External Validity and Multi-Organization Samples: Levels-of-Analysis Implications of Crowdsourcing and College Student Samples

Daniel A. Newman, Dana L. Joseph and Jennifer Feitosa

Industrial and Organizational Psychology / Volume 8 / Issue 02 / June 2015, pp 214 - 220
DOI: 10.1017/iop.2015.28, Published online: 28 July 2015

Link to this article: http://journals.cambridge.org/abstract_S1754942615000280

How to cite this article:
Daniel A. Newman, Dana L. Joseph and Jennifer Feitosa (2015). External Validity and Multi-Organization Samples: Levels-of-Analysis Implications of Crowdsourcing and College Student Samples. Industrial and Organizational Psychology, 8, pp 214-220 doi:10.1017/iop.2015.28

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

DeSimone, J. A., Harms, P. D., & DeSimone, A. J. (2014). Best practice recommendations for data screening. *Journal of Organizational Behavior*. Advance online publication. doi:10.1002/job.1962

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., & DeShon, R. P. (2012). Detecting and deterring insufficient effort respond to surveys. *Journal of Business and Psychology*, *27*, 99–114. doi:10.1007/s10869-011-9231-8

Huang, J. L., Liu, M., & Bowling, N. A. (2014). Insufficient effort responding: Examining an insidious confound in survey data. *Journal of Applied Psychology.* Advance online publication. doi:10.1037/a0038510

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice.*

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *65*, 539–569.

Shoss, M. (2013, September 16). Re: Collecting longitudinal data [Electronic mailing list message]. Retrieved from RMNet listserv: https://lists.unc.edu/

Williams, L. J., & Anderson, S. E. (1994). An alternative approach to method effects by using latent-variable models: Applications in organizational behavior research. *Journal of Applied Psychology*, *79*(3), 323–331.

External Validity and Multi-Organization Samples: Levels-of-Analysis Implications of Crowdsourcing and College Student Samples

Daniel A. Newman
University of Illinois at Urbana–Champaign

Dana L. Joseph and Jennifer Feitosa
University of Central Florida

Here, we expand on Landers and Behrend's (2015) discussion of the external validity of convenience samples. In particular, we note that their focal article failed to mention one important limitation of multi-organization

Daniel A. Newman, Department of Psychology and School of Labor & Employment Relations, University of Illinois at Urbana–Champaign; Dana L. Joseph and Jennifer Feitosa, Department of Psychology, University of Central Florida.

Correspondence concerning this article should be addressed to Daniel A. Newman, Department of Psychology, University of Illinois at Urbana–Champaign, 603 East Daniel Street, Champaign, IL 61820. E-mail: d5n@uiuc.edu

convenience samples (e.g., MTurk samples, student samples): Multi-organization convenience samples tend to confound levels of analysis, which affects the external validity of these samples. Specifically, between-organizations phenomena (i.e., organization-level) and within-organizations phenomena (i.e., individual-level) are distinct and separable (Ostroff, 1993; Robinson, 1950). Unfortunately, multi-organization samples such as those found in MTurk or MBA student samples can confound these two sets of phenomena. The current commentary uses a levels-of-analysis framework to expand on Landers and Behrend's discussion of what external validity is, and then the commentary illustrates how the diversity of convenience samples can actually harm external validity under some common circumstances.

Levels of Analysis: Theorizing about Individuals and Organizations

A primary purpose of organizational research is to estimate theoretically interesting parameters (e.g., the relationship between personality and counterproductive work behavior [CWB] or the relationship between job satisfaction and work withdrawal). Of note, organizational phenomena can reside at different levels of analysis, including the within-group and the between-group levels.

Thorndike (1939) and Robinson (1950) classically demonstrated that within-group correlations and between-group correlations need not be similar (also see discussions by Dansereau, Alluto, & Yammarino, 1984; Ostroff, 1993; Simpson, 1951; Yule, 1903). This is illustrated in Figure 1, where we see that the within-group correlation between X and Y (r_{within}) can be different from the between-group correlation between X and Y ($r_{between}$). What is true within-groups (e.g., each oval has a positive slope, indicating a positive within-group correlation) need not be true between groups (e.g., the slope across ovals is negative, indicating a negative between-group correlation).

As a hypothetical example, suppose we are interested in the relationship between individual-level job satisfaction (X) and individual turnover intentions (Y). We might find that the individual-level relationship between satisfaction and turnover intentions (e.g., the slope of the X-Y line *within* an oval, or r_{within}; cf. Figure 1) is negative. When individuals *within an organization* are more satisfied, they are less likely to plan to quit. At the same time, it is entirely possible that at the *organizational level of analysis* (e.g., the slope of the X-Y line *between* different ovals, $r_{between}$), the satisfaction–turnover intention relationship could be zero or even positive. That is, it is plausible that entire organizations with high organizational-mean levels of satisfaction might experience higher organizational-mean levels of turnover. One theory potentially implying this is the frame-of-reference effect (Hernandez, Newman, & Jeon, in press; Hulin, 1966), in which organizations in poor/less affluent areas experience higher job sat-

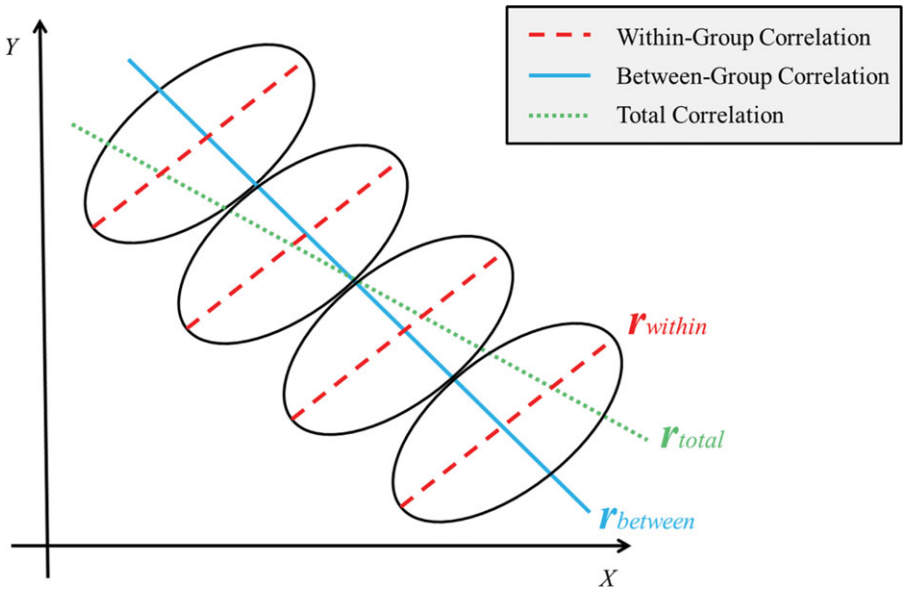

Figure 1. Illustration that the Between-Group *X*-*Y* Relationship Can Differ from the Within-Group *X*-*Y* Relationship

Note. **Each oval represents one group. The total correlation, r_{total}, is the correlation that would have been found if one had ignored group membership, and it is a weighted average of r_{within} and $r_{between}$ (see Ostroff, 1993). (Color online)**

isfaction due to the social comparisons being made by workers to others in the local community. If these same organizations in less affluent geographic areas have workers who intend to leave the poor community for various reasons, then we might observe a *positive* relationship between job satisfaction and turnover intentions, at the organizational level of analysis (due to the fact that both organization-level satisfaction and organization-level turnover intentions are enhanced by low socioeconomic status conditions). Although this example is contrived, it makes the point that individual-level correlations do not imply group-level correlations of the same magnitude. Previous work has labeled inappropriate generalizations from the organizational level to the individual level as "ecological fallacies," whereas inappropriate generalizations from the individual level to the organizational level have been labeled "atomistic fallacies" (see Kozlowski & Klein, 2000).

External Validity

Equipped with an understanding of levels of analysis, we can now articulate a particular, commonly used notion of external validity. That is, we assert that many organizational researchers who use the term "external validity" are essentially defining *external validity* as the extent to which the magnitude of an individual-level relationship observed *within* one or more groups (e.g.,

r_{within}) generalizes to other groups. For example, does a published research finding from a single-organization sample generalize to other companies? This definition treats external validity as the absence of group-level moderators of individual-level effects. In Figure 1, external validity would imply that all ovals in a specified population have roughly the same slope (i.e., if the *X*-*Y* relationship, r_{within}, is sampled in one oval, the slope identified in this oval would be similar to that observed in other ovals). Although our current interpretation of external validity is not the only definition of external validity available (see the excellent discussion of external validity definitions by Landers and Behrend), we assert that it is one of the most common notions of external validity used in industrial–organizational (I-O) psychology (i.e., Do the results from this study generalize to other organizations?).

Multi-Organization Samples and External Validity

For organizational researchers who seek to study individual-level phenomena within organizations (e.g., the relationship between justice perceptions and CWB), there is a choice to make about which type of sample to use: *single-organization samples* versus *multi-organization samples*. According to Landers and Behrend's (2015) typology of convenience samples, *organizational samples* are usually single-organization samples: "the most typical convenience sample found in I-O psychology journals involves a single organization with which the researcher has some prior relationship" (p. 19). These are the samples that Landers and Behrend characterize as most esteemed by organizational researchers. In contrast, *crowdsourced (MTurk) samples*, *college student samples*, *online panels*, and *snowball samples* (all of which are less esteemed by organizational researchers) tend to be multi-organization samples, in which the respondents are affiliated with/employed by many different organizations. For example, in an MTurk sample of 200 respondents, it is likely that nearly 200 different organizations are represented (i.e., it is unlikely that a sizeable portion of MTurk respondents work for the same group or organization). As such, any research conducted on MTurk respondents will inherently be estimating *group-level phenomena (between groups)* rather than within-group/individual-level phenomena (e.g., see Figure 1, but imagine there is only one respondent per group [one respondent per oval]; in such cases, the intraclass correlation [ICC(1)] $\approx$ 1.00, and all variance is group-level variance [Bliese, 2000]; thus, the observed correlations are $r_{between}$ correlations).

Why does it matter whether one's sample is a single-organization sample versus a multi-organization sample? These two types of samples tend to address two fundamentally different sets of phenomena. On the one hand, a single-organization sample does a good job of estimating individual-level relationships (r_{within}) but is unable to assess the existence of group-level mod-

erators (i.e., is unable to assess external validity of r_{within} across groups). On the other hand, a multi-organization sample tends to do a poor job of estimating individual-level relationships (r_{within}), because almost all of the individual-level variance is confounded with group-level (between-groups) variance (i.e., because there is approximately one individual per group). As such, MTurk researchers are in the unfortunate position of studying $r_{between}$ while assuming that they are studying r_{within} (thus committing an ecological fallacy).

Although Landers and Behrend appropriately acknowledge that any use of a convenience sample should include a thoughtful discussion of the external validity of that sample, these authors do not acknowledge that such discussions must recognize that an ecological fallacy has likely been committed in the case of multi-organization convenience samples. Further, with so few respondents per organization, it is typically impossible for researchers with multi-organization convenience samples to meaningfully assess external validity (i.e., to assess the absence of moderators of r_{within} across groups). Thus MTurk researchers (and researchers using other multi-organization convenience samples) should carefully question whether an MTurk sample is an appropriate sample to use for investigating an individual-level phenomenon, and this issue should be discussed in the article itself.

Therefore, we assert that Landers and Behrend's conclusion, "for studies with the goal of generalizing to the global worker pool, MTurk may be ideal" (2015, p. 18; which implies that the diversity of MTurk samples is beneficial for external validity), requires a caveat. MTurk inferences involving the "global worker pool" should often be limited to inferences about $r_{between}$ (or perhaps about r_{total}, which is a combination of $r_{between}$ and r_{within}), but if there are differences across organizations in the variables being studied, MTurk data can only support confounded inferences about what happens within organizations (r_{within}). As such, we suggest that the multi-organizational diversity of MTurk samples can actually harm a common type of external validity inference—because when only one participant from each organization is sampled, it is nearly impossible to infer the extent to which the observed $r_{between}$ relationship represents the targeted r_{within} relationship (i.e., r_{within} relationships are often the focus of I-O psychology, but it is difficult to know whether multi-organization results provide relevant estimates of what happens within organizations).

Conclusions

To us, the best way to study external validity is to use a diverse set of many single-organization samples. This enables researchers to achieve two goals: (a) to precisely estimate individual-level relationships (r_{within}) for each group, unconfounded with group-level relationships ($r_{between}$), and (b) to assess the

common notion of external validity (i.e., whether individual-level relationships have the same magnitude across groups—the absence of group-level moderators). In contrast, a single-organization sample (or "organizational sample"; Landers & Behrend, 2015) only achieves the first goal (i.e., estimating individual-level relationships). Finally, a multi-organization sample (e.g., MTurk sample, employed college student sample, online panel, or snowball sample) typically achieves neither goal. That is, multi-organization samples cannot cleanly estimate individual-level relationships in a way that is unconfounded with group-level relationships, nor can they assess the external validity of individual-level relationships across groups.

The point we are making here is not that crowdsourced samples are useless; indeed, we have used these samples ourselves (Feitosa, Joseph, & Newman, 2015). Rather, we are merely emphasizing that, when crowdsourced samples (or employed student samples, online panels, or snowball samples) are used as an alternative to organizational samples, it is essential that researchers think clearly about the level of analysis of each hypothesis. From our reading of Landers and Behrend's (2015) thoughtful discussion of omitted moderator variables, we assume they might agree with this point. Again, we just want to make sure that we as organizational researchers are testing what we think we are testing (i.e., r_{between} vs. r_{within}) and to ensure that we are making more precise external validity claims.

References

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Dansereau, F., Alluto, J. A., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewood Cliffs, NJ: Prentice Hall.

Feitosa, J., Joseph, D., & Newman, D. (2015). Crowdsourcing survey methods and measurement equivalence: A caveat about countries whose primary language is not English. *Personality and Individual Differences, 75*, 47–52.

Hernandez, I., Newman, D. A., & Jeon, G. (in press). Twitter analysis: Methods for data management and a word count dictionary to measure city-level job satisfaction. In S. Tonidandel, E. King, & J. Cortina (Eds.), *Big data at work: The data science revolution and organizational psychology*. New York, NY: Routledge.

Hulin, C. L. (1966). Effects of community characteristics on measures of job satisfaction. *Journal of Applied Psychology, 50*(2), 185–192.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology: Perspectives on Science and Practice*.

Ostroff, C. (1993). The effects of climate and personal influences on individual behavior and attitudes in organizations. *Organizational Behavior and Human Decision Processes, 56*(1), 56–90.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, 13*(2), 238–241.

Thorndike, E. L. (1939). On the fallacy of imputing the correlations found for groups to the individuals or smaller groups composing them. *American Journal of Psychology, 52*, 122–124.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2*(2), 121–134.

A Convenient Solution: Using MTurk To Sample From Hard-To-Reach Populations

Nicholas A. Smith
The Pennsylvania State University

Isaac E. Sabat
George Mason University

Larry R. Martinez, Kayla Weaver, and Shi Xu
The Pennsylvania State University

We agree with Landers and Behrend's (2015) proposition that Amazon's Mechanical Turk (MTurk) may provide great opportunities for organizational research samples. However, some groups are characteristically difficult to recruit because they are stigmatized or socially disenfranchised (Birman, 2005; Miller, Forte, Wilson, & Greene, 2006; Sullivan & Cain, 2004; see Campbell, Adams, & Patterson, 2008, for a review). These groups may include individuals who have not previously been the focus of much organizational research, such as those of low socioeconomic status; individuals with disabilities; lesbian, gay, bisexual, or trans-

Nicholas Smith, School of Hospitality Management, The Pennsylvania State University; Isaac Sabat, Department of Psychology, George Mason University; Larry Martinez, Kayla Weaver, and Shi Xu, School of Hospitality Management, The Pennsylvania State University.

Correspondence concerning this article should be addressed to Nicholas Smith, School of Hospitality Management, The Pennsylvania State University, 201 Mateer Building, University Park, PA 16802. E-mail: nas277@psu.edu