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.2017.48 Published online by Cambridge University Press

of precision needed to deem an estimate "generalizable" is actually reached at levels of situational specificity that are so high as to (paradoxically) refute an inference of generalizability. This notion highlights the need to move away from claiming that effects are either "generalizable" or "situationally specific" and instead look more critically and less dichotomously at degrees of generalizability, or effect size variability.

We offer an additional application of this perspective and extend the authors' recommendations to the context of adverse impact analysis and the problem of determining when multiple subsamples are "similar enough" (i.e., lacking enough in situational specificity) to justify aggregating data into a single, larger sample for analysis. The multiple-event methods that are commonly used to aggregate adverse impact data bear a close resemblance to meta-analysis, in that a summary statistic is computed for each subsample and then a weighted average is used to represent the overall trend (Morris, Dunleavy, & Lee, 2017). In fact, the most common method for conducting a multiple events test, the Mantel–Haenszel test (Mantel & Haenszel, 1959), is also used as a method for meta-analysis of categorical data (Fleiss & Berlin, 2009). The approach that Tett et al. (2017) recommend for evaluating between-study variability and the generalizability of validity evidence would prove useful in the adverse impact context as well.

Aggregation in Adverse Impact Analysis

Adverse impact refers to statistical disparities in employment outcomes for different protected groups, and these statistical disparities can form the basis for an allegation of illegal discrimination in employment practices (Dunleavy, Morris, & Howard, 2015). Although these statistics do not in and of themselves constitute definitive proof of illegal discrimination, they are often a prerequisite that plaintiffs or agencies must meet when alleging a case of pattern and practice or disparate impact discrimination, and they do create an obligation for the employer to explain the disparities, justify the practices that create them, and/or consider reasonable alternatives (Gutman, Koppes, & Vodanovich, 2010). If this obligation cannot be met, the practices in question may be considered discriminatory. As such, these are extremely high-stakes analyses.

Adverse impact analytics are often complicated by the fact that an organization's selection decisions may be collected across multiple work locations, jobs, and points in time (i.e., subsamples with their own unique characteristics), and sample sizes for any one subsample may be so small that statistical analyses have little power (Morris, 2001). As such, there is frequently a question of whether to analyze data for the subsamples individually or to aggregate them and create a larger aggregate sample with higher power for statistical analysis (i.e., meta-analyze). The results of these analyses can differ

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press

substantially depending on whether data are aggregated or not, and as such, decisions around aggregation are often topics of much contention in the legal arena (e.g., *Wal-Mart Stores, Inc. v. Dukes*, 2011).

Defendants often argue that separate analyses must be conducted for each subsample, alleging that each represents a distinct decision-making process. Conversely, plaintiffs tend to argue for aggregated data as the best evidence for systemic discrimination. (For a more thorough discussion of the legal arguments, see Mehri and Lieder [2017] and Ross and Merrill [2017].) The two competing perspectives bear a close resemblance to the concepts of situation specificity and validity generalization that are discussed by Tett et al. (2017). Given the high stakes of these analyses, it is of considerable importance that the appropriateness of aggregation be carefully evaluated.

Evaluating the Appropriateness of Aggregation

Aggregation of adverse impact data can be justified to the extent that the different subsamples represent a common decision process, which depends on both conceptual and empirical considerations (Morris, Dunleavy, & Lee, 2017). Just as Tett et al. (2017) discuss in the context of validity generalization, the average result can be generalized to the individual settings only if results are fairly consistent across those settings. Empirically, it is common to observe that adverse impact is larger in some settings and smaller in others. But just as in meta-analysis, some of this observed variability is expected due to first-order sampling error. Pattern consistency tests (Biddle, 2011) can determine whether this variability of results across subsamples is larger than what would be expected due to sampling error. The widely used Breslow–Day Test (Breslow & Day, 1980) evaluates homogeneity of the odds ratio, with a significant result indicating a *lack* of consistency in the odds ratio (i.e., in the degree of adverse impact) across subsamples.

Some experts have recommended using the Breslow–Day Test as a preliminary analysis to determine whether aggregation is appropriate (Biddle, 2011; Cohen, Aamodt, & Dunleavy, 2010). In this approach, if the Breslow–Day test is significant, there are differences in the degree of adverse impact across settings (i.e., situation specificity), and an aggregated analysis would typically not be recommended. If, on the other hand, the Breslow–Day is nonsignificant, the results are generalizable from an empirical perspective, and the aggregation would typically be supported. Recently, Morris, Dunleavy, and Lee (2017) proposed a more flexible framework in which a significant Breslow–Day Test signals a need to interpret situational influences on adverse impact but does not act as an absolute bar to conducting an aggregate analysis. This alternative perspective will be further detailed below.

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press

Not Black and White: Degrees of Heterogeneity

In the above debate over when it is and is not statistically appropriate to aggregate, we see the same need to move beyond black-and-white thinking that is highlighted by Tett et al.'s (2017) discussion of the artificial dichotomy between generalizability and situational specificity. A finding of between-study variability does not render the average effect size uninterpretable, but it does limit the generalizability of inferences because it is not clear which effect size applies to a particular situation.

We recommend that rather than interpreting a significant Breslow–Day result as a red light prohibiting aggregation and a nonsignificant result as a green light providing statistical permission to aggregate, as some practitioners have begun to view them, we should consider both the significance or nonsignificance of the result and the actual degree of variability in effect size across subsamples. Here we echo Tett et al.'s (2017) position that we should focus on gathering "multiple, preferably converging lines of evidence" (p. 452) to support the inferences we draw from statistical data. In the case of adverse impact analysis, we argue that average effect sizes should be interpreted in light of both statistical significance and degree of variability (i.e., credibility intervals around the effect size), and all of this statistical evidence should be evaluated in light of other (i.e., conceptual) evidence related to the extent to which the same employment practices have affected individuals in each subsample.

The need to take a closer, less dichotomous look at subsample variability in effect size is especially critical in cases of adverse impact given that, as the Supreme Court noted in *Wal-Mart Stores, Inc. v. Dukes* (2011), statistically significant disparities between groups may arise not from a uniform practice carried out by all decision-making units within an organization but from discriminatory behavior that is specific to just a few of those decision-making units. In light of this, Mehri and Lieder (2017) argue that rather than applying stringent black-and-white rules or insisting that any statistical disparities be identical in each subsample, the courts should critically evaluate whether the disparities are more consistent with the existence of discrimination affecting all subsamples or more consistent with the existence of discrimination in relatively few of the subsamples, given the variation in results that should be expected due to chance. We believe that meta-analysis provides several methods that are well suited to this type of inquiry.

Methods for Evaluating Degree of Heterogeneity

Tett et al. (2017) offer several practical recommendations for reporting and interpreting meta-analytic results that we believe should be adopted in multiple-event adverse impact analysis. First, the researchers should evaluate and report the precision and certainty of generalizability estimates.

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press

Second, when substantial variability across samples is found, the interpretation should shift from a focus on the mean to a focus on the credibility interval. Here, we offer some suggestions for how practitioners can apply these recommendations when evaluating the appropriateness of aggregation in adverse impact analysis.

As noted, many practitioners have come to rely on the Breslow–Day Test to provide a yes-or-no answer to the question of whether data should be aggregated. Here, we offer some methods that may be used as supplements or alternatives to the Breslow–Day, allowing the practitioner to develop a more comprehensive picture of the variability that exists in the pattern and degree of adverse impact across subsamples.

1. *Examine the actual pattern of heterogeneity.* Following a significant Breslow–Day Test, it is often informative to look at the distribution of results across settings (Mehri & Lieder, 2017; Morris et al., 2017). Consider a situation where the adverse impact ratio varies from .2 to .6 across locations, and this variability results in a significant Breslow–Day Test. Although it is clear that the level of adverse impact is not identical in every setting, it is also the case that substantial adverse impact exists across all locations. The presence of variability in the size of the disparity may not necessarily justify analyzing each setting separately and requiring that each show a statistically significant disparity in order to make some empirically driven conclusions.
2. *Identify relatively homogeneous subgroups.* Tett et al. (2017) demonstrate that subgroup analyses that focus on more similar situations tend to show smaller between-study variability. Similarly, in adverse impact analysis, it will often be useful to identify moderator variables or substrata (e.g., geographical regions, applicant population characteristics) that account for some of the situation specificity. Tett et al. also caution that the improved generalizability with more narrow subgroups often comes at the cost of decreased certainty (i.e., larger confidence intervals on the mean effect size) due to the reduced number of studies. Thus, it is important to avoid improperly defining the situation so narrowly that power becomes too low to detect disparities (Bielby & Coukos, 2007).
3. *Apply meta-analytic methods such as random effects models, credibility intervals, and empirical-Bayes estimates to evaluate subsample heterogeneity.* Methods developed for exploring situation-specificity in meta-analysis are likely to prove useful in the context of adverse impact as well (Morris et al., 2017). Application of random effects models (obtained, for example, through generalized linear modeling; Huang & Morris, 2013) could provide information regarding both the mean and variability of the odds ratio. Constructing a credibility interval from the results

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press

of a random effects model would provide a clear picture of the degree of heterogeneity and help to determine whether variability in effect size is a matter of degree or a matter of kind. That is, do the endpoints of the credibility interval both point to similar conclusions about the pattern of adverse impact (e.g., relatively small differences in magnitude across subsamples, or substantial differences in magnitude that are all in the same direction), or do they point to fundamentally different conclusions (e.g., adverse impact favoring the majority group in some settings but favoring the minority group in other settings)?

Another potentially useful methodology from the meta-analytic toolbox may be the empirical-Bayes estimate of the study-specific effect size (Brannick, 2001). When a meta-analysis results in a wide credibility interval, we do not know which effect size in that interval applies to a particular situation (Tett et al., 2017). At the same time, due to small sample size, results for individual subsamples may be imprecise. By combining data from an aggregate analysis with sample-specific information, empirical Bayes estimates provide a more stable estimate of the degree of adverse impact in a particular setting.

Conclusion

In the focal article, Tett et al. (2017) make a compelling case for paying greater attention to the variability of results across settings (e.g., situational specificity) in meta-analysis of selection procedure validity. This perspective is also relevant in the context of adverse impact analytics. Due to a variety of factors influencing employment outcomes, ranging from applicant pool characteristics to nuances of the decision-making process (Arthur, Doverspike, Barrett, & Miguel, 2013), adverse impact statistics are likely to show some degree of situational specificity. When evaluating adverse impact in such settings, we need to move beyond sole reliance on significance testing to provide a simple yes-or-no answer to this complex and nuanced question. Methods developed for examining heterogeneity in meta-analysis provide potentially useful tools for exploring patterns of adverse impact in large-scale employment practices as well.

References

Arthur, W., Doverspike, D., Barrett, G. V., & Miguel, R. (2013). Chasing the Title VII holy grail: The pitfalls of guaranteeing adverse impact elimination. *Journal of Business and Psychology, 28*(4), 473–485.

Biddle, D. A. (2011). *Adverse impact and test validation: A practitioner's handbook* (3rd ed.). Scottsdale, AZ: Infinity.

Bielby, W. T., & Coukos, P. (2007). “Statistical dueling” with unconventional weapons: What courts should know about experts in employment discrimination class actions. *Emory Law Journal, 56*, 1563–1612.

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press

Brannick, M. T. (2001). Implications of empirical Bayes meta-analysis for test validation. *Journal of Applied Psychology, 86*, 468–480.

Breslow, N. E., & Day, N. E. (1980). Statistical methods in cancer research, Volume 1: The analysis of case-control studies (Vol. 32). Lyon, France: IARC Scientific.

Cohen, D. B., Aamodt, M. G., & Dunleavy, E. M. (2010). *Technical advisory committee report on best practices in adverse impact analyses*. Washington, DC: Center for Corporate Equality.

Dunleavy, E., Morris, S., & Howard, E. (2015). Measuring adverse impact in employee selection decisions. In C. Hanvey & K. Sady (Eds.), *Practitioner's guide to legal issues in organizations* (pp. 1–26). Cham, Switzerland: Springer International Publishing.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–254). New York: Russell Sage Foundation.

Gutman, A., Koppes, L. L., & Vodanovich, S. J. (2010). *EEO law and personnel practices*. Abingdon, UK: Psychology Press.

Huang, J., & Morris, S. B. (2013, April). *HGLM and Mantel–Haenszel tests for adverse impact*. Poster presented at the 28th Annual Conference of the Society for Industrial and Organizational Psychology, Houston, TX.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719–748.

Mehri, C., & Lieder, M. (2017). Addressing the ever increasing standards for statistical evidence: A plaintiff attorney's perspective. In S. B. Morris and E. M. Dunleavy (Eds.), *Adverse impact analysis* (pp. 298–329). New York: Routledge.

Morris, S. B. (2001). Sample size required for adverse impact analysis. *Applied H.R.M. Research, 6*, 13–32.

Morris, S. B., Dunleavy, E. M., & Lee, M. (2017). Many 2x2 tables: Understanding multiple events in adverse impact analyses. In S. B. Morris and E. M. Dunleavy (Eds.), *Adverse impact analysis* (pp. 147–166). New York, NY: Routledge.

Ross, D. B., & Merrill, G. (2017). Disparate impact, trial by statistics: Thoughts from a defense attorney's perspective. In S. B. Morris and E. M. Dunleavy (Eds.), *Adverse impact analysis* (pp. 330–348). New York Routledge.

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.

Wal-Mart Stores, Inc. v. Dukes, 131 S. Ct. 2541 (2011).

https://doi.org/10.1017/iop.2017.48 Published online by Cambridge University Press