers' perspective on "An updated meta-analysis": Comment on Van Iddekinge, Roth, Raymark, and Odle-Dusseau (2012). *Journal of Applied Psychology, 97*(3), 531–536.

Hezlett, S. A., Kuncel, N. R., Vey, M. A., Ahart, A., Ones, D. S., Campbell, J. P., & Camara, W. (2001, April). The predictive validity of the SAT: A comprehensive meta-analysis. In D. S. Ones & S. A. Hezlett (Chairs), *Predicting performance: The interface of I-O psychology and educational research*. Symposium presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72–98. doi: 10.1037/0033-2909.96.1.72

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research finding* (2nd ed.). Newbury Park, CA: Sage.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, *86*(4), 721–735. doi: 10.1037/0033-2909.86.4.721

Ones, D. S., Dilchert, S., Deller, J., Albrecht, A.-G., Duehr, E. E., & Paulus, F. M. (2012). Cross-cultural generalization: Using meta-analysis to test hypotheses about cultural variability. In A. M. Ryan, F. T. L. Leong, & F. L. Oswald (Eds.), *Conducting multinational research projects in organizational psychology: Challenges and opportunities* (pp. 91–122). Washington, DC: American Psychological Association.

Ones, D. S., Mount, M. K., Barrick, M. R., & Hunter, J. E. (1994). Personality and job performance: A critique of the Tett, Jackson, and Rothstein (1991) meta-analysis. *Personnel Psychology, 47*, 147–156.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (1993). Comprehensive meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *Journal of Applied Psychology [Monograph], 78*, 679–703.

Ones, D. S., Viswesvaran, C., & Schmidt, F. L. (2012). Integrity tests predict counterproductive work behaviors and job performance well: A comment on Van Iddekinge et al. *Journal of Applied Psychology*, *97*(3), 537–542. doi: 10.1037/a0024825

Ones, D.S., Viswesvaran, C., & Schmidt, F. L. (2017). Realizing the full potential of psychometric meta-analysis for a cumulative science and practice of human resource management. *Human Resource Management Review*, *27*(1), 201–215.

Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C., Buster, M. A., Robbins, S. B., & Campion, M. A. (2014). Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology, 99*, 1–20.

Roth, P. L., Le, H., Oh, I.-S., Van Iddekinge, C. H., & Robbins, S. B. (2017). Who r u? On the (in)accuracy of incumbent-based estimates of range restriction in criterion-related and differential validity research. *Journal of Applied Psychology, 102*(5), 802–828. doi: 10.1037/apl0000193

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529–540.

Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Thousand Oaks, CA: Sage.

Schmidt, F. L., Law, K., Hunter, J. E., Rothstein, H. R., Pearlman, K., & McDaniel, M. (1993). Refinements in validity generalizations methods: Implications for the situational specificity hypothesis. *Journal of Applied Psychology, 78*, 3–13.

Schmidt, F. L., & Oh, I.-S. (2013). Methods for second order meta-analysis and illustrative applications. *Organizational Behavior and Human Decision Processes, 121*, 204–218.

Tett, R. P., Hundley, N. A., & Christiansen, N. D. (2017). Meta-analysis and the myth of generalizability. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 10*(3), 421–456.

Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703–742.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557–574.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (2016). Comparing rater groups: How to disentangle rating reliability from construct-level disagreements. *Industrial and Organizational Psychology, Perspectives on Theory and Practice, 9*, 800–806.

Viswesvaran, C., Ones, D. S., Schmidt, F. L., Le, H., & Oh, I.-S. (2014). Measurement error obfuscates scientific knowledge: Path to cumulative knowledge requires corrections for unreliability and psychometric meta-analysis. *Industrial and Organizational Psychology: Perspectives on Theory and Practice, 7*, 507–518.

Generalizability Versus Situational Specificity in Adverse Impact Analysis: Issues in Data Aggregation

Elizabeth Howard and Scott B. Morris
Illinois Institute of Technology

Eric Dunleavy
DCI Consulting Group

Tett, Hundley, and Christiansen (2017) argue that the concept of validity generalization in meta-analysis is a myth, as the variability of the effect size appears to decrease with increasing moderator specificity such that the level

Elizabeth Howard, Illinois Institute of Technology; Scott B. Morris, Illinois Institute of Technology; Eric Dunleavy, DCI Consulting Group.

Correspondence concerning this article should be addressed to Elizabeth Howard, Illinois Institute of Technology, 3105 S. Dearborn, Chicago, IL 60616. E-mail: ehoward3@iit.edu

https://doi.org/10.1017/iop.2017.47 Published online by Cambridge University Press